

New big data collection, storage, and processing requirements as identified from the EU-SysFlex use cases

Deliverable 5.3



EU-SysFlex

PROGRAMME	H2020 COMPETITIVE LOW CARBON ENERGY 2017-2-SMART-GRIDS
GRANT AGREEMENT NUMBER	773505
PROJECT ACRONYM	EU-SysFlex
DOCUMENT	Deliverable 5.3
TYPE (DISTRIBUTION LEVEL)	<input checked="" type="checkbox"/> Public <input type="checkbox"/> Confidential <input type="checkbox"/> Restricted
DATE OF DELIVERY	28/10/2020
STATUS AND VERSION	V1 - FINAL
NUMBER OF PAGES	229
Work Package / TASK RELATED	WP5 / Task 5.3
Work Package / TASK RESPONSIBLE	Kalle Kukk – Elering / Alan Tkaczyk – University of Tartu (UTartu)
AUTHOR (S)	<p>Main authors: Alan Tkaczyk (UTartu), Stanislav Sochynskyi (UTartu), Kalle Kukk (Elering)</p> <p>Contributing authors: Philippe Szczech (AKKA), Florentin Dam (AKKA), Riccardo Benedetti (AKKA), Gunay Abdullayeva (UTartu), Oleksandr Kurylenko (UTartu), Grzegorz Gucwa (PSE), Simon Siöstedt (AFRY), Ulf Roar Aakenes (Enoco), Nick Good (Upside), Mitchell Curtis (Upside), Steve Wattam (Upside), Katharina Brauns (IEE), Nicolas Kuhaupt (IEE), Angela Sakh (Cybernetica), Ville Sokk (Cybernetica), Wiebke Albers (innogy), Jan Budke (innogy), Carmen Calpe (innogy), Maik Staudt (Mitnetz Strom), Karin Maria Lehtmetts (Elering)</p>

DOCUMENT HISTORY

VERS	ISSUE DATE	CONTENT AND CHANGES
V1	28/10/2020	Report submitted to EC

DOCUMENT APPROVERS

PARTNER	APPROVER
Elering	Kalle Kukk – Work Package Leader
EDF	Marie-Ann Evans – Technical Manager
EirGrid, EDF, SONI, VITO, innogy,	EU-SysFlex Project Management Board
Elering, EDP, EurActiv, Zabala	
EirGrid	John Lowry – Project Coordinator

TABLE OF CONTENTS

EXECUTIVE SUMMARY	11
INTRODUCTION	19
1. BIG DATA FRAMEWORK	20
1.1 ABSTRACT	20
1.2 INTRODUCTION	20
1.3 METHODOLOGY AND APPROACH	21
1.4 RESULTS AND CONCLUSIONS	26
2. BIG DATA REQUIREMENTS	34
2.1 IDENTIFICATION OF TECHNICAL REQUIREMENTS	34
2.1.1 ABSTRACT	34
2.1.2 INTRODUCTION	34
2.1.3 METHODOLOGY AND APPROACH	35
2.1.4 RESULTS AND CONCLUSIONS	35
2.2 COMPARATIVE STUDY OF EXISTING SOLUTIONS	40
2.2.1 ABSTRACT	40
2.2.2 INTRODUCTION	41
2.2.3 METHODOLOGY AND APPROACH	42
2.2.4 RESULTS AND CONCLUSIONS	43
3. COST OF DATA EXCHANGE FOR ENERGY SERVICE PROVIDERS	49
3.1 ABSTRACT	49
3.2 INTRODUCTION	50
3.3 METHODOLOGY AND APPROACH	54
3.4 RESULTS	57
3.5 DISCUSSION AND CONCLUSIONS	60
4. CASE STUDIES	62
4.1 BASELINE MODELS AND RESILIENCE OF SERVICE DELIVERY	62
4.1.1 ABSTRACT	62
4.1.2 INTRODUCTION	63
4.1.3 METHODOLOGY AND APPROACH	64
4.1.4 TESTING SELECTED BASELINE METHODOLOGIES TO REAL DATASETS	68
4.1.5 CONCLUSIONS	75
4.2 PREDICTION OF AVAILABILITIES AND QUANTITIES OF FLEXIBILITY SERVICES	76
4.2.1 ABSTRACT	76
4.2.2 INTRODUCTION	77
4.2.3 METHODOLOGY AND APPROACH	78
4.2.4 RESULTS AND CONCLUSIONS	81
4.3 NEAR REAL-TIME RESIDUAL LOAD FORECASTING AT GRID POINTS	89
4.3.1 ABSTRACT	89
4.3.2 INTRODUCTION	90
4.3.3 METHODOLOGY AND APPROACH	94
4.3.4 RESULTS AND CONCLUSIONS	98
4.4 DATA EXCHANGE BETWEEN DSO AND TSO	103
4.4.1 ABSTRACT	103
4.4.2 INTRODUCTION	104
4.4.3 CONTEXT AND APPROACH	105
4.4.4 RESULTS AND CONCLUSIONS	109
4.5 FORECASTING IN INTEGRATED ENERGY SYSTEMS	117
4.5.1 ABSTRACT	117
4.5.2 INTRODUCTION	118
4.5.3 AIM	118
4.5.4 USE CASE #1: INVESTIGATING ROBUST LSTM ARCHITECTURE FOR ENERGY TIME SERIES FORECASTING	119
4.5.5 USE CASE #2: DEVELOPMENT OF CNN-BASED MODELS FOR SHORT-TERM LOAD FORECASTING	123
4.5.6 GENERAL CONCLUSIONS	129
4.6 PRIVACY-PRESERVING DATA ANALYSIS	130
4.6.1 ABSTRACT	130
4.6.2 INTRODUCTION	131
4.6.3 METHODOLOGY AND APPROACH	133

4.6.4	RESULTS AND CONCLUSIONS	134
4.7	DEVELOPMENT OF A BIG DATA SYSTEM FOR THE ELECTRICITY MARKET	135
4.7.1	ABSTRACT	135
4.7.2	INTRODUCTION	136
4.7.3	METHODOLOGY AND APPROACH.....	136
4.7.4	USE CASE DESCRIPTION, TECHNICAL IMPLEMENTATION AND RESULTS	138
4.7.5	CONCLUSIONS.....	145
BIBLIOGRAPHY		147
COPYRIGHT.....		151
ANNEX I – BIG DATA FRAMEWORKS: SUPPLEMENTARY INFORMATION ABOUT COMPONENTS		152
ANNEX II – IDENTIFICATION OF TECHNICAL REQUIREMENTS.....		160
ANNEX III – COMPARATIVE STUDY OF EXISTING SOLUTIONS: DETAILED ESTIMATION OF SELECTED SOLUTION		173
ANNEX IV – DATA EXCHANGE BETWEEN DSO AND TSO: SUPPLEMENTARY INFORMATION ABOUT EU REGULATIONS		212
ANNEX V - PRIVACY-PRESERVING DATA ANALYSIS: PROOF OF CONCEPT.....		220

LIST OF FIGURES

FIGURE 1.1 EU-SYSFLEX TASK 5.3 DOMAIN MODEL OF IDENTIFIED REQUIREMENTS.....	27
FIGURE 1.2 IDENTIFIED BIG DATA FEATURES.....	28
FIGURE 1.3 REFERENCE ARCHITECTURE.....	31
FIGURE 2.1 NUMBER OF IDENTIFIED BIG DATA REQUIREMENTS PER EACH USE CASE.....	40
FIGURE 2.2 COMPARATIVE ANALYSIS PROCEDURE.....	42
FIGURE 2.3 COMPARISON OF ANALYZED SOLUTIONS (RADAR CHART).....	45
FIGURE 2.4 COMPARISON OF ANALYZED SOLUTIONS (BAR GRAPH).....	46
FIGURE 2.5 RESULTS OF EVALUATION OF EXISTING SOLUTIONS BY GROUP OF REQUIREMENTS (RADAR CHARTS).....	47
FIGURE 3.1 DSO PREDICTING FLEXIBILITY IN THE LONG TERM (INVESTMENT) TIMESCALE.....	52
FIGURE 3.2 SYSTEM OPERATOR PREDICTING FLEXIBILITY IN MEDIUM TERM (OPERATIONAL) TIMESCALE.....	53
FIGURE 3.3 THE BIG DATA ARCHITECTURE TO ENABLE 100 000 DEVICES TO COMMUNICATE AND PROVIDE FLEXIBILITY SERVICES.	58
FIGURE 4.1 CLOUD-BASED REAL-TIME MONITORING SYSTEM WITH AGGREGATED DATA	64
FIGURE 4.2 A) POWER CONSUMPTION HOUR-BY-HOUR B) HOURLY CONSUMPTION 24H, 90 DAYS	70
FIGURE 4.3 CONSUMPTION PATTERNS FOR SELECTED DAYS OF FIGURE 4.1	70
FIGURE 4.4 ABSOLUTE PERCENT DEVIATION FROM THE ESTIMATED BASELINE AND THE ACTUAL CONSUMPTION	70
FIGURE 4.5 A) HOUR-BY-HOUR CONSUMPTION B) 24H X 10 DAYS	71
FIGURE 4.6 A) ENERNOC BASELINE B) ABSOLUTE VALUE OF THE DEVIATION.....	72
FIGURE 4.7 RESULTS OF APPLYING UK MODEL	72
FIGURE 4.8 RESULTS OF APPLYING THE AVERAGE MODEL.....	72
FIGURE 4.9 RESULTS OF APPLYING THE DAILY PROFILE MODEL.....	72
FIGURE 4.10 VARIATIONS IN DAILY CONSUMPTION PATTERNS	73
FIGURE 4.11 VALUES OF MAPE ON ALL DATASETS IN THE STUDY (KURYLENKO, 2020).	75
FIGURE 4.12 PLANNING TIMESCALES	79
FIGURE 4.13 FORECAST SYSTEM FOR CALCULATING THE RESIDUAL LOAD FORECASTUSED IN THE GERMAN DEMONSTRATOR IN WP6....	92
FIGURE 4.14 OVERVIEW BIG DATA CLUSTER AND THE RESOURCES	95
FIGURE 4.15 DEEP NEURAL NETWORK ARCHITECTURE WITH LSTM LAYERS.....	98
FIGURE 4.16 SCALING RESULTS OF EVALUATING A LSTM MODEL ON THE MNIST TEST DATA	99
FIGURE 4.17 EVALUATION OF THE MEAN PROCESSING TIME FOR 1 TO 10,000 MODELS FOR 48H FORECAST	101
FIGURE 4.18 REDISPATCH PROCESS APPROACH BASED ON CONNECTT+ APPROACH AND IN LINE WITH GERMAN EU-SYSFLEX DEMONSTRATOR.....	110
FIGURE 4.19 HIGH-LEVEL ARCHITECTURE OF FLEXIBILITY PLATFORM CONCEPT.....	113
FIGURE 4.20 DATA MODEL OF FLEXIBILITY PLATFORM CONCEPT	115
FIGURE 4.21 THE AVERAGE SCALED RMSE ERRORS ARE RELATIVE TO PERSISTENCE	121
FIGURE 4.22 THE 36 HOURS TIME STEPS PREDICTIONS OF THE PERSISTENCE, ARIMA, AND UNIVARIATE STACK LSTM MODELS	122
FIGURE 4.23 VALUES OF RMSE (KWH) ON ALL DATA SETS	127
FIGURE 4.24 VALUES OF MAE (KWH) ON ALL DATA SETS.....	127
FIGURE 4.25 VALUES OF MAPE (%) OF ALL DATA SETS.....	128
FIGURE 4.26 DIAGRAM OF COMPONENTS PARTICIPATING IN THE BASELINE CALCULATION PROCESS.....	134
FIGURE 4.27 METHODOLOGY OVERVIEW.....	137
FIGURE 4.28 OVERVIEW OF USE CASE DATA FLOWS.....	138
FIGURE 4.29 CORRELATION MATRIX	141
FIGURE 4.30 ENERGY CONSUMPTION IN ESTONIA	142
FIGURE 4.31 MSE BY EPOCH ON THE VALIDATION SET	143
FIGURE 4.32 THE INITIAL BIG DATA CLUSTER	144

ANNEX III: FIGURE A.1 ENTSO-E MODULES.....	174
ANNEX III: FIGURE A.2 ESTFEED MODULES.....	177
ANNEX III: FIGURE A.3 ESTFEED MODULES.....	177
ANNEX III: FIGURE A.4 DOWNLOAD MY DATA SCENARIO.....	180
ANNEX III: FIGURE A.5 CONNECT MY DATA SCENARIO	180
ANNEX III: FIGURE A.6 GREEN BUTTON ACTORS.....	181
ANNEX III: FIGURE A.7 OPENESPI SOFTWARE ARCHITECTURE	182
ANNEX III: FIGURE A.8 OPENESPI FRAMEWORKS	182
ANNEX III: FIGURE A.9 METER DATA HANDLING AND END USER ACCESS PERSONAL DATA PROCESS.....	183
ANNEX III: FIGURE A.10 SUPPORTING MARKET PROCESSES	184
ANNEX V: FIGURE A.11 DIAGRAM OF COMPONENTS PARTICIPATING IN THE BASELINE CALCULATION PROCESS	224
ANNEX V: FIGURE A.12 SEQUENCE DIAGRAM OF BASELINE CALCULATION PROCESS.....	225

LIST OF TABLES

TABLE 1.1 LIMITS OF TRADITIONAL ICT SOLUTIONS VS BIG DATA SOLUTIONS.....	23
TABLE 1.2 BIG DATA FUNCTIONALITIES.....	24
TABLE 1.3 VELOCITY AND VOLUME ESTIMATION OF EU-SYSFLEX WP9 NEEDS.....	29
TABLE 1.4 LIST OF THE SELECTED BIG DATA COMPONENTS.....	30
TABLE 2.1 BIG DATA REQUIREMENTS IN SYSTEM USE CASES.....	38
TABLE 2.2 RESULTS OF EVALUATION OF EXISTING SOLUTIONS BY GROUP OF REQUIREMENTS	43
TABLE 2.3 SCALE OF ASSESSMENT OF REQUIREMENTS.....	44
TABLE 2.4 NORMALIZED RESULTS OF EVALUATION OF EXISTING SOLUTIONS BY GROUP OF REQUIREMENTS	44
TABLE 3.1 MONTHLY COST OF THE DIFFERENT SERVICE TYPES IN THE BIG DATA ARCHITECTURE FOR TWO CASES WITH DIFFERENT STORAGE CAPABILITY	59
TABLE 3.2 PERCENTAGE OF THE TOTAL MONTHLY COST FOR TWO CASES WITH DIFFERENT STORAGE CAPABILITY	60
TABLE 3.3 PERCENTAGE OF THE TOTAL MONTHLY COST FOR TWO CASES WITH DIFFERENT STORAGE CAPABILITY	61
TABLE 4.1 METHODS DESCRIPTION	65
TABLE 4.2 THE CHARACTERSITICS OF THE FOUR MODELS IN THE BALTIC TSO TEST.....	68
TABLE 4.3 THE RESULTS APPLYTING REGULAR BASELINE MODELS ON THE CONSUMPTION OF AN INDUSTRY PLAYER.....	71
TABLE 4.4 THE RESULTS OF APPLYING REGULAR BASELINE MODELS ON THE CONSUMPTION OF AN INDUSTRY PLAYER	73
TABLE 4.5 INTRODUCTION OF APPLIED METHODS AND USED METRICS	74
TABLE 4.6 ESTIMATION OF ANNUAL NUMBER OF RECORDS NEEDED FOR PREDICTION OF QUANTITIES AND AVAILABILITIES OF FLEXIBILITY SERVICES.....	86
TABLE 4.7 OVERVIEW OF BIG DATA RESOURCES ON THE HADOOP CLUSTER	95
TABLE 4.8 AVERAGE SCALED EVALUATION METRICS RESULTS OF MULTIVARIATE STANDARD LSTM, UNIVARIATE STANDARD LSTM AND MULTIVARIATE STACK LSTM FOR EACH TIME SERIES.....	123
TABLE 4.9 HIERARCHY OF THE MODELS.....	126
TABLE 4.10 MEAN VALUES OF RMSE (KWH) ON ALL DATA SETS	128
TABLE 4.11 MEAN VALUES OF MAE (KWH) ON ALL DATA SETS.....	129
TABLE 4.12 MEAN VALUES OF MAPE (%) ON ALL DATA SETS	129
ANNEX I: TABLE A.1 BIG DATA COMPONENTS EXAMPLES	153
ANNEX II: TABLE A.2 TECHNICAL REQUIREMENTS FOR SELECTED SUCS	160
ANNEX III: TABLE A.3 ENTSO-E OPDE DESCRIPTION.....	173
ANNEX III: TABLE A.4 ESTFEED DESCRIPTION.....	176
ANNEX III: TABLE A.5 GREEN BUTTON DESCRIPTION	180
ANNEX III: TABLE A.6 ELHUB DESCRIPTION	183
ANNEX III: TABLE A.7 EVALUATION RATES	186
ANNEX III: TABLE A.8 ENTSO-E OPDE EVALUATION TABLE.....	186
ANNEX III: TABLE A.9 ESTONIAN DATA EXCHANGE PLATFORM ESTFEED, DATAHUB, E-ELERING PORTAL EVALUATION TABLE	193
ANNEX III: TABLE A.10 GREEN BUTTON EVALUATION TABLE	199
ANNEX III: TABLE A.11 NORWEGIAN ELHUB.....	206
ANNEX IV: TABLE A.12 GUIDELINE ON SYSTEM OPERATION	212
ANNEX IV: TABLE A.13 TSO PROPOSALS RELATED TO DATA EXCHANGE	215
ANNEX IV: TABLE A.14 GUIDELINE ON ELECTRICITY BALANCING	216
ANNEX IV: TABLE A.15 NETWORK CODE ON DEMAND CONNECTION	216
ANNEX IV: TABLE A.16 NETWORKCODE ON REQUIREMENTS FOR GRID CONNECTION OF GENERATORS.....	217
ANNEX IV: TABLE A.17 DIRECTIVE ON COMMON RULES FOR THE INTEARNAL MARKET IN ELECTRICITY	217
ANNEX IV: TABLE A.18 REGULATION ON THE INTERNAL MARKET FOR ELECTRICITY	219

ABBREVIATIONS AND ACRONYMS

AEMO	Australian Electricity Market Operator
aFRR	Automatic Frequency Restoration Reserve
API	Application Programming Interface
ARIMA	Autoregressive Integrated Moving Average
AIS	All-Island Power System
AS	Ancillary Services
BRP	Balance Responsible Party
BSP	Balancing Service Provider
CE	Continental Europe
CEP	Clean Energy Package
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DCC	Network Code on Demand Connection
DEP	Data Exchange Platform
DER	Distributed Energy Resources
DR	Demand Response
DRR	Dynamic Reactive Response
DSO	Distribution System Operator
DSR	Demand Side Response
DS3	Delivering a Secure, Sustainable Electricity System
EB GL	Guideline on Electricity Balancing
EC	European Commission
ECCo SP	ENTSO-E Communication & Connectivity Service Platform
e.g.	<i>exempli gratia</i> : for example
EHV	Extra High Voltage
EIC	Energy Identification Code
ENTSO-E	European Network of Transmission System Operators for Electricity
ESCO	Energy Service Company
etc.	<i>et cetera</i> : and other similar things
EU	European Union
EU-SysFlex	Pan-European System with efficient, coordinated use of flexibilities for the integration of a large share of Renewable Energy Source
FCR	Frequency Containment Reserve
FFR	Fast Frequency Response
FPFAPR	Fast Post-Fault Power Recovery
FRR	Frequency Restoration Reserve
FRT	Fault ride through
FSP	Flexibility Service Provider
GB	Gigabyte
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
h	Hour

HDD	Hard Disk Drives
HDFS	Hadoop File System
HV	High Voltage
HFoT	High Five of Ten
ID	Identifier
IEA	International Energy Agency
IOPS	Input/output operations per second
IoT	Internet of Things
IT	Information Technology
kB	Kilobyte
KORRR	Key Organisational Requirements, Roles and Responsibilities
kV	Kilovolt
LMP	Locational Marginal Pricing
LSI	Largest System Infeed
LSTM	Long-short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MB	Megabyte
mFRR	Manual Frequency Restoration Reserve
min	Minute
MVA	Mega Volt Ampere
MVAr	Mega Volt Ampere Reactive
MW	Megawatt
NAESB	The North American Standard Board
N/A	Not Applicable
PB	Petabyte
PET	Privacy-Enhancing Technologies
PoC	Proof of Concept
POR	Primary Operating Reserve
RES	Renewable Energy Sources
ReLU	Rectified Liner Unit
RfG NC	Network Code on Requirements for Grid Connection of Generators
RR	Replacement Reserve
RRD	Replacement Reserve Desynchronised
RNN	Recurrent Neural Network
RRS	Replacement Reserve Synchronised
RMSE	Root Mean Square Error
RoCoF	Rate of Change of Frequency
S2S	Sequence to Sequence
SCADA	Supervisory Control and Data Acquisition
sec	Second
SGU	Significant Grid User
SIR	Synchronous Inertial Response
SLA	Service Level Agreement
SNSP	System Non-Synchronous Penetration

SO	System Operator
SOGL	Guideline on System Operation
SONI	System Operator Northern Ireland
SOR	Secondary Operating Reserve
SPFoT	Similar Profile five of Ten
SSD	Solid State Drives
SSRP	Steady State Reactive Power
STOR	Short Term Operating Reserve
STLF	Short-Term Load Forecasting
SMAPE	Symmetric Mean Absolute Percentage Error
SUC	System Use Case
TB	Terabyte
tbd	To be defined
TiB	Tebibyte
TOR	Tertiary Operating Reserve
TSO	Transmission System Operator
VDIFD	Voltage Dip-Induced Frequency Deviation
VM	Virtual Machine
WP	Work Package

EXECUTIVE SUMMARY

As an output of Task 5.3 within the EU-SysFlex project, this deliverable report describes “big data”¹ considerations and solutions for flexible energy systems. Task 5.3 is a forward-looking innovation task addressing particular big data needs and requests of several EU-SysFlex work packages. There are relevant links of conducted studies to ongoing activities within coordination of flexibilities connected to distribution, coordination of centralised and decentralised flexibilities, cross-border and cross-sectoral data management WPs. The work of Task 5.3 seeks to facilitate quick and safe operation in an increasingly decentralised situation with numerous stakeholders, with the needs of managing data in very close to real-time.

Task 5.3 brought on board and started an active dialogue between TSOs, DSOs, technology providers, consultants, aggregators and researchers, resulting in several case studies aligned with the EU-SysFlex demonstration goals. To fulfil ambitions and the needs of tomorrow’s flexible energy system, all the participants of Task 5.3 contributed to a comprehensive study of data collection, storage, and processing requirements and functionality.

There are nine Key Messages based on the work of Task 5.3:

Key Message #1: Forty-eight (48) big data related requirements were identified in the EU-SysFlex data exchange use cases. These requirements are currently addressed only partly in selected data platform solutions around the globe, as these platforms were initially developed to address different needs. The largest gaps occur in the area of support for flexibility services and near real-time communication with SCADA systems. *(For more details related to this Key Message, please see Chapter 2 of this report.)*

Key Message #2: A big data framework can be designed to match all aforementioned 48 big data requirements identified in the EU-SysFlex data exchange use cases. Nevertheless, the framework relies on a combination of various open-source components and not only one unique multi-purpose component. *(For more details related to this Key Message, please see Chapter 1 and Section 4.7 of this report.)*

Key Message #3: An assessment of data exchange cost for energy service providers reveals that the flexibility service start-up cost is dominant over the data storage capacity cost where high-throughput capacity is necessary. However, the storage capacity cost becomes dominant when a larger amount of storage is required; therefore, the sizing of the storage needs to be investigated thoroughly. *(For more details related to this Key Message, please see Chapter 3 of this report.)*

Key Message #4: Traditional assessments of baseline electricity load during a demand response event are based on analytical calculations which assume repeating patterns and/or regularity. To increase accuracy during irregular periods, more advanced models are needed. Machine learning with input from influencing ambient

¹ Data is considered to be big data in the moment when it becomes difficult to process it with traditional information technology systems.

factors can contribute to a significant improvement in baseline modelling. *(For more details related to this Key Message, please see Section 4.1 of this report.)*

Key Message #5: Data requirements for prediction and monitoring flexibilities are likely to increase significantly due to increased demand for flexibility services and the trend to provision by smaller units. *(For more details related to this Key Message, please see Section 4.2 of this report.)*

Key Message #6: By applying the latest machine learning methods, it is possible to compute and deliver to DSOs more than 1500 residual load forecasts for transformer stations every 15 minutes, employing a marginal amount of computational power. *(For more details related to this Key Message, please see Section 4.3 of this report.)*

Key Message #7: As the European power system evolves toward increasing complexity and decentralisation, the need for system flexibility and therefore DSO-TSO data exchange increases accordingly. This work reports on two EU-SysFlex demonstrators which validate two different approaches to improve the DSO-TSO data exchanges for flexibility usage. *(For more details related to this Key Message, please see Section 4.4 of this report.)*

Key Message #8: Neural network-based machine learning methods can be applied to short-term load forecasting in the energy sector, with high performance compared to industry-standard baseline models. Multivariate LSTM models exhibited high accuracy while univariate LSTM models exhibited high robustness in various scenarios; CNN-based models exhibited high accuracy in forecasts of future flexibility. *(For more details related to this Key Message, please see Section 4.5 of this report.)*

Key Message #9: Adopting privacy by design and privacy-enhancing technologies will enable adherence to data protection laws, increase consumer trust and enable new business models. However, the technologies may be disruptive to current approaches, meaning that privacy and the security to ensure it should already be considered from the early stages of (re-)designing a system. *(For more details related to this Key Message, please see Section 4.6 of this report.)*

This deliverable report from EU-SysFlex Task 5.3 is structured into the following sections to systematically address this complex issue. First, the “Big data framework” chapter discusses big data components or frameworks to support the energy system’s flexibility. Then, the “Big data requirements” chapter consists of 2 sections. First, an overview identifies and sets performance, functionality, security, privacy and big data requirements for big data systems. This overview is complemented with the comparative study of existing data exchange solutions to identify the gaps in top-notch systems. Next, the “Cost of data exchange” chapter presents the cost of maintaining data exchange by analysing the case of aggregator that plays a role between the DSO and a flexible platform to provide flexibility. Finally, the “Case studies” chapter contains seven studies that address several challenges such as privacy and ownership of the data, confidentiality, security, business, and GDPR restrictions. Short summaries of the mentioned chapters and sections are provided below.

Summary of Chapter 1 – Big Data Framework

The purpose of big data framework chapter is to identify a list of big data components or frameworks fulfilling the set of requirements linked to the data exchanges and more broadly to the data management, needed for supporting flexibility services.

For each requirement, a set of pre-selected big data components is reviewed and compared, and one of them is finally selected to privilege the advanced state-of-the-art and open-source component. The study concludes with the design of a big data architecture based on the lambda architectural pattern and the chosen components.

The big data architecture designed in this chapter represents an example of a highly scalable and fault-tolerant system, intended to handle the fast-ever increasing data volume encountered in the electricity domain.

Summary of Chapter 2 – Big Data Requirements

Section 2.1 – Identification of technical requirements

The overview aimed to identify technical requirements for the data exchanges based on the data exchange system use cases described in Task 5.2. These requirements relate to performance, functionality, security, privacy, and big data with the focus is on the latter category. The big data requirements identified serve as input for further works in the following chapters for more detailed analyses.

Around 70 technical requirements were identified, out of which 48 relate to big data. The largest number of big data requirements are related to use-cases on data collection, sub-meter data management, flexibility activation, flexibility prequalification and bidding, and DER-SCADA data exchange. Use cases on personal data and a listing of suppliers and ESCOs involve no big data requirements. Description of data is based on the needs of WP9 demonstrators. It means that testing of some big data requirements does not refer to the need to process massive data there because demonstrators remain on the proof-of-concept level. However, if implementing several use cases on a commercial level, including cross-border, it would result in actual amounts of big data.

Section 2.2 – Comparative study of existing solutions

The comparative study aimed to explore already implemented solutions on meeting the technical requirements for the data exchange described in Identification of technical requirements in order to enhance further design and development of data exchange solutions with an extended range of functionalities. The following four solutions were analysed: OPDE (ENTSO-E), Estfeed + Data Hub (Estonia), Elhub (Norway) and Green Button (USA). Each of the solutions was evaluated in terms of meeting the requirements for data exchange.

The aggregated evaluation indicated that none of the analysed solutions meets all the requirements for data exchange defined in the EU-SysFlex project. The most supported areas are in meter data exchange handling,

security and privacy management. The largest gaps occur in the area of support for flexibility services and near real-time communication with SCADA systems. The results presented in the form of tables and graphs illustrating the degree of compliance of the solutions with the groups of requirements.

Summary of Chapter 3 – Cost of Data Exchange for Energy Service Providers

The cost of data exchange case study investigated the cost of handling data communication by analysing the cost of data exchange in the case of an aggregator that plays a role between the DSO and a flexible platform that enables flexibility provision. The specific use case for the aggregator considered 100 000 devices located in the same region with the same service types (such as Spark, Kafka) but in one case with 10 TB and in another case with 1000 TB storage capacity.

The results from the cost assessment showed that the cost for the public cloud-based deployment with 10 TB of storage capacity would be 5 700 EUR/month where the storage capacity itself only accounted for 5% of the total cost. On the other hand, the cost with the same data architecture and with the same services but with 1000 TB of storage would cost 23 400 EUR/month where the storage capacity itself accounted for 77% of the total cost.

The cost distributions between the two cases show that the service types excluding the storage capacity are dominant for a case with limited requirement for storage. In order to provide flexibility services as an aggregator, all discussed service types are needed in data architecture design.

Therefore, two main conclusions can be stated:

1. The flexibility service start-up cost is dominant over the storage capacity where efficient processing of data, load forecasting and high-throughput capacity is necessary to run the services rather than providing a large amount of storage capacity.
2. It is essential to determine stakeholders' responsibility for managing storage capacity and storage volume that is needed to run the services to prevent over-dimension of the application. Otherwise, it will lead to a higher cost for an end-user.

Summary of Chapter 4 – Case Studies

Section 4.1 – Baseline models and resilience of service delivery

Quantifying the resilience of service delivery presents evaluation and testing results of models that estimate what the consumption would have been without a Demand Response event (DR) – baseline models.

An essential element in a flexibility market is that a consumer should be able to offer a reduction (or increase) in their consumption in order to release capacity for other more critical consumers. A payment shall reward released capacity (or consumption of excess), and it is, therefore, necessary to document that the reduction is delivered as agreed. Verification of contracted reduction of consumption can be done in many ways.

Testing the different baseline models on real datasets reveals the models' ability to calculate correctly during "very irregular" consumption patterns. The EnerNOC, the UK Model, the Average and the Daily Profile models are widely used, and representatives of such models are tested. Advanced deep learning models have also been tested on the same real datasets. The tests show that the simplest models, such as Average and Daily profiles, are the most accurate and often outperform the more complex ones. The models have also been tested on single large consumers, where the result shows that none of them can estimate adequately. In such cases of a DR request, baselines produced before event combined with metering data and real-time monitoring is a better solution. The focus has been on baseline models that meet the requirements of simplicity and transparency. Payment is involved, and therefore such characteristics are essential to be able to avoid attempts of gaming and to reduce the burden of administration.

Section 4.2 – Prediction of availabilities and quantities of flexibility services

The prediction case study explored estimates of the data requirements (in terms of the number of records) for predicting availabilities and quantities of flexibility services to support power system, e.g., frequency response services.

After describing the different timescales over which prediction of availabilities and quantities of flexibility services are conducted, estimates of such quantities are presented through case studies, which demonstrate how these predictions are made in practice and the volumes and types of data associated with those predictions. Based on estimations of existing data requirements and forecasts of increased flexibility requirements, the future data requirements to predict availabilities and quantities of flexibility services are shown to be significant. The number of individual data records required for prediction of availabilities and quantities of flexibility services in real-time for the case study with the highest requirements (Great Britain) was estimated to be 11 038 million/year. The challenge associated with dealing with such a large amount of data may be ameliorated if aggregation of data before reaching the system operator is allowed. But that is dependent on the rules prescribed by each system operator.

Besides the indication of the scale of the data requirements required for prediction of availabilities and quantities of flexibility services, a major finding was the need for clarity and transparency on the methodologies for prediction. Particularly at the investment and operational planning timescales, the methodologies (and hence data requirements) were unclear. Clarity on these methodologies could encourage potential flexibility providers (especially those with long lead times, or for those whose primary purpose is not a provision of flexibility services) to make their equipment suitable for providing flexibility.

Section 4.3 – Near real-time residual load forecasting at grid points

The load forecasting case study examined three different approaches to measuring the processing time for the timely provision of all forecasts of a residual load to the DSO in a near real-time system. Near real-time means that the forecasts are continuously delivered every 15 minutes to the DSO calculated in the forecast system of the German demonstrator. It means the delivering of the residual load forecast of a large number of transformer stations around 1500 in time under at least 15 minutes for the active and reactive power.

The forecasts are generated using a Long Short-Term Memory (LSTM) machine learning approach. It involves using a big data approach with a Hadoop Cluster which is compared to the usage of a stand-alone server by using at first up to 32 central processing units (CPU) and in a second evaluation phase 2 graphics per units (GPUs). The main challenge was that the focus is on evaluation in a near real-time system rather than on the probably more widely used variant by training a variety of forecasting models. In this case, the forecast model is used to calculate the forecast for a one-time step that only includes a small input data set instead of massive data sets with the purpose for training a deep neural network. However, for the Hadoop cluster and the GPU approach, there is still a certain amount of traffic that needs to be taken into account, which is time consuming compared to the fast calculation of the forecast itself. Finally, it was demonstrated that these forecasts could be generated with all three approaches.

The comparison showed that the Hadoop Cluster and the GPU did not outperform the usage of CPUs. For the delivery of about 3000 forecasts (including active and reactive power) under 15 minutes, the usage of a stand-alone server with 5 CPUs is still sufficient.

Section 4.4 – Data exchange between DSO and TSO

This section describes two approaches based on EU-SysFlex demonstrators in the context of EU regulation in terms of data exchange for flexibility usage – German demonstrator and Flexibility Platform demonstrator.

The rising need for system flexibility creates new requirements for data exchanges. These requirements mainly refer to the data exchange between DSOs and TSOs, as more and more flexible resources are connected to the distribution grid, and both stakeholders are in a rising need for system flexibility.

Several clauses can be identified in both pre- and post-Clean Energy Package (CEP) EU regulations, which concern DSO-TSO data exchange for flexibility usage. Regulations approved already before CEP include several network codes. CEP itself has resulted in amended electricity market directive and electricity market regulation, potentially followed by new network codes and implementing acts still to be established.

In German demonstrators' approach, each system operator selects needed flexibilities to solve congestions in its grid (subsidiarity principle) and determines the maximum flexibility potential for the upper system operator. Only the grid data relevant for re-dispatch, including costs, sensitivities, and flexibility limitations, is exchanged with

the upstream system operator. The connecting system operator initiates the flexibility activation in his grid based on his own need and the received request by the upstream system operator.

The starting point of Flexibility Platform demonstrator is to maximise the liquidity of and easy access to the flexibility market through a single flexibility market concept. Such concept implies massive flows of data: in terms of several stakeholders, services/products as well as from data granularity perspective, and up to very-near-real-time exchanges. In a single market, several marketplaces or market platforms can coexist and even compete with each other, and therefore, it is essential to ensure interoperability. TSO-DSO data exchanges result from the need to ensure that flexibilities are procured and activated most efficiently, including a case of joint procurement and that all flexibilities have access to the market regardless of where they are physically connected.

Section 4.5 – Forecasting in integrated energy systems

The forecasting case study investigated Demand Response (DR) mechanism performance assessment on the two use cases with the application of recurrent neural network (RNN) and convolutional neural networks (CNN) with various configurations.

DR mechanisms facilitate balancing the demand-supply ratio and provide greater flexibility within the electric grid. When the demand needs to be reduced, a DR event is activated on the market, and the amount of reduced electricity consumption is measured to assess the DR performance.

The results of the networks are compared to the Naïve and ARIMA benchmark models. Additional to these benchmark models, CNN based models are also compared to the industry-standard baseline models (Asymmetric HFoT, SPFoT, Average, Daily Profile).

The conducted experiments have shown both LSTM and CNN based models outperform the baseline models in most time series based on RMSE, MAE, and MAPE evaluation metrics. Stack LSTM and CNN-LSTM models show more stable results over all-time series.

Section 4.6 – Privacy-preserving data analysis

Growing data exchange in the energy sector between different systems increases the intentional/unintentional storage of personal data across them. The ultimate goal of every system that works with personal data is to protect customers and lower risks associated with unsuitable data storage and processing. The privacy-preserving data analysis chapter presented the results of including privacy-enhancing technologies (PETs) in the electricity market by conducting a case study. Additionally, the proof of concept implementation is turned into a demonstrator under WP9 to better showcase PETs within the project.

The case results confirm the possibility to use PETs to various use cases in the electricity market. Results of this work highlight the importance to include PETs to protect consumer data on the early stage of designing or re-designing of existing systems and functionalities as it will enable innovative ways to execute approaches and

processes. Laws and regulations will start impacting the electricity sector more and more, as it uses and stores highly sensitive data: during the study, the privacy issues highlighted by The European Consumer Organization in regard to consumer data availability for aggregators was identified.

Section 4.7 – Development of a big data system for the electricity market

The case study presents the particular use of big data components and architectural design patterns identified in the Chapter 1 on “Big Data Framework” for the development of a big data system that could reinforce the electric market. Two use cases were chosen, implemented and deployed over big data system: the computation of the near real-time prediction of electrical consumption based on streaming data and the batch measurement of the prediction of consumption for a longer time scale. Both use cases are representative for the electric market, which wants to leverage large quantities of data from smart meters or various sensors in a real-time manner and generate different types of predictions such as consumption, flexibility availability. These use cases were described with technical details of the big data systems built for them.

Also, it has been experimented an improved version of the Seq2Seq prediction algorithm in with residual LSTM network supported by attention mechanisms. The obtained results show sufficient accuracy for the prediction of electricity consumption in the context of the data used.

INTRODUCTION

The EU-SysFlex project seeks to enable the pan-European power system to utilise efficient coordinated flexibilities in order to integrate a large share of renewable energy sources. As part of the EU-SysFlex project, Work Package 5 aims at providing recommendations for data management in flexibility services when applied in a large scale (on an IT perspective) and developing customer-centric data exchange models for flexible market design serving all stakeholders (transmission system operators, distribution system operators, suppliers, flexibility providers, energy service companies, etc.) and enabling data exchange across borders.

As an output of Task 5.3 within the EU-SysFlex project, this deliverable report describes “big data”² considerations and solutions for flexible energy systems. Specifically, Task 5.3 investigates the options for implementing massive data exchanges, with appropriate data storage and data processing as required for extensive use of flexibility services, with increasing number of flexibility providers (including decentralised generation and prosumers). It proposes solutions to enhance existing architectures and develop data exchange platforms in the energy domain. Some proposed solutions will be tested in the data exchange demonstrators in Work Package 9 of EU-SysFlex. The objectives of Task 5.3 addressed in this deliverable report are as follows:

- A. Identification of technical requirements for the data exchanges based on the use cases from Task 5.2 (e.g. requirements relating to data exchange, storage and processing volume, time constraints, security and privacy) and comparison of existing solutions (such as ENTSO-E’s “OPDE” and national data exchange platforms) regarding the identified requirements;
- B. Elicitation of applicable methodologies and big data frameworks for effective data exchange, data storage and processing of streaming and historical data, and estimation of resources and costs;
- C. Consideration of massive data analysis tasks essential for the success of flexibility services, e.g., quantifying the reliability of service delivery of technologies and solutions – it will be crucial to characterize the extent to which flexibility service providers deliver the response they have contracted to provide; prediction of availabilities and quantities of flexibility services – it will be necessary for the system operators to know how much flexibility will be available; estimation of missing grid measurements – e.g. due to outages or meter failures; data exchange optimization between DSO and TSO for flexibility benefits calculation;
- D. Implementation and demonstration of some of the above data exchange, data storage and data processing functionalities required for the success of cross-border and cross-sector demonstrations with WP9, adhering to the requirements of volumetry, time, security, privacy.

The identified requirements, elicited methodologies and new functionalities for data exchanges, data storage and data processing contribute to formulating the flexibility roadmap for the European grid.

² Data is considered to be big data in the moment when it becomes difficult to process it with traditional information technology systems.

1. BIG DATA FRAMEWORK

Main section authors: Riccardo Benedetti (AKKA), Philippe Szczech (AKKA), Florentin Dam (AKKA)

1.1 ABSTRACT

The purpose of this chapter is the identification of a list of big data components or frameworks fulfilling the set of requirements linked to the data exchanges, and more broadly to the data management, needed for supporting flexibility services.

The study starts with the analysis of the requirements to determine the 3 V's key concepts: Volume, Velocity and Value, meant respectively to answer the following questions: *what is the amount of data? What is the minimum processing rate? What are the business issues to solve?* Based on the answers, it is possible to outline the main features of the requested components. It has been figured out that they cover all the requirements of a complete big data system: *ingestion, storage, processing, querying, governance and security*. For each requirement, a set of pre-selected big data components is reviewed and compared, and one of them is finally selected to privilege the advanced state-of-the-art and open-source component. The study concludes with the design of a big data architecture based on the lambda architectural pattern and the chosen components.

The big data architecture designed in this chapter represents an example of a highly scalable and fault-tolerant system, intended to handle the fast-ever increasing data volume encountered in the electricity domain.

1.2 INTRODUCTION

1.2.1 AIM

Big data framework hereafter aims at identifying some big data frameworks fulfilling a set of technical requirements regarding data exchange and big data topics. It is related to the following statement of the EU-SysFlex project DoA: *"Elicitation of applicable methodologies and big data frameworks for effective data exchange, data storage and processing of streaming and historical data (...) to achieve the identified requirements."* The mentioned "big data framework" expression has been interpreted as a set of technical components or specific tools implemented to address the needs of the big data domain. In the following paragraphs, the terms "big data framework" and "big data component" are used indifferently.

The output of this work is a selection of big data components and a reference architecture combining all of them into a consistent IT system.

1.2.2 CONTEXT

This chapter is linked to Identification of technical requirements work of Task 5.3 (see Chapter 2.1). The latter provides the list of technical requirements to be used for the elicitation of the big data frameworks. Those

requirements come mainly from the various data exchange system use cases of Task 5.2, which gives the context for their interpretation. Each requirement comes with insights on the volume and types of data which could be encountered in WP9 demonstrators where they will be implemented.

This report is also tied to the Cost of data exchange for energy service providers (see Chapter 3). Indeed, the big data architecture designed will be costed for the specific case of an aggregator which would deploy it to support its various processes.

Finally, the results of this work will also be used for the needs of WP9 demonstrators by developing partially the big data architecture and by connecting this one to the demonstrators through the data exchange platform Estfeed. The work related to implementation is done as part of the Development of a big data system for the electricity market in Chapter 4.7.

1.3 METHODOLOGY AND APPROACH

1.3.1 OVERVIEW

The overview presents the course of main action that were performed in order to identify a set of big data frameworks fulfilling the requirements provided by the Identification of technical requirements in Chapter 2.1.

Firstly, requirements analysis was conducted according to the 3 Vs criteria met in the big data landscape. More precisely, this analysis has been conducted to answer the following questions:

- **Volume:** what is the amount of data the big data framework should be able to handle?
- **Velocity:** what minimum processing rate should the big data framework handle?
- **Value:** what business issue do the requirements refer to?

Analysis was completed within the Identification of the domain model and the Identification of the big data features linked to these requirements, such as “data collection”, “processing”, “querying”.

Secondly, the first list of IT big data components was selected which could cover the big data features identified in the first step and according to the current state-of-the-art of the big data domain. This list has been refined to take into account the different constraints expressed by the requirements to provide finally, for each requirement, a set of big data components which can enable its implementation.

Thirdly, the elicited big data components were gathered in a reference architecture after having compared two overall architectural patterns: the Lambda and the Kappa ones.

1.3.2 DISCUSSION ON INNOVATION

This section introduces the added-values of big data solution and the specific characteristics of those challenges which can fall in the big data domain, selecting the big data technologies suitable for their resolution. These

elements were used during the requirements analysis in order to verify that they are linked to a big data problem but also to help the selection of the most beneficial big data components.

CHARACTERISTICS OF A BIG DATA PROBLEM

What is a big data problem? In a broad sense, data is considered to be big data in the moment when it becomes difficult to process it with traditional information technology systems.

A big data problem is also recognized when it involves one or all of these characteristics:

- **High volume:** interpreted as the size of the amount of data which is massive in case of big data, usually involving datasets of terabytes to petabytes.
- **High velocity:** a characteristic related to streaming data. It refers to the capability to handle fast streams in order to limit the loss of information.
- **High variety:** a capability to manage different types of data such as structured (e.g. tables in relational databases), semi-structured (e.g. XML or JSON) and unstructured data (e.g. data logs) as well as unstructured data represented in many formats including text, images, videos, audios.

WHAT CAN BIG DATA SOLUTIONS DO? WHAT ARE THE POTENTIAL BENEFITS?

The adoption of a big data solution leads to several benefits which implicitly solve many non-functional requirements that are hard to get over with traditional solutions. The main benefits of a big data solution are:

Flexibility

The term flexibility refers to the ability to handle heterogeneous data format. Traditionally, data have always been stored in a well-structured database where each instance had to respect a fixed schema. The added value of big data is the capability of managing also unstructured data, which nowadays are becoming even more widespread, and perform on them high-speed data transformation.

Scalability

The most popular big data platforms, such as Hadoop and Spark, offer the possibility to scale efficiently. Comparing to traditional SQL database, potential growth of data does not undermine the analytical performance, thanks to the possibility of adding additional nodes (workers) to the cluster.

Real-time computation

Big data offers the possibility to perform real-time computation. While some tasks do not necessarily need for a fast result and so they can be easily managed with traditional batch approaches. At the same time, other tasks such as anomaly detection, reactive notification systems and real-time prediction may depend on the responsiveness of the system.

Machine learning

Modern Machine Learning applications, in particular Deep Learning, rely on a vast dataset. In order to be able to manage these volumes of data, it might be necessary to have a working parallel cluster on which to perform machine learning on big data or for big compute tasks. Big data solutions offer this capability.

Break data locality

Whereas traditionally storage systems used to deal with conventional tapes and disk drives (physical data locality), nowadays, there is a migration towards distributed and fault-tolerant cloud systems. Cloud technologies provide a sort of abstracted data locality because the user can access the data as they reside in his file system, even though they are physically spread over the network.

Since this transition is not final yet and many solutions still rely on the traditional approach, the goal of big data is not only to provide support for the cloud but also to fill the gap between traditional storage and next-generation storage.

Merge data silos

A data silo is a collection of information, or data storage, in an organization which is isolated and not easily accessible. Removing data silos can facilitate the retrieval of the right information in a reasonable time and reduce the costs of eventual duplicate. In big data, this is accomplished by gathering all data in a single central data warehouse.

TABLE 1.1 LIMITS OF TRADITIONAL ICT SOLUTIONS VS BIG DATA SOLUTIONS

Limits of traditional IT architecture	The 'big data' proposal
Storage cost and complex scalability The approach used in the traditional system is the shared storage based commonly on technologies such as Storage Area Network (SAN) or Network Attached Storage (NAS). The limitations arise when the volume of data starts to increase, leading to OPEX or CAPEX costs.	Distributed Storage big data components like the Hadoop Distributed File System (HDFS) provides a high-level distributed storage where the cost per GB significantly drops. This solution also provides data replication in order to improve the availability and implement the fault-tolerance mechanism.
Enterprise hardware and software licensing In term of scalability, the cost of proprietary hardware can be burdensome. As the organizations grow, the consequent hardware adaptation can be costly for what concern both software licenses and physical resources.	Off-the-shelf hardware and software open-source Hadoop allows building a high-performant distributed infrastructure based on Off-the-shelf hardware (i.e. common IT components broadly used, interchangeable) instead of enterprise hardware and this with a more reasonable cost. Similarly, the pricing to scale a Hadoop cluster is significantly cheaper respect a proprietary cluster. Moreover, the building of a big data system can be based on completely free and open-source components.
Organizational complexity Term complexity means the difficulties in administrating	Administrative simplicity The common Hadoop big data infrastructure is very intuitive

massive modular architectures, which are often based on the integration of many different heterogeneous tools. Such administration typically requires a multitude of competencies such as system administrators, DBAs, application server teams, storage teams, and network teams.	and allows to manage thousands of distributed data nodes with just one administrator.
Skimping on data quality Traditional systems usually try to improve the performance by pre-aggregating data and filtering in order to reduce the volume to analyse. This approach inevitably leads to loss of information which can impact negatively on the resulting accuracy and confidence.	Boost the data quality Data stored in HDFS can be easily analysed with high performant big data processing tools. There is no more need for pre-processing, and so the data remain atomics “as-is”. It increases the possibility of finding correlation and so produce more accurate results. In addition to that, the time for data loading in a Hadoop solution is lower.
Moving Data to the Programs Traditional solutions based in relational databases rely on static applications in which data must be loaded and transported to them. Data transportation has to take care of network bandwidth limitations which can often represent a potential bottleneck.	Moving Programs to the Data Hadoop solution exploits parallel computation. Data in HDFS are spread over the disks, and the applications run on each one in parallel. It implies that the application move to the data and not vice-versa. It is also no secret the benefit of parallel programming over the sequential paradigm.

LIMITS OF CURRENT ICT ARCHITECTURE

In addition to the benefits mentioned before, another reason to encourage the transition towards a big data solution can be even more evident by highlighting the limits of the traditional IT architecture. It also describes how a big data solution would overcome these limitations.

BIG DATA FUNCTIONALITIES

This paragraph describes the standard features encountered in a general-purpose big data architecture. It also precise some terms, definitions and concepts frequently used in the big data landscape. Besides, it is mentioned, as well as the technical components traditionally used to implement the mentioned features.

TABLE 1.2 BIG DATA FUNCTIONALITIES

FEATURE	SUB-FEATURES	DESCRIPTION	Example of ICT components supporting the feature
INGESTION	BROKER	The set of frameworks used for collecting and transferring data from different sources. A broker allows buffering the data coming from different kind of IoT sources in order not directly to access them.	Apache Kafka, Apache Flume, RabbitMQ, Apache ActiveMQ, Artemis
	INTEGRATION	The set of frameworks used for ingesting data which reside in heterogeneous sources.	Apache Sqoop, Apache Kafka, Apache Nifi, Apache Gobblin
PROCESSING	BATCH	The process or action of transforming a given amount of previously collected data within a single job.	Hadoop MapReduce, Apache Spark, Apache Tez, Apache Flink
	STREAMING	The process or action of transforming a real-	Apache Spark (Spark Streaming API),

FEATURE	SUB-FEATURES	DESCRIPTION	Example of ICT components supporting the feature
		time stream of incoming data through a steady job.	Apache Kafka (Kafka Streams API), Apache Storm, Apache Samza, Apache Flink
STORAGE	RELATIONAL	The set of traditional databases for storage and retrieval of structured data based on SQL syntax.	MySQL, PostgreSQL, SQLite
	NO-SQL	The set of databases for storage and retrieval of unstructured data.	Hadoop File System (HDFS), MongoDB, Apache HBase, Apache CouchDB, Apache Cassandra
	NEW SQL	The set of modern database management system designed to provide atomicity, consistency, isolation and durability properties and NO-SQL performances.	MariaDB
GOVERNANCE & SECURITY	AUTHENTICATION	The process or action of verifying the identity of a user or process (proving or showing something to be correct, genuine, or valid).	Kerberos protocol, Apache Snort, Apache Knox
	AUTHORIZATION	The process or action of verifying if an authenticated user or process has the right to access a specific resource.	Apache Ranger, Apache Knox, Apache Sentry
	ANONYMIZATION	The process or action of de-identify data by removing or masking any personal information in order to accomplish the GDPR.	ARX
	GOVERNANCE	The data management tool which enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data.	Apache Atlas
ANALYTICS	MACHINE LEARNING	The branch of Artificial Intelligence based on the optimization of mathematical models in order to extract knowledge from a set of data.	scikit-learn, Apache Spark MLlib, Apache SystemML, Weka
	DEEP LEARNING	A subset of Machine Learning algorithms that concern the usage of models based on neural networks.	TensorFlow + Keras, PyTorch, DL4J
VISUALIZATION	MONITORING	The set of tools which provide a user-friendly interface to show analytical results, charts and performance indicators.	Kibana, Elastic Search
QUERYING	OLAP QUERIES	The set of frameworks to perform interactive and fast queries on massive multidimensional data.	Apache Drill, Apache Druid, Apache Kylin, Pentaho BI
	OLTP QUERIES	Transactional queries in relational DBs with ACID properties	(functionality integrated with all relational DBMS – see ACID properties)
INFRASTRUCTURE	CLOUD	Technology which allows managing, through a remote server, a pool of hardware and software resources. The service is usually offered by a provider, through subscription.	AWS Cloud, Microsoft Azure, Google Cloud, IBM Cloud, OpenStack
	HARDWARE	The set of invariant physical components (computers, processors, storage media, GPUs) which compose a data processing system.	Nvidia GPU, Graphcore, Mythic, Intel Core processor Family, Kingston SSD
OTHER	COLLABORATION & DEVELOPMENT	Set of tools used for AGILE team development.	Git, Anaconda, Jupyter Notebook, Spyder, Apache Zeppelin, Watson Studio, IntelliJ Idea

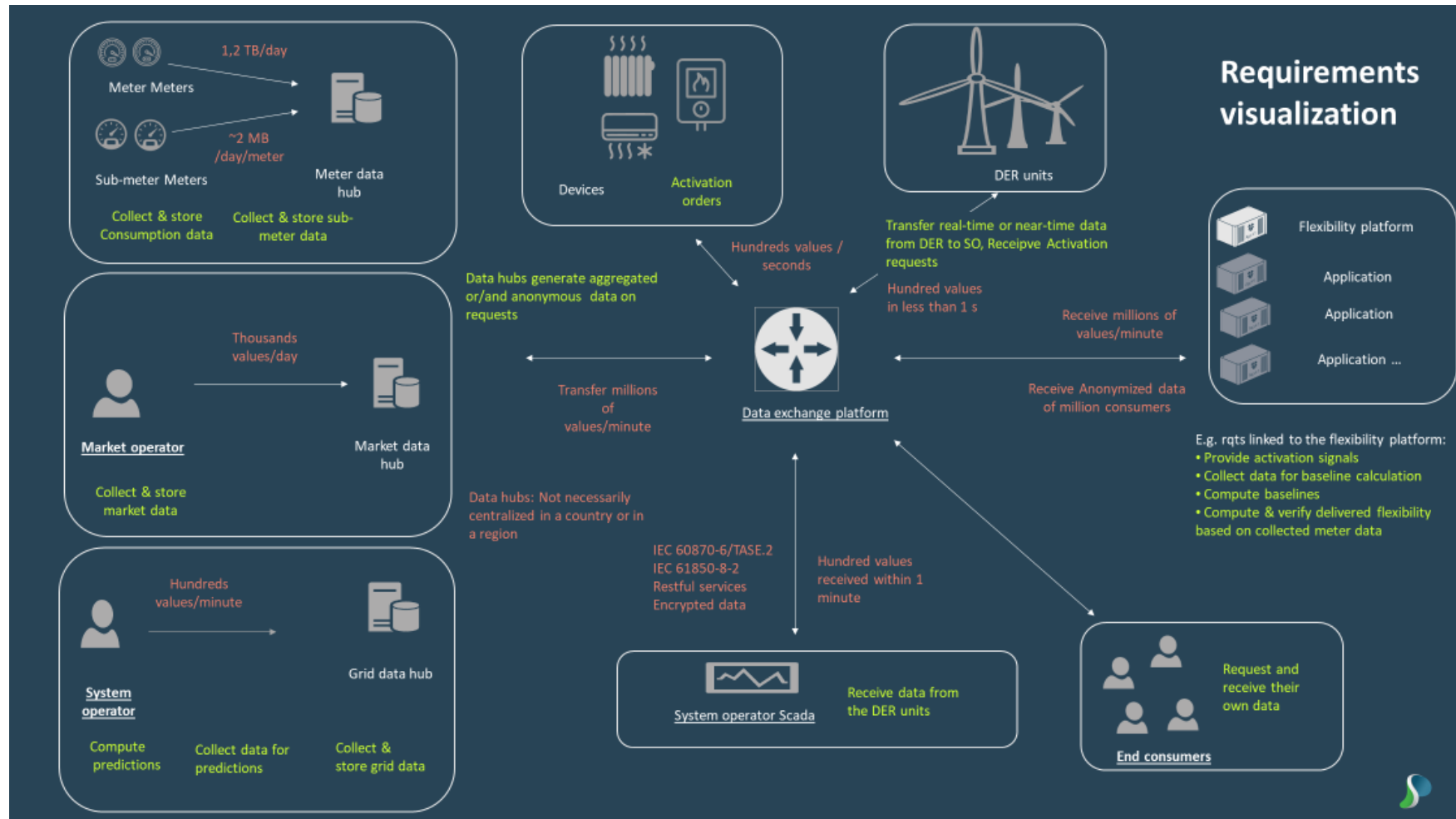
1.4 RESULTS AND CONCLUSIONS

1.4.1 RESULTS

REQUIREMENTS ANALYSIS

EU-SYSFLEX DOMAIN MODEL

One result of the requirement analysis is the Identification of the systems mentioned by the requirements as well as the principal data flows occurring between them. It results in the schema below, which includes graphically represented elements with the description of the most critical requirements.



Note: in this schema, green typography is used for functional requirements and red one for technical requirements.

FIGURE 1.1 EU-SYSFLEX TASK 5.3 DOMAIN MODEL OF IDENTIFIED REQUIREMENTS

IDENTIFICATION OF REQUESTED BIG DATA FEATURES

One objective of the requirements analysis was the Identification of the big data features (taken from Table 1.2 big data functionalities) referred by the requirements. In essence, the following features have been identified:

- **Data ingestion:** it is mentioned mainly in the *SUC Data Collection* and by some requirements of *SUC Data Transfer*, *SUC DER-SCADA data exchange*, *SUC Flexibility activation*, *SUC Flexibility baseline*, *SUC Flexibility bids* and *SUC Sub-meter data*;
- **Data storage:** it is a feature often implicit in some requirements belonging to Data Collection, e.g. *SUC Data Collection* and *SUC Sub-meter data*;
- **Data analytics:** it comes from the requirements of *SUC Flexibility prediction* and *SUC Flexibility baseline*;
- **Data processing:** for batch processing, it refers to requirements in *SUC Aggregate data* and *SUC Anonymize data*, and for real-time processing, to some requirements related to the Flexibility Platform, like *SUC Flexibility activation*, *SUC Flexibility baseline* and *SUC Flexibility bids*;
- **Data Querying:** it emerges from all those requirements aimed to make information available to data owners and external applications in different SUCs;
- **Data Governance & Security:** it comes from requirements of the *SUC Authentication of data users*, *SUC Access permissions' management* and *SUC Data logs*.

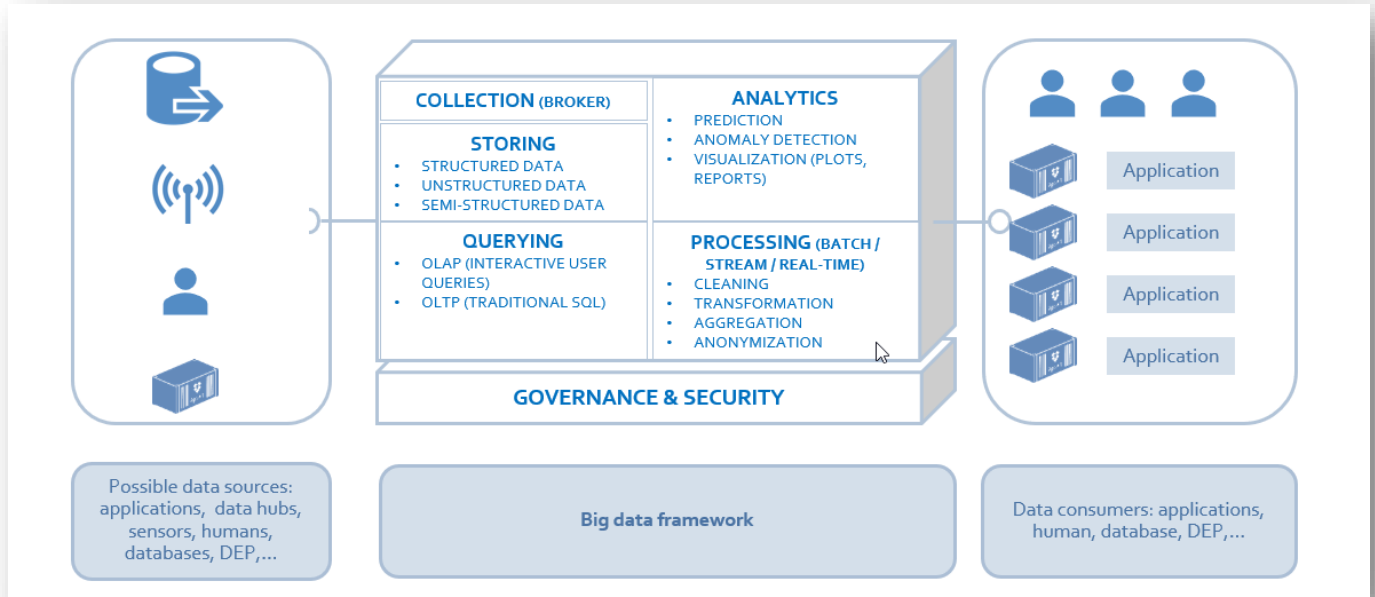


FIGURE 1.2 IDENTIFIED BIG DATA FEATURES

This list of features represents the central reference point to select the big data components which can better support the EU-SysFlex system use cases.

VELOCITY AND VOLUME ASSESSMENT

The previous sections have provided some valid qualitative motivations why one would need a big data solution. In this section, a qualitative estimation of the problem is proposed. The Identification of technical requirements in Chapter 2.1 provides information about the volume and the velocity which could be encountered **in the context of the WP9 demos** for each requirement. A few of these requirements will be developed concerning the elicited big data framework. All the individual figures were added up to get the maximum constraints in terms of volume and velocity the big data framework should face. The traffic of data across the current platform was calculated.

From each requirement was given the data velocity in MB/s and from the latter, which made it possible to derive an approximation of the volume needed per year (in TB).

Two different types of data exchange were identified: the one which concerns the ingestion aspect, where the data from the source are gathered into the data hubs, and the one which involves the traffic across the DEP (Data Exchange Platform).

TABLE 1.3 VELOCITY AND VOLUME ESTIMATION OF EU-SYSFLEX WP9 NEEDS

	Velocity (MB/s)	Volume (TB per year)
From data sources to data hubs:	165 ± 15 MB/s	5200 ± 150 TB
Grid data (SUC DC)	60 ± 5	1900 ± 50
Meter data (SUC DC)	55 ± 5	1700 ± 50
Market Data (SUC DC)	50 ± 5	1600 ± 50
Through the DEP:	50 ± 2 MB/s	1700 ± 70 TB
Between DER & SO's SCADA (SUC DER-SCADA)	0,30 ± 0,05	9 ± 1
Between data hubs, data owner & applications (SUC DT)	49 ± 1	1600 ± 50
Sharing security logs (SUC LOGS)	3 ± 0,5	95 ± 5
Sharing authentication info & access permissions (SUC AUTH & SUC AUTHZN)	non-significant	non-significant
Flexibility baselines (SUC FB)	non-significant	non-significant
Flexibility bids (SUC FBIDS)	non-significant	non-significant
Flexibility activations (SUC FA)	non-significant	non-significant
Flexibility verifications (SUC FVERIF)	non-significant	non-significant
Flexibility predictions (SUC FVERIF)	non-significant	non-significant

The estimations from the table would be useful to prove the minimum hardware capabilities required to handle one year of data, especially in terms of memory, with the current architecture. It is essential to point out that

many big data technologies rely on data replication to support fault-tolerance. Therefore, for the big data solution storage capabilities should be provided significantly higher from the ones mentioned in the table.

In conclusion, the magnitude of the values shown in the table is proof that a big data solution is required to address the Identification of technical requirements, even just on the level of WP9 demonstrators.

THE LIST OF THE SELECTED BIG DATA COMPONENTS

Table 1.4 presents the big data framework selected after the requirement analysis. The rationale and the complete description of those components are in the Annex I – big data frameworks: Supplementary information about components. Those components could be used to implement the different technical requirements, details regarding this point are in Annex I.

TABLE 1.4 LIST OF THE SELECTED BIG DATA COMPONENTS

big data feature	Selected big data component	What is?
Data Ingestion	Apache Kafka	A general publish-subscribe based messaging system (Broker)
	Apache NiFi	A data flow manager between software systems
Data Storing	Apache HDFS	A distributed file system designed to run on commodity hardware
	MongoDB	A consistent and fault-tolerant non-relational storage system
	Apache Cassandra	An available and fault-tolerant non-relational storage system
Batch data processing	Apache Spark	A unified analytics engine for big data processing
Stream data processing	Spark Streaming	A Spark library specific for near-real-time processing
Data Querying	Apache Hive	A SQL-like data warehouse software running over HDFS
	Apache Drill	A distributed SQL query engine for data-intensive for interactive analysis of large-scale datasets
	Apache Presto	A high-performing distributed SQL query engine
Data analytics	TensorFlow + Keras / PyTorch / Deeplearning4j	Programming libraries for machine and deep learning.
Data security	Apache Ranger	A framework to enable, monitor and manage comprehensive data security across Hadoop
	Apache Knox Gateway	A security perimeter for interacting with the big data platform through the REST APIs
	ARX	Data anonymization tool to secure sensitive personal data.
cluster & resource management	Apache YARN / Apache Mesos	Resource management and job scheduling technology in the distributed big data cluster
	Apache Zookeeper	A centralized service for providing configuration information, naming, synchronization and group services over large clusters in distributed systems
	Apache Oozie	A workflow scheduler system to manage Hadoop batch jobs

THE BIG DATA REFERENCE ARCHITECTURE

It is possible to put together and interface these big data components in order to produce a system which can serve multiple purposes related to the EU-SysFlex domain. In practice, this system could be partially or entirely implemented at any business domain (such as system operator, market operator) and be integrated with the already existing systems as well as in the new ones. This system could be interfaced with a data exchange platform such as Estfeed.

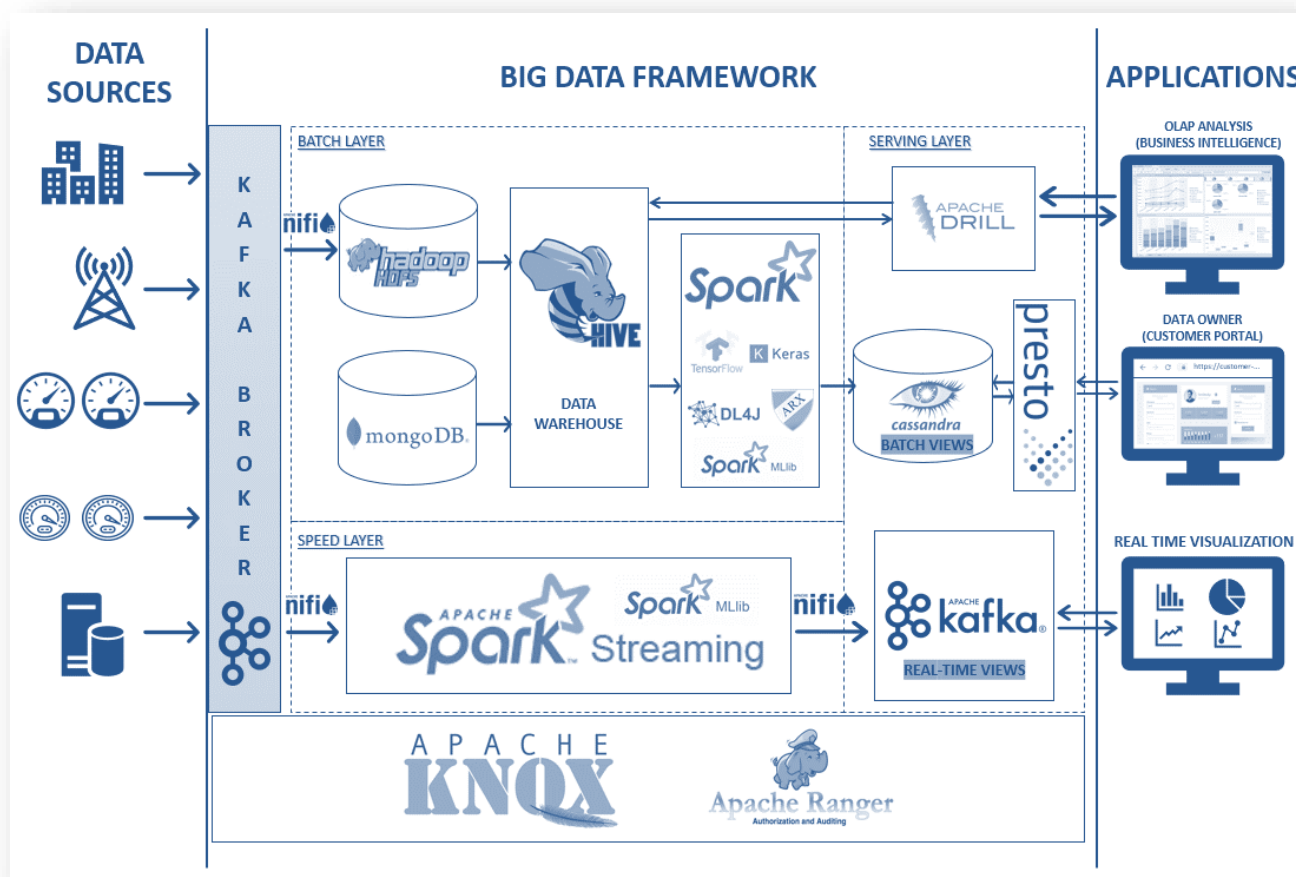


FIGURE 1.3 REFERENCE ARCHITECTURE

On Figure 1.3 the final big data architecture designed after having described and compared all the big data frameworks proposed to accomplish the tasks expressed by the EU-SysFlex requirements, in particular, it consists of a concrete solution of Figure 1.2 introduced at the beginning of the chapter.

In contrast with the DEP, the big data system does not represent only a middleware for data exchange between data sources, data owners and applications, but a solution to address **data ingestion, storing, processing, analysis, querying, security and governance in a large geographical scale**.

In order to **collect** all the data generated from the external world (e.g. data sources and market domain), an ingestion layer was proposed. It consisted of a distributed and **horizontally scalable** cluster of Kafka brokers.

Taking into account the lambda architecture principles, the collected data are consumed in parallel by both a **batch layer**, for the **periodical** and transactional jobs, and a **speed layer**, for the **near-real-time** use cases. This data transfer can be facilitated with the usage of NiFi to redirect the flow towards different destinations.

Another essential difference between the batch and speed layer concerns the presence of storage technology. In the speed layer, data must be handled and processed **on-the-fly**. In many cases, this can be achieved with a solution based only on Kafka (i.e. Kafka Streaming). However, as explained in the Data Processing section, the usage of Spark Streaming allows to perform many additional tasks, such as real-time **machine learning** and eventually provides structured streaming. The results of the real-time processing (**real-time views**) can be remitted into a Kafka topic to be dynamically consumed by external applications.

On the other side, the batch layer relies strongly on the presence of a **data lake** where the IoT data flow goes. The Hadoop Distributed File System (HDFS) is currently the best state-of-art solution to handle **high-throughput**. MongoDB was added to store those data that are not inherently dynamic and mutable, such as metadata, user information, and authorizations.

The core of the big data system is the **data warehouse**, where both new and historical data are archived, cleaned and structured. Data warehouse, Apache Hive, sits over the data like to provide summarized SQL-like view of data and facilitate their analysis and querying.

The way to explore the data warehouse is essentially through Online Analytical Processes (**OLAP**). Apache Drill is the OLAP gold standard that business analysts use to perform data mining, interactive queries and analytics at a massive scale.

On the other hand, one of the objectives of the EU-SysFlex scenario is to provide a way for data owner and application to get their data (e.g. their details about consumption) through a customer portal. Giving them the possibility to query the data warehouse directly can be awkward in term of performances and response time. For that reason, a batch processing engine was used, like Apache Spark, to **pre-compute** these **queries** and store the result into **batch views** in the serving layer. Apache Cassandra can host these batch views. This storage solution aims to privilege **high-availability** from a reading perspective. Finally, a fast querying engine like Presto can be employed to extract the requested data from these batch views.

Apache Spark is not limited to load data into a batch view “as is” but can also run processing task like data aggregation and anonymization. In these cases, there is a need for **aggregated views** and **anonymized views**. Moreover, Spark can be used to perform an **extensive calculation** (i.e. flexibility baselines) and run deep learning frameworks (such as TensorFlow, PyTorch, DL4J) to train **predictive models** (i.e. flexibility predictions).

Under the hood of these complex big data system, it was discovered that some frameworks responsible for ensuring data **governance** and **security**, in particular **authentication mechanisms**, **authorization management** and **access control**. These frameworks are Apache Knox and Ranger.

1.4.2 CONCLUSIONS

In this research, a list of big data components was introduced, which can be used to implement most of the requirements defined in the Identification of technical requirements in Chapter 2.1. Each component is described in terms of features requested, and clear rationales have been provided to explain why a specific component has been selected. Finally, all the components have been put together into a reference architecture. The latter constitutes a system which could be used partially or totally in the different domains described by the Identification of technical requirements.

From a functional perspective, the designed solution enables global data exchange between data sources (such as IoT data, market data) and data consumers (end-users or applications) by collecting the data, processing them and making them available; all of this at a massive scale. This architecture suits also some non-functional requirements which are mainly: the usage of state-of-the-art **open-source components**, a very **highly-scalable** solution enabling the big data implementation at different levels (local level, regional level or national level) and the **security integration** (additionally to the possible security mechanisms brought by the infrastructure hosting or surrounding the big data system such as network infrastructure, data exchange platform).

2. BIG DATA REQUIREMENTS

2.1 IDENTIFICATION OF TECHNICAL REQUIREMENTS

Main section author: Kalle Kukk (Elering)

2.1.1 ABSTRACT

Data exchange requirements may relate to performance, functionality, security, privacy, and big data. Around 70 technical requirements were identified based on the system use cases described in Task 5.2, out of which 48 relate to big data. The largest number of big data requirements are related to use cases on data collection, sub-meter data management, flexibility activation, flexibility prequalification and bidding, and DER-SCADA data exchange. Use cases on personal data and a listing of suppliers and ESCOs involve no big data requirements. Description of data is based on the needs of WP9 demonstrators. It means that testing of some big data requirements does not refer to the need to process actual massive data there because demonstrators remain on the proof-of-concept level. However, if implementing several use cases on a commercial level, including cross-border, it would result in real amounts of big data.

2.1.2 INTRODUCTION

2.1.2.1 AIM

This work aims to identify technical requirements for the data exchanges based on the system use cases described in Task 5.2. These requirements relate to performance, functionality, security, privacy, and big data, the focus is on the latter category. The big data requirements identified serve as input for a few of the following chapters for more detailed analyses.

2.1.2.2 CONTEXT

In Task 5.2 of EU-SysFlex, sixteen system use cases relevant to the data exchange have been identified and described. The use cases range from flexibility market-specific ones (such as flexibility prequalification, bidding, activation, baseline calculation) to generic ones relevant not only for the flexibility market but for other business processes (such as data collection, data transfer, authentication, access permission management).

Data description is based on the needs of WP9 demonstrators. It means that testing of some big data requirements does not refer to the need to process actual massive data there because demonstrators remain on the proof-of-concept level. However, if implementing several use cases on a commercial level, including cross-border, it would result in real amounts of big data. This fact has to be addressed for scalability and replicability analysis.

2.1.3 METHODOLOGY AND APPROACH

2.1.3.1 OVERVIEW

Technical requirements identified apply to the following systems to be demonstrated in WP9:

- DEP (data exchange platform) – Estfeed platform
- Data hub – Elering’s data hub for electricity meter data
- Flexibility platform – to be developed in Task 9.2
- Aggregator’s application – “Affordable Tool” to be developed in Task 9.1
- System operators’ application to exchange data with flexibility platform – to be developed in Task 9.2
- Baseline calculation tool to be developed in Task 5.3 (integrated with DEP in Task 9.3)
- big data tool to be developed in Task 5.3 (integrated with DEP in Task 9.3)

Categories addressed to describe data related to requirements:

- Volume of data to be collected by period
- Volume of data to be processed by period
- Type of processing of data (e.g., prediction, reformatting, anonymization)
- Type of data (e.g., structured, semi-structured, unstructured data, times series, streaming, sequence, graph, spatial)
- Accuracy (is it necessary to complete, filter, transform, to correct the data)

Additionally, requirements were categorized based on the following types:

- Performance
- big data
- Personal data
- Security
- Functional

2.1.4 RESULTS AND CONCLUSIONS

2.1.4.1 RESULTS

Table 2.1 summarizes the results of the analysis presenting the identified big data requirements for all system use cases from Task 5.2. Few examples of the requirements describing the data volumes involved follow. Data description indicates what could be demonstrated in WP9 based on the needs of partners involved in concerned demonstrators (but does not mean that these volumes would be tested).

Ability to share access permissions:

- Nature of data considered – natural and legal persons of Estonia, Lithuania and Norway who want to give consent to their meter or sub-meter data access by other parties

- Volume of data to be collected – each time when it is necessary to authorize the user – millions of users, thousands of access permissions per day
- Volume of data to be processed – thousands of access permission per day

Collection of near-real-time meter data (up to 1 hour):

- Nature of data considered – hourly meter data from all Estonian metering points
- Volume of data to be collected – 20 million hourly values per day. One message containing 24 hourly values for one metering point = 3kB

Transfer of data, data owner's and application's access to data through DEP:

- Nature of data considered – parties and systems involved in data exchanges involve data owners, data hub of Elering, flexibility platform, aggregator's tool, system operator's flexibility application, Transparency Platform of ENTSO-E, baseline calculation tool, big data tool
- Volume of data to be processed – thousands of values per second, millions of values per minute; a message with grid (outage) data depending on the number of values may be 5-50 kB; a message with meter data containing 24 hourly values for one metering point is 3kB

Ability of DEP to forward real-time data from DER's to System Operators:

- Nature of data considered – exchange of some real-time data between aggregator's tool (and customers linked to this aggregator) on one hand and SCADA systems of system operator (in demonstrator different system could be used instead of SCADA) on the other hand
- Volume of data to be collected – hundreds of real-time values
- Volume of data to be processed – hundreds of values exchanged in less than 1 second

Exchange of flexibility activation requests through DEP and flexibility platform:

- Nature of data considered – flexibility activation requests delivered by system operator's application to flexibility platform and forwarded by flexibility platform to FSPs
- Volume of data to be collected – hundreds of values per minute
- Volume of data to be processed – hundreds of values exchanged in less than 1 minute (high-speed products have to be activated as the response to the frequency deviations in the grid; but otherwise, for slower products the activation request can be sent via DEP)

Flexibility platform's ability to collect bids from FSPs (through DEP):

- Nature of data considered – flexibility bids submitted by FSPs to flexibility platform
- Volume of data to be collected – few values per minute/hour, size of the bid – 120 kB
- Volume of data to be processed – few values exchanged in less than 1 hour, size of the bid – 120 kB

Collection of data from sub-meters:

- Nature of data considered – data collected by aggregator's tool

- Volume of data to be collected – hundreds of values per second
- Volume of data to be processed – ~1 kb/sec per meter

Storing sub-meter data in a data hub:

- Nature of data considered – data stored by aggregator's tool
- Volume of data to be collected – hundreds of values per second
- Volume of data to be processed – ~2 MB/day/meter

TABLE 2.1 BIG DATA REQUIREMENTS IN SYSTEM USE CASES

SUCs / Requirements	
SUC: Aggregate energy data	
1	Data source (e.g. meter data hub) ability to aggregate data
2	DEP ability to forward aggregated data from a data source to a data user
SUC: Anonymize energy data	
3	Data source (e.g. meter data hub) ability to anonymize data
4	DEP ability to forward anonymized data from a data source to a data user
SUC: Authenticate data users	
5	Ability to share information related to representation rights between data users and concerned Customer Portals
6	Ability to share authentication information between data users, Customer Portal and Authentication Service Provider
SUC: Manage access permissions	
7	Ability to share access permissions between data owners, concerned DEPs, applications and data sources
SUC: Collect energy data	
	Collection of meter data
8	<ul style="list-style-type: none"> Get near-real-time data (up to 1 hour) from meters
9	<ul style="list-style-type: none"> Get historical data (monthly) from conventional meters
10	<ul style="list-style-type: none"> Store data in a meter data hub
	Collection of market data
11	<ul style="list-style-type: none"> Get near-real-time (up to 1 hour) data from market
12	<ul style="list-style-type: none"> Get historical data from market
13	<ul style="list-style-type: none"> Store data in a market data hub
	Collection of grid data
14	<ul style="list-style-type: none"> Get very-near-real-time (up to 1 minute) data from grid
15	<ul style="list-style-type: none"> Get near-real-time (up to 1 hour) data from grid
16	<ul style="list-style-type: none"> Get historical data from grid
17	<ul style="list-style-type: none"> Store data in a grid data hub
SUC: Transfer energy data	
18	Transfer of data must be secured, through encryption or communication protocol
19	Data owner's access to data through DEP (and foreign DEP)
20	Application's access to data through DEP (and foreign DEP)
SUC: Exchange data between DER and SCADA	
21	Ability of DEP to forward real-time data from DER's to System Operators
22	Ability of DEP to forward very-near-real-time (up to 1 minute) data from DER's to System Operators
23	Ability of DEP to forward near-real-time (up to 1 hour) data from DER's to System Operators
24	Ability of DEP to forward activation requests from System Operators to DER
25	Encrypted data exchange
SUC: Manage flexibility activations	
26	Exchange of activation requests through DEP and flexibility platform

SUC: Calculate flexibility baseline	
27	Ability of flexibility platform to collect input for baseline calculation, incl. through DEP
28	Ability of flexibility platform to compute baseline
SUC: Manage flexibility bids	
29	Ability to exchange information on System Operators' flexibility need and FSPs' flexibility potential through flexibility platform (and DEP)
30	Algorithm for prequalification of flexibility providers
31	Flexibility platform's ability to collect bids from FSPs
32	Selection of successful bids
33	Calculation of grid impacts (congestion, imbalance)
SUC: Predict flexibility availability	
34	Collection of data for prediction (long term - years)
35	Computation of predictions (long term - years)
36	Collection of data for prediction (medium-term - days to years ahead)
37	Computation of predictions (medium-term - days to years ahead)
38	Collection of data for prediction (short term - intraday operation)
39	Computation of predictions (long term - intraday operation)
SUC: Verify and settle activated flexibilities	
40	Calculation of actually delivered flexibility as a response to an activation request
41	Verification that flexibility delivered matches with flexibility requested
SUC: Provide a list of suppliers and ESCOs – no big data requirements identified	
SUC: Erase and rectify personal data – no big data requirements identified	
SUC: Manage data logs	
42	Ability to share information related to data logs between data owners, concerned DEPs, applications and data sources
SUC: Manage sub-meter data	
43	Collection of data from sub-meters
44	Storing sub-meter data in a data hub
45	Ability of DEP to forward sub-meter data from data hub to customer (data owner) and application (energy service provider)
46	Ability of DEP to forward activation orders from a customer (data owner) or application (energy service provider) to devices
47	Data format of sub-metering
48	Transmission protocols of sub-metering

2.1.4.2 CONCLUSIONS

Around 70 technical requirements were identified based on the described system use cases, out of which 48 relate to big data. Annex II – Identification of technical requirements presents a detailed description of all these requirements. The biggest number of big data requirements related to use cases on data collection, sub-meter

data management, flexibility activation, flexibility prequalification and bidding, and DER-SCADA data exchange. Use cases on personal data and listing of suppliers and ESCOs involve no big data requirements. Figure 2.1 summarizes the big data requirements use case by use case.

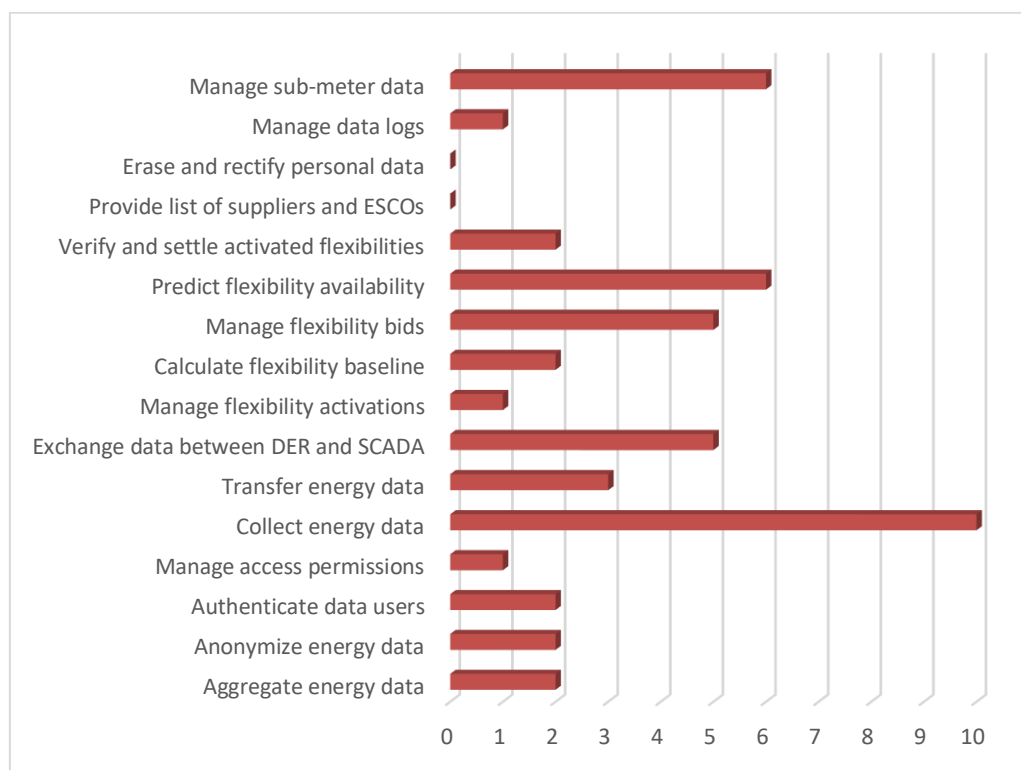


FIGURE 2.1 NUMBER OF IDENTIFIED BIG DATA REQUIREMENTS PER EACH USE CASE

2.2 COMPARATIVE STUDY OF EXISTING SOLUTIONS

Main section author: Grzegorz Gucwa (PSE)

2.2.1 ABSTRACT

Designing and implementing systems for data exchange can be carried out from scratch, it can also use the ideas, experiences and solutions implemented in the existing systems. The comparative study aims to examine the existing solutions in terms of meeting the requirements for data exchanges described in Chapter 2.1 on Identification of technical requirements.

Four different solutions implemented in the EU, Europe and the USA were selected for the comparative analysis. These solutions were evaluated in terms of meeting each the defined data exchange requirements. The results of the evaluation were aggregated and presented in the form of tables and graphs illustrating the degree of compliance of the solutions with the groups of requirements.

As a result of the analysis, it was found that the groups of requirements most strongly supported by the existing solutions are: handling and exchange of meter data, the data security and privacy. Most of the gaps were found in the areas of flexibility services (flexibility bids, baseline calculations, prediction, activation and verification) and real-time or near-real-time communication required for the activation of flexibility services.

Analysis of existing solutions confirmed that new needs related to data exchange solutions appeared. The implementation of flexibility services requires extending the functionality of existing platforms or building new ones.

2.2.2 INTRODUCTION

2.2.2.1 AIM

Comparative analysis aims to analyse existing solutions using technical requirements for the data exchange. The primary goals of the analysis are:

- investigation of what options already exist, how other actors or other industries have already started to address similar data exchange needs;
- identification which existing solutions meet requirements and which are not;
- determination of the suitability of existing solutions for the identified requirements.

2.2.2.2 CONTEXT

In task Identification of technical requirements about 70 technical requirements were identified and described. These requirements relate to data exchange, storage and processing volumes, time limits, security and privacy.

Identified technical requirements in Chapter 2.1 were taken for comparison criteria. A comparative analysis of four selected solutions existing in the energy market was carried out.

The following solutions were selected for comparative analysis:

- ENTSO-E OPDE,
- Estonian DEP,
- Green Button,
- Norwegian Elhub.

A description of the solutions selected for analysis can be found in the Annex III – Comparative study of existing solutions: detailed estimation of selected solution.

2.2.3 METHODOLOGY AND APPROACH

2.2.3.1 OVERVIEW

The procedure for comparative analysis of existing solutions and required actions are presented in the Figure 2.2.

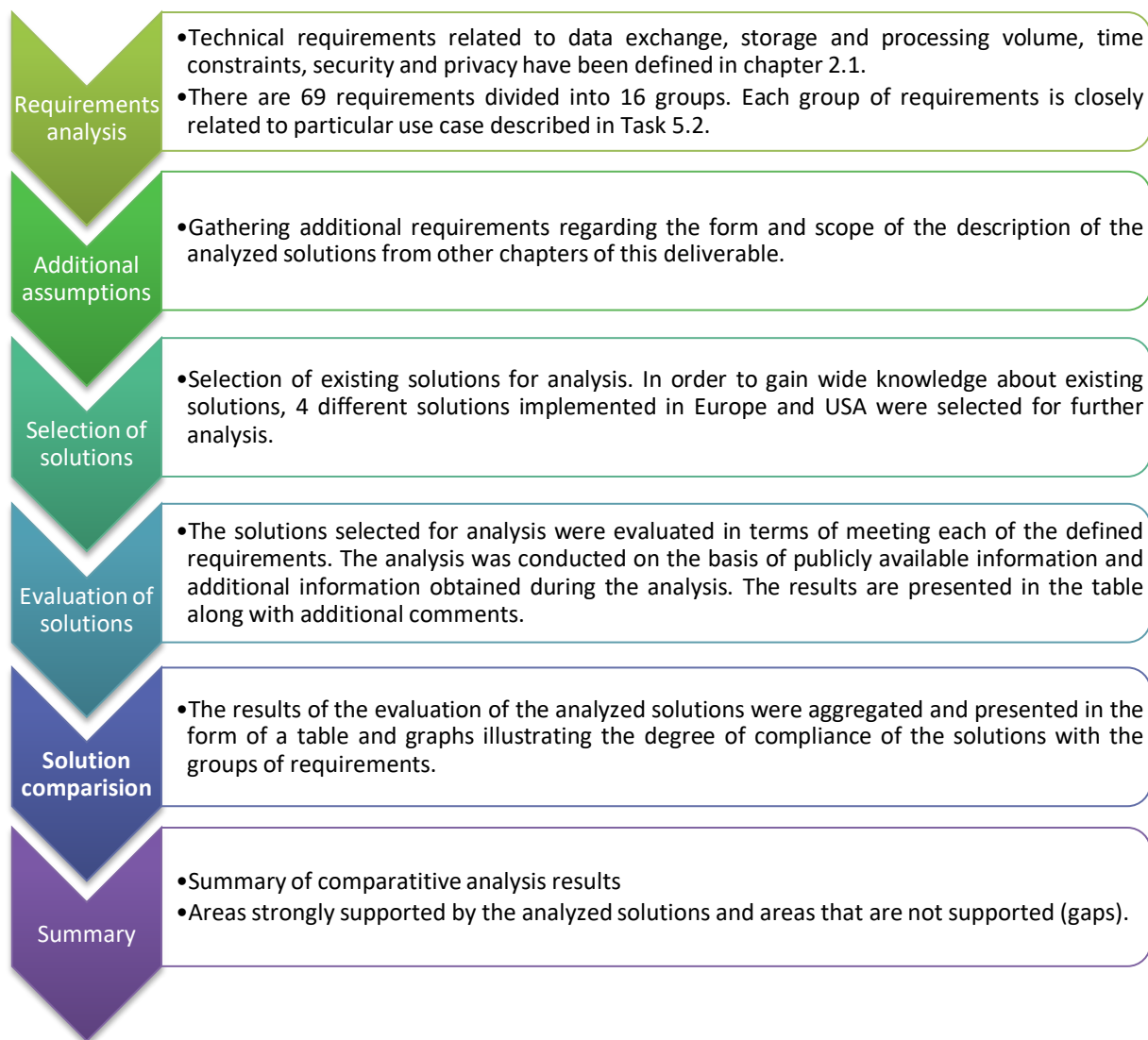


FIGURE 2.2 COMPARATIVE ANALYSIS PROCEDURE

2.2.4 RESULTS AND CONCLUSIONS

2.2.4.1 RESULTS

Table 2.2 summarizes the results of the analysis of existing solutions in terms of meeting the requirements for data exchange. The results are aggregated into groups of requirements which correspond to the use cases defined in Task 5.2. Therefore the shown rating per solution per use case (requirement group) is the sum of all requirements in this group. The detailed results how the solutions are rated for every single requirement are found in Annex III – Comparative study of existing solutions: detailed estimation of selected solution.

TABLE 2.2 RESULTS OF EVALUATION OF EXISTING SOLUTIONS BY GROUP OF REQUIREMENTS

Requirements group				Rating			
ID	Name	No. of req. in group	Max. rating ³	ENTSO-E OPDE	Estonian DEP	Green Button	Norwegian Elhub
AGG-ED	SUC: Aggregate energy data	4	12	3	12	4	12
ANO-ED	SUC: Anonymize energy data	4	12	3	9	7	0
AUTH	SUC: Authentication of data users	4	12	1	12	8	12
AUTHZN	SUC: Access permission management	3	9	1	9	9	9
DC	SUC: Data collection	10	30	0	9	6	6
DT	SUC: Data transfer	4	12	4	12	12	10
DER-SCADA	SUC: DER-SCADA data exchange	6 ⁴	18	8	6	0	0
FA	SUC: Flexibility activation	2	6	1	2	0	0
FB	SUC: Flexibility baseline	2	6	1	1	0	0
FBIDS	SUC: Flexibility bids	9	27	7	7	0	0
FPRED	SUC: Flexibility prediction	6	18	3	3	0	0
FVERIF	SUC: Flexibility verification	3	9	0	2	0	0
ESCO	SUC: List of suppliers and ESCOs	1	3	1	3	3	3
PERSO-DATA	SUC: Personal data	2	6	2	6	1	6
LOGS	SUC: Data logs	1	3	1	3	1	3
SUBMET	SUC: Sub-meter data	7	21	4	3	13	0
Total		68	204	40	99	64	61

Solution rating for each group of requirements is calculated as the sum of assessments for individual requirements. The requirements are assessed according to the scale in Table 2.3.

³ The maximum rating for each requirement group is calculated as follows: [number of requirements in the group] x [maximum score of a single requirement].

⁴ One of the requirements in the DER-SCADA group (DER-SCADA-REQ3) was omitted in the calculations because it is out of the scope of the analysis. See ANNEX III for details.

TABLE 2.3 SCALE OF ASSESSMENT OF REQUIREMENTS

Rating	Description
N/A	Not analysed. Requirement applies to components out of scope of analysis.
0	The analysed solution is not designed for this kind of requirement
1	The analysed solution does not meet the requirement, but some of its functionality can be used to meet this kind of requirements
2	The solution partially meets the requirement (no more than 75%)
3	The solution meets the requirement (75% and more)

Due to the scale used to evaluate the solution, the maximum score means that analysed solution meets the requirements between 75% and 100% (not always 100%).

Table 2.4 summarizes the normalized results of the analysis. The results are normalized by calculating the average value of the results in each group.

TABLE 2.4 NORMALIZED RESULTS OF EVALUATION OF EXISTING SOLUTIONS BY GROUP OF REQUIREMENTS

Requirements group		Rating (normalized)			
ID	Name	ENTSO-E OPDE	Estonian DEP	Green Button	Norwegian Elhub
AGG-ED	SUC: Aggregate energy data	0,75	3,00	1,00	3,00
ANO-ED	SUC: Anonymize energy data	0,75	2,25	1,75	0,00
AUTH	SUC: Authentication of data users	0,25	3,00	2,00	3,00
AUTHZN	SUC: Access permission management	0,33	3,00	3,00	3,00
DC	SUC: Data collection	0,00	0,90	0,60	0,60
DT	SUC: Data transfer	1,00	3,00	3,00	2,50
DER-SCADA	SUC: DER-SCADA data exchange	1,33	1,00	0,00	0,00
FA	SUC: Flexibility activation	0,50	1,00	0,00	0,00
FB	SUC: Flexibility baseline	0,50	0,50	0,00	0,00
FBIDS	SUC: Flexibility bids	0,78	0,78	0,00	0,00
FPRED	SUC: Flexibility prediction	0,50	0,50	0,00	0,00
FVERIF	SUC: Flexibility verification	0,00	0,67	0,00	0,00
ESCO	SUC: List of suppliers and ESCOs	1,00	3,00	3,00	3,00
PERSO-DATA	SUC: Personal data	1,00	3,00	0,50	3,00
LOGS	SUC: Data logs	1,00	3,00	1,00	3,00
SUBMET	SUC: Sub-meter data	0,57	0,43	1,86	0,00
Total		0,64	1,81	1,11	1,32

Figures 2.3-2.5 contain graphs that are a representation of the normalized results of comparative analysis. Detailed information on the compliance of the analysed solutions with the defined requirements can be found in Annex III – Comparative study of existing solutions: detailed estimation of selected solution.

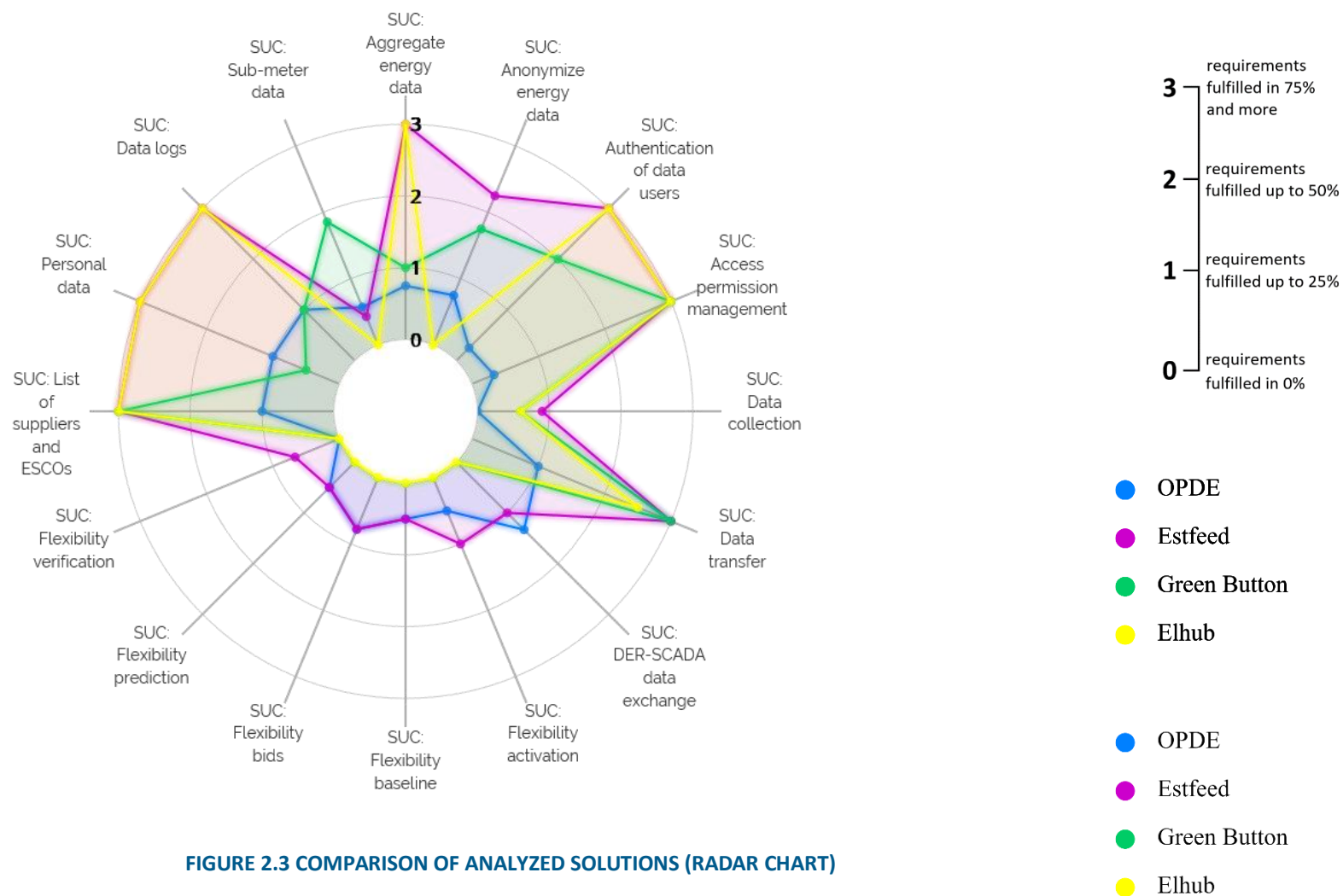


FIGURE 2.3 COMPARISON OF ANALYZED SOLUTIONS (RADAR CHART)

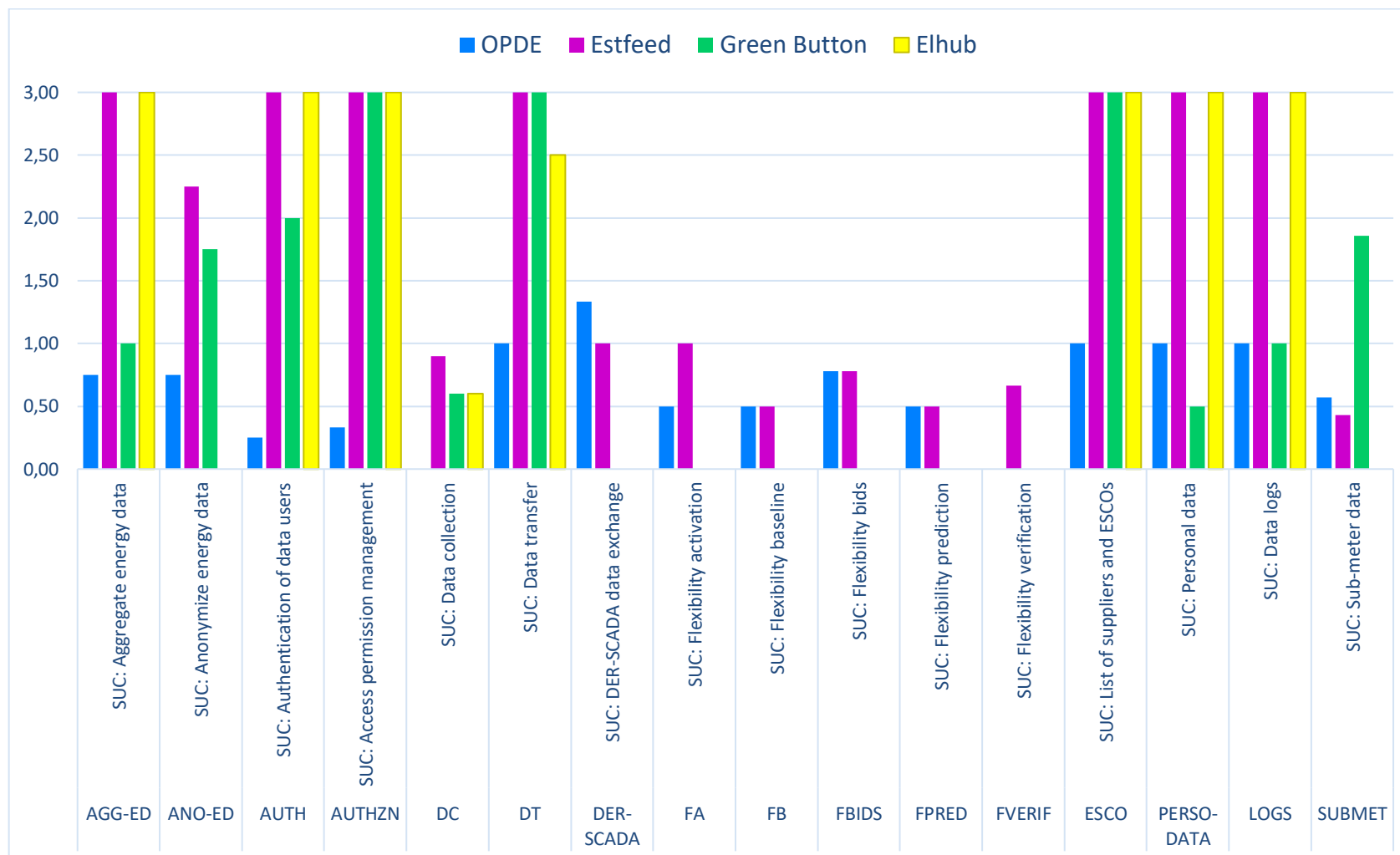
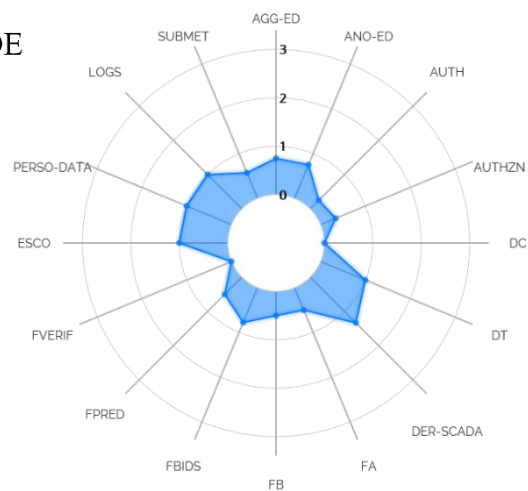
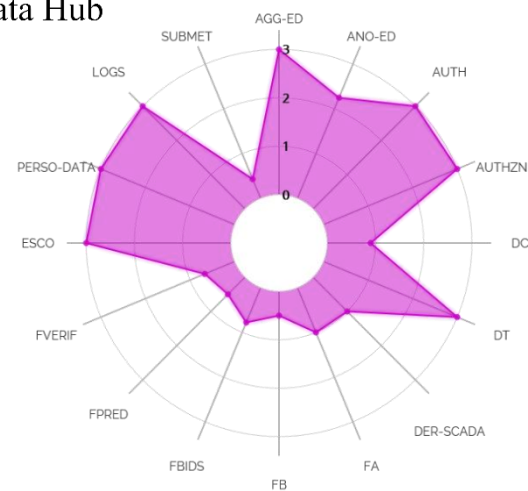


FIGURE 2.4 COMPARISON OF ANALYZED SOLUTIONS (BAR GRAPH)

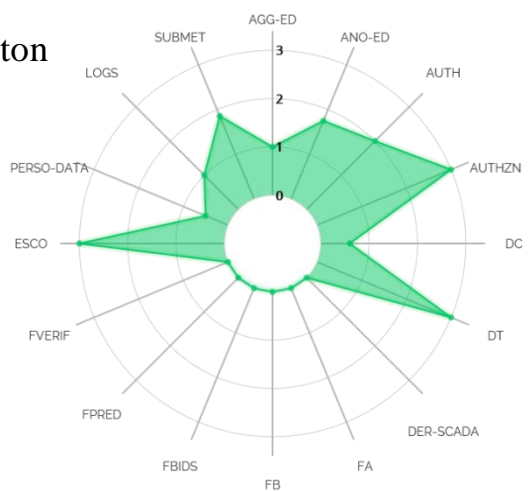
OPDE



Estfeed + Data Hub



Green Button



Elhub

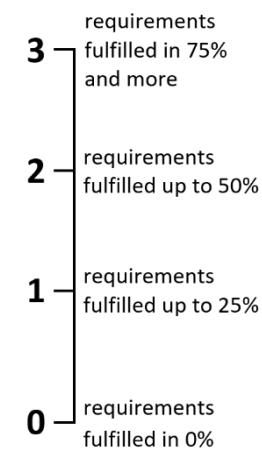
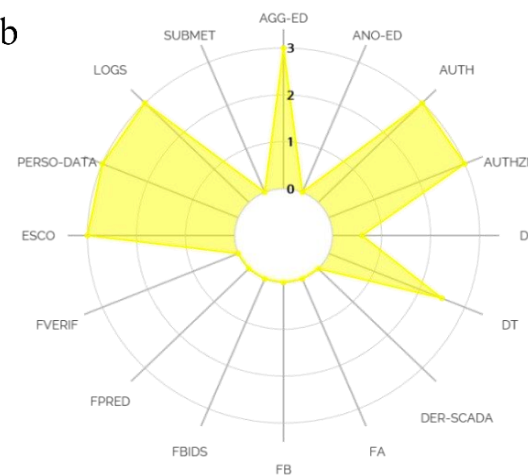


FIGURE 2.5 RESULTS OF EVALUATION OF EXISTING SOLUTIONS BY GROUP OF REQUIREMENTS (RADAR CHARTS)

2.2.4.2 CONCLUSIONS

All the analysed solutions were created to fulfil the requirements specified for the purpose for which they were built. These goals and requirements differ from the ones defined in the EU-SysFlex project, and therefore the analysed solutions implement only part of the EU-SysFlex requirements.

Analysed solutions strongly support the areas of handling and exchanging meter data (aggregation, anonymization). All solutions strongly care for security. The area of authentication, user rights management and personal data (GDPR compliance) is served very well, too.

Most gaps are in the area of flexibility services. None of the analysed solutions supports functionalities related to this area (flexibility bids, baseline calculations, prediction, activation and verification). Although some of them can be used to implement a layer of secure communication that is the basis for the implementation of flexibility services management, they do not support flexibility business functions explicitly.

All the analysed solutions support the transfer and storage of meter data. However, data from sub-meters is supported only by Green Button. Neither of analysed solution support Grid-data and market-data.

The analysed solutions do not currently support real-time or very-near-real-time (up to 1 minute) communication required for the activation of flexibility services in real-time. It is probably technically possible to use the Estfeed platform for such communication, but currently, there are no confirmed implementations in this area. There are also no confirmed cases of using analysed solutions for communication with the SCADA system.

The EU-SysFlex project as an innovation project shows new needs to be considered. None of the analysed existing solutions meets all the data exchange requirements derived from use cases defined under Task 5.2, because they were not built for this. The implementation of the Clean Energy Package, which results in the need for a different approach to obtaining flexibility services, requires extending the functionality of existing platforms or building new ones. Pilot projects to support flexibility services are already underway, e.g. Piclo Flex, NODES, GOPACS, ETPA, and EPEX SPOT. However, it is noteworthy that these are market platforms to address specific business processes, rather than data platforms agnostic to business processes. Further research may include mentioned solutions, as well as solutions implemented in other industries (e.g. telecommunications, banking).

3. COST OF DATA EXCHANGE FOR ENERGY SERVICE PROVIDERS

Main section authors: Simon Siöstedt (AFRY), Steve Wattam (Upside), Philippe Szczech (AKKA)

3.1 ABSTRACT

This section addresses the cost of data exchange for energy service providers, specifically considering an aggregator that provides flexibility.

The integration of large amounts of renewable energy sources to the European grid requires more flexibility, both for the generation and the consumption. With the introduction of “smart devices,” it has been possible to enable two-way communication with all types of grid-connected devices, which in turns has the potential to provide flexibility services to the grid. The questions are, what kind of big data architecture is needed to handle this data communication, and what would the cost be.

This chapter has investigated the latter of the two questions by analysing the cost of data exchange in the case of an aggregator that plays a role between the DSO and a market platform that provides flexibility.

The method that has been used considered the option of a public cloud-based deployment and referred to a Microsoft Azure IaaS solution and exclusively refers to the remote provisioning of 'raw' IT resources. The specific use case for the aggregator considered 100 000 devices located in the same region with the same service types but one case with 10 TB and one case with 1000 TB storage capacity.

The results from the cost assessment showed that the cost for the public cloud-based deployment with 10 TB of storage capacity would be 5 700 EUR/month where the storage capacity itself only accounted for 5% of the total cost. On the other hand, the cost with the same data architecture and with the same services but with 1000 TB of storage would cost 23 400 EUR/month where the storage capacity itself accounted for 77% of the total cost.

The cost distributions between the two cases show that the service types excluding the storage capacity are dominant for a case with limited requirement for storage. In order to provide flexibility services as an aggregator, all of the service types in the data architecture are needed, therefore two main conclusions can be stated:

1. The start-up cost to provide flexible service is dominant over the storage capacity where efficient processing of data, load forecasting and high-throughput capacity is necessary to run the services rather than providing a large amount of storage capacity.
2. It is essential to determine which stakeholder should account for the storage capacity and how much storage is needed to run the services to not over-dimension the application and thereby increase the cost which in the end will lead to a higher cost for the end-user.

3.2 INTRODUCTION

3.2.1 BACKGROUND

The successful integration of large amounts of renewable energy sources (RES) in the European electricity grid requires the entire power system to handle increased electricity production variability. At present, the grid is still predominantly designed and constructed to distribute electricity efficiently from large conventional plants with stable baseload power production to all connected end-users. The future distributed and weather dependent power plants give rise to new challenges for both transmission and distribution grids and one of the main high-level solutions is to increase the flexibility both in the power production and on the end-user side.

New technologies on both generation and demand-side incorporating distributed devices that can communicate with each other have increased the possibility to introduce dynamic flexibility in the power system. The rollout of "smart meters" required for high granularity measurement and control of such dynamic flexibility services has been performed in most of the European countries where the consumer or the appliance itself can now see the own energy consumption in near-real-time. It is possible to combine this information with market or other information and then, for example, act to reduce their average or peak consumption, to perform different adjustments as required for to optimize on an economic basis. The possibility to change the consumption patterns manually would increase the flexibility of the entire power system. To unleash the full potential of dynamic flexibility, devices need to be able to communicate and respond to price differences in the electricity market or signals from the grid companies in case of overload or other constraints. By enabling automated energy consumption control, it could both save money for the consumer and decrease the system cost for grid companies and as an overall benefit leads to the most efficient utilization patterns of all CAPEX intensive transmission and distribution infrastructure assets. Introducing communication devices to enhance the power system flexibility provides many new opportunities but several aspects need to be analysed and clarified to provide a complete picture.

3.2.2 AIM

For a large scale rollout of communication devices to provide flexibility to the power system, it is essential to analyse the required data architecture and its specific cost drivers related to the data volumes in the future. The two questions below address this concern:

- 1) What type of unstructured data is needed to enable data communication for flexibility providers to apply across the entire power system and between countries?
- 2) How much data would be generated for these applications, and what would be the cost of this data exchange?

Question 1 is analysed in Chapter 1 on Big data framework which provides the basis for current chapter, while the aim here is to answer question 2 by analysing the cost of data exchange in the case of an aggregator that plays a role between the DSO and a market platform that provides flexibility.

3.2.3 CONTEXT

Digitalisation: A new era in energy? This question was stated by the international energy agency (IEA) 2017 and today the question mark should be replaced with an exclamation mark! The digitalisation of the energy sector in general and for the electricity sector in particular, opens new opportunities where the integration of more renewable generation to the electricity grid is one of the most important work of our time to comply with the Paris agreement and reduce the carbon footprint. Increased consumption and generation flexibility is important when increasing intermittent electricity generation to the electricity grid and the digitalisation will play a major role to enable this flexibility. New IoT-solutions could for an instance enable two way communication and respond to challenges at any level in the grid and by any type of device which then has the potential to be a flexible resource rather than just static load. By aggregating the response from the two way communication devices, it is possible to optimise the flexibility resources which opens new business opportunities. Third parties with a business model to aggregate, analyse and use the data in a coordinated way to provide flexibility services will increase in the future, therefore this type of stakeholder is extra interested to understand data architecture and cost drivers.

In other EU-SysFlex work and deliverables, incl. in Chapter 4.2 on Prediction of availabilities and quantities of flexibility services where the use case of an aggregator is derived from, the various data flows required for predicting and using flexibility services has been assessed. Identified timescales over which flexibility may be predicted/managed are the following:

1. Investment — 3 years plus
2. Operational Planning — from days to years
3. Real-time — intraday

Each of these problems has a particular data access pattern that significantly impacts the cost of IT systems such as storage or software as a service (SaaS) provision, for example, lower latencies are usually required for systems that access operational planning or real-time data, compared to those used for investment decisions.

This work identified the data requirements of day-to-day operation for an aggregator providing flexibility services to a TSO/DSO, working from two scenarios (investment timeframe and operational timeframe) drawn from Task 5.2 respective SUC (system use case) description. These two scenarios cover the vast majority of data access patterns (as the real-time demands closely follow daily optimization/prediction patterns). They are slightly modified from the scenarios compared to the original ones in flexibility prediction SUC, by assuming that the aggregators' current operations are described by the abstract system operator, i.e. TSO/DSO.

3.2.3.1 SCENARIO 1 - DSO PREDICTING FLEXIBILITY AVAILABILITY FOR INVESTMENT PLANNING

In this scenario, the business (labelled as the DSO) is primarily concerned with longer-term trends in generation and system (hence market) volatility. These trends are derived from expert opinion and careful examination of large volumes of observed data, describing long timescales.

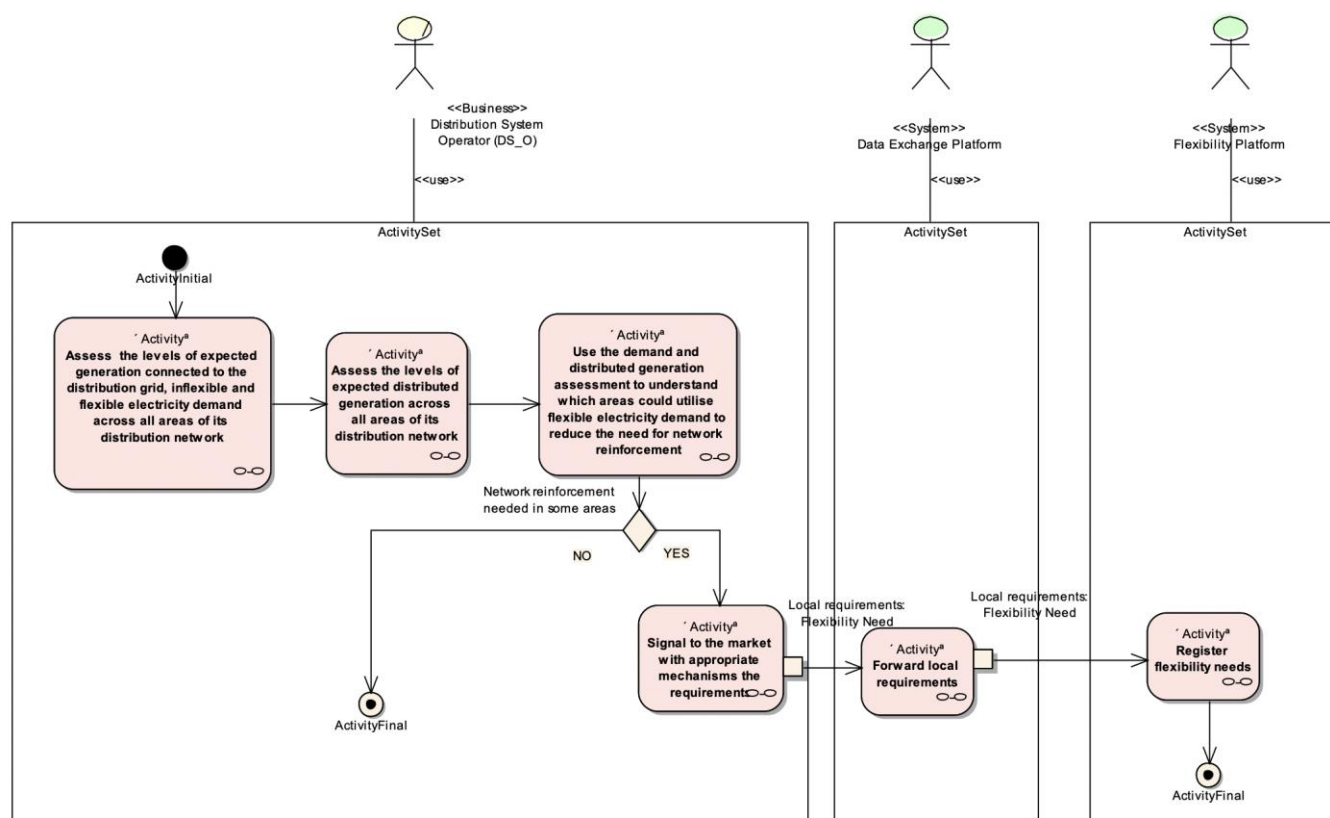


FIGURE 3.1 DSO PREDICTING FLEXIBILITY IN THE LONG TERM (INVESTMENT) TIMESCALE

3.2.3.2 SCENARIO 2 - SYSTEM OPERATOR PREDICTING FLEXIBILITY AVAILABILITY FOR OPERATIONAL PLANNING

This more immediate scenario requires examination of immediately available flexibility. This problem is much more concrete, as many participants in the system have committed to offering given volumes and others have reported their physical state (e.g. state of charge of assets, maintenance windows). This scenario is, therefore, much more a problem of synthesizing these data to perform useful predictions over the coming day(s).

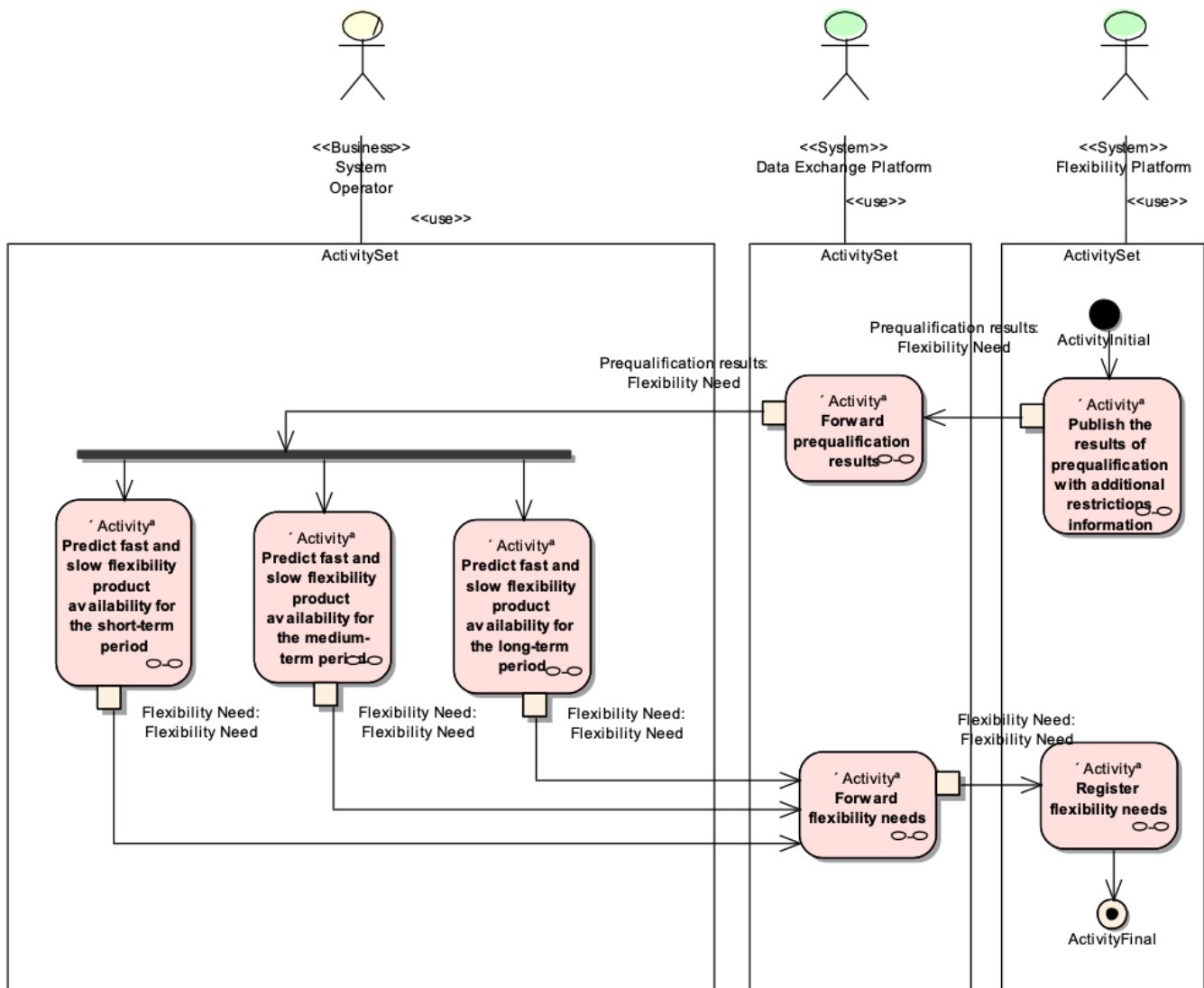


FIGURE 3.2 SYSTEM OPERATOR PREDICTING FLEXIBILITY IN MEDIUM TERM (OPERATIONAL) TIMESCALE

3.2.4 THE ROLE OF AGGREGATOR AND DATA SOURCES

The aggregator is assumed to play as a part-way between the flexibility platform and the DSO in the scenarios described above. The business model is built upon providing a platform to optimize and dispatch flexible assets, and this means that the aggregator is often responsible for the prediction task, as well as curating the data sources required to execute this regularly. The "data broker" role in the figure ('Data Exchange Platform') effectively does not exist in the current electricity system, and as such the data are stored and handed between the aggregator (as flexibility provider) and asset owners who must report to the DSO or TSO.

Essentially, the lack of a data broker today reduces the "Data Exchange Platform" actions to application programming interface (API) without secure handling of private data and lack of data hubs for separate data storing facility which both belong either to the customer or the aggregator.

As described earlier, the aggregator is processing data sources and data volumes in three different timescales, long-term data (grid planning and investments), mid-term (operational data) and short-term (intraday and real-time). The following data sub-categories are assumed for each of the time scale categories:

1. Long-term data stores
 - a. Market data
 - b. Service and Control Logs
2. Mid-term data stores
 - a. Device Schedules and Configuration
 - b. Price, Site Load Forecasts
3. Short-term data stores
 - a. Device/Site Telemetry
 - b. Control Instructions

3.3 METHODOLOGY AND APPROACH

This section describes a cost estimation of a big data solution architecture which could be used to implement the previously described scenarios. This big data solution is based on the reference architecture designed in the 'Big data framework (Chapter 1).

The methodology considered the option of a public cloud-based deployment and referred to a Microsoft Azure IaaS solution as an example. The costing introduced here refers exclusively to the remote provisioning of 'raw' IT resources such as virtual servers, software programs, storage devices and outbound/inbound network interfaces. It does not include the labour cost of IT administrators to implement and manage the solution in the IaaS model, the customer being responsible for configuring, managing and provisioning the resources.

The main factors which have affected the cost are:

- **The selected service.** Each cloud provider has its portfolio of services, including different options. Each cloud provider also has a set of price schemes for these services.
- **The characteristics of chosen instances.** An instance is a virtual machine which performs some tasks in the cloud. Often the terms *VM*, *Virtual Machine* and *Instance* are used indifferently. VMs are roughly grouped into those categories:
 - *Standard* instances that are used for addressing most of the use cases such as web applications, medium-database. They provide a good trade-off between CPU power and memory.
 - *High CPU* instances that are used when a lot of processing power is requested, like in cases of batch processing or data analytics.
 - *High memory* instances that are used for processing memory-intensive tasks such as near or strict real-time data ingestion or big data analytics.

- *GPU* instances (Graphics processing unit) that provide intensive processing capability and are used, for example, in cases of deep learning algorithms or scientific simulations.
- **Data centre geographical localization.** When a customer orders a cloud-based resource, the location of the data centre where the resource will be provisioned has to be selected. This choice could depend on customer requirements such as latency performance, regulatory constraints. From a cost perspective, selecting different data centres from the same cloud provider could lead to a different price.
- **Resource utilization.** The resources utilization rate is another critical factor which affects the cost. When a computing resource is used, the customer will be invoiced based on the duration of the usage. When a storage resource is used, he will be invoiced based on the data volume stored.
- **Support.** In general, different levels of support are available in the portfolios of cloud providers with different prices. A strategic application does not require the same level of support as a test application. Therefore, the level of support must be decided accordingly, and this choice will impact the price.

The estimated cost comes from the Azure tool simulation provided by Microsoft (Azure Microsoft, u.d.), in which parameters and hypothesis were entered deduced from the analysis of the scenarios. The price depends directly on the design and the sizing tasks of the chosen solution.

3.3.1 EXPLANATION OF THE AZURE CLOUD

Hereafter, some explanation about some concepts or wording used in the results section of Table 3.1:

Managed disks: These are the storage disks used by the Azure VMs. The term 'managed' refers to the fact they are managed by Azure which simplifies their use for the customer, for example, these disks can be easily resizable, attached to a VM, and so forth. They come in different characteristics: size, unit (SSD or HDD), throughput, etc. which influence their prices.

Pay-as-you-go vs. 1-year reserved option pricing: Pay-as-you-go is the payment option when the customers are billed on the actual resource usage and uptime. In this case, customers have not committed the usage of the resource. Such pricing is the opposite of the reserved resource option, which is an upfront commitment offering in return price reductions. In principle, this option is less expensive in case of 24/7 usage resource or processing a massive amount of data. In the cost assessment, the choice between those two options has been made in function of these both criteria.

VM service: This service enables the provisioning of a VM. The VM differs in processing/CPU power, memory, and storage capacity. They are mentioned in the cost table with this kind of names: *A3*, *D13V2*, *D12V2*, *D4V2*, *D4s*.

HDInsight service: In Azure, HDInsight service enables the usage and the management of popular open-source frameworks for data analytics such as Hadoop, Apache Spark, Apache Hive, Apache Kafka, Storm. When an HDInsight is provisioned, the customer can select only one of the previous frameworks for a given cluster, in other words, if a customer needs two of those frameworks, it is necessary to provide two different clusters. It should be

noted that Microsoft does not currently support this service in all its data centres, and especially it is not available in the one located in the United Kingdom. Therefore the "Northern Europe region" was selected while developing the cost estimate.

Bandwidth: Microsoft charges customer only for the outbound data transfers, meaning the data moving from the cloud to the external world, while the opposite flows (inbound data transfers) are not charged. The outbound traffic can have a strong impact on the final cost, for example, a monthly data transfer of 1 TB is charged approximatively 74 euros, but it costs 24 000 euros to transfer 500 TB. Besides, it seems that in Azure is not possible to exceed 500 TB per month for the outbound data traffic.

Storage account: Data Lake Storage Gen2 is a data storage solution offered by Microsoft to build a data lake which is accessible through an HDFS-API compatible. This storage capacity is scalable, meaning that it can provide up to several exabytes with a throughput measured in gigabits per second.

3.3.2 DETAILS OF THE SOLUTION

Hereafter, some explanation of the different clusters in the big data architecture used in the results section and illustrated in Figure 3.3 follows.

3.3.2.1 CASSANDRA CLUSTER

In the proposed architecture, a Cassandra database is used to create the mid-term data store containing "device schedule and configuration", "price, site load forecast" data of the use case. Cassandra is a distributed and fault-tolerant system which requests a cluster of machines. For this reason, Azure VMs was selected which are considered as Memory-optimized ones with 1-TiB disks providing the highest possible combined throughput and IOPS. This cluster should enable Cassandra to quickly respond to the requests coming from the external applications demanding low-latency access.

3.3.2.2 SPARK CLUSTER

Spark is chosen because of its fast engine for large-scale data processing to compute the price and site load forecasts, transform and manipulate massive amount of raw data, provides a solution for the later stages such as integration, machine learning and interactive querying. Moreover, Spark provides a Spark Streaming API for the near-real-time processing use cases.

In the solution, Spark is provisioned through an HDInsight cluster composed of the Apache Spark library and some VMs. Since Spark runs in-memory parallel processes, memory-optimized Linux VMs was selected. Currently, the Memory-optimized Linux VMs for Azure are D12 v2 or greater.

3.3.2.3 KAFKA CLUSTER

Kafka is the data ingestion component used to improve further the high-throughput capacity of the big data system between the IoT data sources and the long-term/real-time data stores of the use case. Kafka is a distributed and fault-tolerant broker (streaming platform) which can temporarily buffer massive data streams by replicating the ingested information all over the cluster nodes. Likewise Spark, Kafka is also delivered through an HDInsight cluster composed of the Apache Kafka library and some VMs. HDInsight allows user to change the number of worker nodes (Kafka-broker) after cluster creation.

3.3.2.4 AZURE DATA LAKE

Azure Data Lake storage solution is used to gather the aggregators' real-time data store, including device/site telemetry and control instructions as well as long-term data store containing market data and service/control logs.

This solution can be used as a Hadoop File System (HDFS) for collecting, managing and accessing real-time data transferred from the Kafka broker. The Azure Storage Data Lake is scalable and can store and serve up too many exabytes of data, ingested with a throughput measured in gigabits per second. In the cost assessment, there were two options for the initial size of this database (10 TB and 1 PB) to illustrate the price of the scalability.

3.3.2.5 DRILL CLUSTER

The Drill is the OLAP solution proposed for querying interactively data residing in the data Lake. It enables users to explore and analyse long-term as real-time data without sacrificing the flexibility and agility offered by these datastores.

Drill is a low latency distributed query engine for large-scale datasets. It is designed to scale to several nodes and query petabytes of data that business intelligence and data mining contexts might require. The solution included VMs to create with this capability of scalability and high-performance execution engine.

3.4 RESULTS

The costing evaluation introduced hereafter is related to a scenario of **100k devices** located in the same region (Northern Europe in this case).

Besides, in Figure 3.1 and Figure 3.2, the same data architecture has been used for the '100K devices' scenario:

- The outbound traffic (such as data transferred from the cloud to the devices, users, application): 6 TB every month
- The sum of data volumes contained in the long-term and real-time data stores: 1 PB
- The data volume in the mid-term data store: 1 TB
- The inbound traffic from the assets/devices: 6666 messages every second, equivalent to 70 kB/s

Eventually, it turns out that these metrics strongly impacted the final costing.

Figure 3.3 illustrates the service types for the big data architecture to enable the 100k devices scenario which has been used to assess the monthly cost of data exchange quantified in Table 3.1.

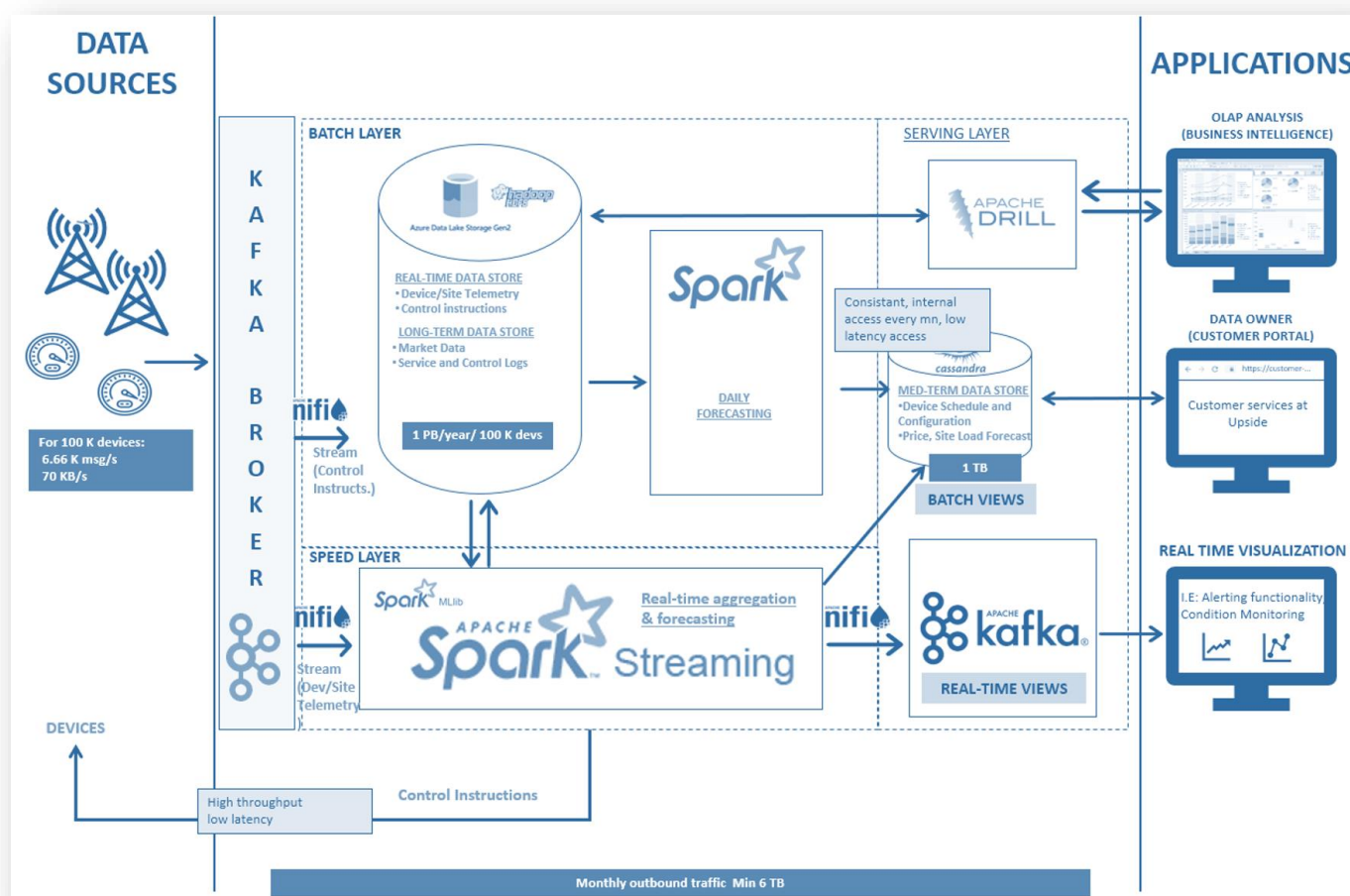


FIGURE 3.3 THE BIG DATA ARCHITECTURE TO ENABLE 100 000 DEVICES TO COMMUNICATE AND PROVIDE FLEXIBILITY SERVICES.

3.4.1 MONTHLY COST ASSESSMENT

In Table 3.1, the cost assessment per service and month are presented for a case with a small storage capability (10 TB) and a large storage capability (1 PB). The service types are illustrated in Figure 3.3 and described in the table.

TABLE 3.1 MONTHLY COST OF THE DIFFERENT SERVICE TYPES IN THE BIG DATA ARCHITECTURE FOR TWO CASES WITH DIFFERENT STORAGE CAPABILITY

Service type	Custom name	Region	Description	Option data lake long-term: 10 TB Estimated monthly cost	Option data lake long-term: 1 PB Estimated monthly cost
VMs	Cassandra database	North Europe	3 DS13 v2 (8 vCPU(s), 56 GB RAM); Linux – Ubuntu; 1 year reserved; 3 managed OS disks – P30	€1,185.26	€1,185.26
HDInsight	Spark real-time	North Europe	Spark Component: 2 A3 (4 cores, 7 GB RAM) Head nodes x 730 Hours, 4 D13V2 (8 cores, 56 GB RAM) Region nodes x 730 Hours, 0 D4V2 (8 cores, 28 GB RAM) Edge nodes x 730 Hours	€2,217.67	€2,217.67
HDInsight	Kafka broker	North Europe	Kafka Component: 2 A3 (4 cores, 7 GB RAM) Head nodes x 730 Hours, 4 D12V2 (4 cores, 28 GB RAM) Region nodes x 730 Hours, 3 A1 (1 cores, 1.75 GB RAM) Zookeeper nodes x 730 Hours, 0 D4V2 (8 cores, 28 GB RAM) Edge nodes x 730 Hours, 0 Standard disks	€1,438.49	€1,438.49
Storage Accounts	Data Lake storage long-term and real-time HDFS-API compatible	North Europe	Data Lake Storage Gen2, Standard, LRS Redundancy, Hot Access Tier, Flat Namespace File Structure, Capacity: See adjacent columns 'Option data lake capacity' - Pay-as-you-go Write operations: 4 MB x 10,000,000 operations, 100,000 List and Create Container Operations, Read operations: 4 MB x 100,000 operations, 100,000 Iterative write operations, 100,000 Other operations. 1,000 GB Data Retrieval, 1,000 GB Data Write	€274.80	€17,955.76
VMs	Drill	North Europe	2 D4s v3 (4 vCPU(s), 16 GB RAM); Linux – Ubuntu; 1 year reserved; 0 managed OS disks – E15, 100 transaction units	€172.91	€172.91
Data Transfers	Outbound traffic	North Europe	Zone 1: North America, Europe, 6 TB	€450.40	€450.40
Support			Support	€0.00	€0.00
			Licensing Program	Microsoft Online Services Agreement	
			Total (Monthly cost)	€5,739.54	€23,420.50

Table 3.1 shows that the cost for 100k devices would be 5 740 EUR/month with a storage capability of 10 TB and 23 420 EUR/month with a storage capability of 1 PB. The major cost difference means that the necessary storage has a significant impact on the cost. Apart from the storage, the monthly cost is the same for the other service types for both cases.

Table 3.2 presents the percentage of the cost for each of the service types, which further show how big impact the storage capability has on the total cost. For the 10 TB case, the cost of the storage only accounts for 5% of the total cost compared to 1 PB storage which accounts for 77% of the total cost.

TABLE 3.2 PERCENTAGE OF THE TOTAL MONTHLY COST FOR TWO CASES WITH DIFFERENT STORAGE CAPABILITY

Service type	Custom name	Option data lake long-term: 10 TB	Option data lake long-term: 1 PB
		% of total cost	% of total cost
VMs	Cassandra database	21%	5%
HDInsight	Spark real-time	39%	9%
HDInsight	Kafka broker	25%	6%
Storage Accounts	Data Lake storage long-term and real-time HDFS-API compatible	5%	77%
VMs	Drill	3%	1%
Data Transfers	Outbound traffic	8%	2%

3.5 DISCUSSION AND CONCLUSIONS

3.5.1 DISCUSSION

"The cost of data exchange" is a general term which has been approached by deriving the cost of data exchange for an aggregator that plays a role between the DSO and a market platform that provides flexibility. The cost has been assessed and presented on a monthly cost run base and considers the core data infrastructure for public cloud-based deployment and does not include administrative costs. A scenario of 100k devices communicating with the aggregator with different time scales for the data has been assessed together with a storage capacity of 10 TB and 1 PB (1000 TB).

The results show that the storage capacity has a material impact on the cost. The cost of the storage service account for a relatively shy 5% at 10 TB capacity but increase to a dominant 77% at 1 PB storage capacity. Storage capacity is a potent price driver and has a significant impact on the total price. Relevant to note is that the mentioned prices do not take in account the possible renegotiation with the cloud provider. In this case, the price curve could be different from the one found in the cloud provider service catalogue.

The storage capacity sensitive price relationship begs the question of what data and how much data needs to be stored. Today the aggregator in the use case is often responsible for the prediction task of flexibility assets, as well as curating the data sources but not to store and share data, and as such the data are stored and handed between the aggregator (as flexibility provider) and asset owners. They must report to the DSO or TSO. It means that the data are stored and that the cost should be accounted for but not necessarily burden the aggregator business case. Consider also that in a big data scenario with Hadoop technology, data are not merely stored, but they are also replicated across the cluster nodes to guarantee fault-tolerance. This fact should clarify why the storage seems to be a bit overestimated, to handle the replicas.

Another question revolves around the amount of data that is efficient to store. An increased monthly cost for service will, in the end, reach the end customer bill, which is unwanted if it counterweighs the benefits achieved

by the services/flexibility provided. Different flexibility services change the nature of the cost elements where a few services would require fast processor speed but limited storage capacity as well as the opposite. For an aggregator it is important to weight the cost of their applications with the benefits and the revenue stream to provide valuable services at an acceptable price for all parties.

3.5.2 CONCLUSIONS

The cost of data exchange for an aggregator that plays a role between the DSO and a flexibility platform has been assessed from the big data architecture developed in Chapter 1 on 'big data requirements for one case with a storage capacity of 10 TB and with a storage capacity of 1 PB. According to the cost estimation from Microsoft Azure IaaS solution, the total monthly cost with a storage capacity of 10 TB would be 5,739.54 euros and 23,420.50 euros with a storage capacity of 1 PB.

Table 3.3 presents the percentage cost for each service type for the two cases where the cost of storage in the 10 TB case only accounts for 5% of the total cost compared to 1 PB storage, which accounts for 77% of the total cost. For the case with only 10 TB of storage the Cassandra database, Spark real-time and Kafka broker account for 85% of the total cost compared to the 1 PB case where these service types only account for 20% of the total cost.

TABLE 3.3 PERCENTAGE OF THE TOTAL MONTHLY COST FOR TWO CASES WITH DIFFERENT STORAGE CAPABILITY

Service type	Custom name	Option data lake long-term: 10 TB	Option data lake long-term: 1 PB
		% of total cost	% of total cost
VMs	Cassandra database	21%	5%
HDInsight	Spark real-time	39%	9%
HDInsight	Kafka broker	25%	6%
Storage Accounts	Data Lake storage long-term and real-time HDFS-API compatible	5%	77%
VMs	Drill	3%	1%
Data Transfers	Outbound traffic	8%	2%

The cost distributions between the two cases shows that the service types excluding the storage capacity is dominant for a case with limited requirement for storage. In order to provide flexibility services as an aggregator, all of the service types in the data architecture are needed, therefore two main conclusions can be stated:

1. The start-up cost to provide flexibility service is dominant over the storage capacity where efficient processing of data, load forecasting and high-throughput capacity is necessary to run the services rather than providing large amount of storage capacity,
2. It is essential to determine which stakeholder should account for the storage capacity and how much storage is needed to run the services to not over dimension the application and thereby increase the cost which in the end will lead to a higher cost for the end user.

4. CASE STUDIES

Chapter 4 is divided into seven case studies which apply the concepts introduced in the first chapters of this document.

- 1) Baseline models and resilience of service delivery.
- 2) Prediction of availabilities and quantities of flexibility services.
- 3) Near real-time residual load forecasting at grid points.
- 4) Data exchange between DSO and TSO.
- 5) Forecasting in integrated energy systems.
- 6) Privacy-preserving data analysis.
- 7) Development of a big data system for the electricity market.

4.1 BASELINE MODELS AND RESILIENCE OF SERVICE DELIVERY

Main section author: Ulf Roar Aakenes (Enoco), Karin Maria Lehtmetts (Elering)

4.1.1 ABSTRACT

An essential element in a flexibility market is that a consumer should be able to offer a reduction (or increase) in their consumption (Demand Response Event - DR) in order to release capacity for other more critical consumers. A payment shall reward released capacity (or consumption of excess), and it is, therefore, necessary to document that the reduction is delivered as agreed. Verification of contracted reduction of consumption can be done in many ways. However, in this part of the study, *models* that estimate what the regular consumption would have been without a DR event – **baseline models** - are evaluated and tested on real data sets.

Most of the baseline-models use data from the nearest preceding days in order to calculate the baseline for the next few hours. Nevertheless, consumption patterns are not regular in shape, large natural variations can be seen from one week to the next, and abnormal (but real) variations can also be observed. Testing the different baseline models on real datasets reveals the model's ability to calculate correctly also during irregular consumption patterns. Also, the focus has been on baseline models that meet the requirements of simplicity and transparency. Payment is involved, and therefore such characteristics are essential to be able to avoid attempts of gaming and to reduce the burden of administration. The EnerNOC, the UK Model, the Average and the Daily Profile models are widely used, and representatives of such models are tested in this chapter. Advanced deep learning models have also been tested on the same real datasets. The tests will show that the simplest models like Average and Daily profiles are the most accurate. The models have also been tested on single large consumers, where the result shows that none of them can estimate adequately. In such cases, regularly produced ex-ante (before the event) baselines combined with real-time monitoring in case of a DR request is a better solution combined with metering data.

The work concludes that simple baseline models often outperform the more complex ones. The production planning much more governs that consumption pattern at single large consumers than by repetitive hour-by-hour patterns. However, at such consumers, their baseline (based on the production plan) can be combined with metering data in order to document DR deliveries.

4.1.2 INTRODUCTION

4.1.2.1 AIM

Quantifying the resilience of service delivery of technologies aims to describe the ability of baseline models to quantitatively describe the degree to which FSP delivers the response they have promised to deliver. It has been realized that the key to address this issue is to find a proper baseline.

4.1.2.2 CONTEXT

The nature of a demand-side flexibility service is that a consumer shall respond to the balancing market by reducing or increasing their consumption by a contracted portion for a specific period. The challenge is then to decide as precise as possible that this consumer responded as agreed. One must have an idea of how the consumption load would look like if the customer did not release or consume as agreed (the baseline), to be able to say if the flexibility service has been delivered as contracted. If one can estimate the baseline precisely – then the difference between the real consumption measurement and the baseline will constitute the real flexibility offered. The problem is that there is no way to be sure what the baseline would be if there were no demand response (DR) event.

Well-designed baseline methodologies enable grid operators and utilities to measure the performance of the flexibility providers.

One would think that this was a task that called for complex algorithms and artificial intelligence. However, in addition to accuracy and integrity, simplicity and transparency are essential characteristics for a baseline. The accuracy is vital in order to evaluate if the flexibility provider delivered as contracted. However, at the same time, the methodology should be simple enough for all stakeholders to calculate and understand. The preferred methodology should minimize the availability of data manipulation and also minimize unintended consequences such as inadvertently penalizing real curtailment efforts.

The consumption patterns differ widely. The consumption can vary based on parameters such as ambient temperature, work hours, weekends. However, it can also be strictly driven by the production planning for one or several large industry players in the grid. One algorithm can succeed in describing the correct baseline for a given consumption pattern, but ultimately fail to describe a different one, especially since simplicity and transparency are essential characteristics of the algorithm. Patterns variety indicates that there is a need for several baseline algorithms suitable for different consumption patterns.

This chapter focuses on the testing of well-established baseline models (where the principles of simplicity and transparency are taken into account) on *real* consumption datasets from single large consumers, and three city-regions. The results of applying Deep Learning Models on the same datasets will be reviewed and evaluated. However, higher complexity results in lower transparency and lower attraction.

Baselines are anyhow only a calculated estimate. Real-time monitoring can give valuable input in order to adjust and improve the estimate. Such monitoring can be done at the consumer side (from smart-meters or added energy meter) or the distribution side (grid transformers). Data is collected and aggregated in cloud-based Energy Surveillance Systems. The meaning of real-time monitoring must also be defined. A DR event may last for single hours, or perhaps also only 15 minutes. For large and dominating loads, milliseconds can be critical.

Figure 4.1 shows an example of a cloud-based real-time monitoring system (more information available at <https://eurora.cloud>). The figure presents real-time monitoring at the second resolution. Data from such a system can be used in real-time evaluation and adjustments of baselines. Eurora.cloud is prepared and applied in DR operations in EU-SysFlex.



FIGURE 4.1 CLOUD-BASED REAL-TIME MONITORING SYSTEM WITH AGGREGATED DATA

4.1.3 METHODOLOGY AND APPROACH

4.1.3.1 MODEL OVERVIEW

Several methodologies for detecting a load reduction have been developed over the years, and more will come. Five basic business practice standards have been defined by the NAESB:

- A. **Maximum Base Load:** A performance evaluation methodology based solely on a Demand Resource's ability to maintain its electricity usage at or below a specified level during a Demand Response Event.

- B. **Meter Before / Meter After:** A performance evaluation methodology where electricity Demand over a prescribed period before Deployment is compared to similar readings during the Sustained Response Period.
- C. **Baseline Type-I:** A Baseline performance evaluation methodology based on a Demand Resource's historical interval meter data, which may also include other variables such as weather and calendar data.
- D. **Baseline Type-II:** A Baseline performance evaluation methodology that uses statistical sampling to estimate the electricity usage of an Aggregated Demand Resource where interval metering is not available on the entire population.
- E. **Metering Generator Output:** A performance evaluation methodology in which the Demand Reduction Value is based on the output of a generator located behind the Demand Resource's revenue meter.

Baseline Type-I and Baseline Type-II are the most common performance evaluation methodologies in use. During this task, baseline methods were evaluated, that all are in category C (Baseline Type-I). In calculations, to define the **window** (W) that will form the baseline (in example last ten days, non-event days/hours), and **exclusion rules** (EX) (for example, previous days/hours with DR events, days with an outage, extreme weather) and **calculation types** (CT) (in example average value, maximum, rolling average). Also, **adjustment rules** (AR) may be defined. Additive or scalar adjustments can be used to bring the curve (with a similar shape) to the same magnitude as the reference curve.

A short description of the methods applied in this study are listed in Table 4.1.

TABLE 4.1 METHODS DESCRIPTION

Method	Short description
EnerNOC	A baseline is equal to the average consumption of 5 corresponding hours with the highest consumption within ten last (W) non-event days (EX). (X of Y) A baseline is adjusted upwards by the average difference between the last two hours' actual consumption and their baseline (AR). (EnerNOC, 2009)
	Extended explanation: Creating a baseline for the hour between 10 and 11, compose the average of the five highest hours (between 10-11) among the last ten days without any special events. Then find out if the baseline calculated for hours 8-9 and 9-10 has a discrepancy compared to today's actual consumption for those hours. If so, correct today's baseline for hours 10-11 with this discrepancy - but ONLY upwards. A baseline will always be positive for the player responsible for DSR. (Asymmetric High five of ten – HFoT)
	Formula: $b_t = \frac{c_1 + c_2 + c_3 + c_4 + c_5}{5} + \max[\frac{c_{t-1} - b_{t-1} + c_{t-2} - b_{t-2}}{2}; 0]$
UK model	A baseline is equal to the average consumption of 5 corresponding hours within five days with the highest daily consumption (out of 10 last non-event days). A baseline is adjusted

	upwards and downwards by the difference between the last two hours' actual consumption and their baseline. (Imperial College, 2009)
	Extended explanation: Among the last ten days (<i>W</i>) without special events (<i>EX</i>), select the five days with the highest total consumption. Use the selected days to calculate the average for the desired hour (e.g., from 10-11). Then adjust up / down (Symmetric) with the deviation between the baseline and the reality of the averages of the two preceding hours within this day (<i>AR</i>). AS the similarity here is the 5 days of highest daily consumption, this is a 'Similar Profile X of Y' model.
	Formula: $b_t = \frac{C_1 + C_2 + C_3 + C_4 + C_5}{5} + \frac{C_{t-1} - b_{t-1} + C_{t-2} - b_{t-2}}{2}$
Average	A baseline is equal to the average of consumption one hour before and one hour after the DR event. (DNV KEMA, 2013)
	Similar to sliding scaffolding.
	Formula: $b_t = \frac{c_{t-1} + c_{t+1}}{2}$
Daily profile	A baseline is equal to the consumption within the past hour multiplied by the fraction of increase/decrease of consumption in the corresponding hours a day before the event. (DNV KEMA, 2013)
	Follow yesterday's curvature, with the baseload of to-day
	Formula: $b_t = \frac{c_{d, t-1} * c_{d-1, t}}{c_{d-1, t-1}}$

b_t –baseline at
hour t ;

c_1 –highest
hourly consumption within 10
last non-event days;

C_1 –highest corresponding hourly consumption
in a day with the highest daily consumption
within 10 last non-event days.

There is also a wide range of other methods such as Naive Model, Persistent Model, ARIMA (Autoregressive Integrated Moving Average) as well as Deep Learning Models like CNN (Convolutional Neural Networks) and LSTM (Long Short-Term Memory). These models were indirectly tested on the present datasets, and the results will be presented later.

4.1.3.2 BASELINE STUDIES PERFORMED ON BEHALF OF AEMO

In 2013 the Australian Energy Market Operator (AEMO) was requested to lead the work with establishing a new Demand Response Mechanism (DRM) and baseline studies. AEMO commissioned DNV KEMA for the work. In Phase 1, DNV KEMA conducted a literature review and interviewed 6 ISOs (Independent System Operators) around their implementation and experience with baseline methodologies. Phase 2 of this work contains evaluated results of different baselines in use.

Some of the key findings from Phase 1 of the study are summarized below.

- That is, the more complex the baseline method, the less likely the demand response mechanism will attract resources to register and participate. The administrative burden of the System Operator to implement the baseline methodology (or methodologies) should be taken into consideration during the baseline selection process.
- Concerns about gaming or strategic behaviour are valid and have been observed in some instances, but these concerns must be weighed against raising unnecessary barriers to entry.

The Phase 2 analyses used a large, robust sample of likely program participants, over a multiple-year frame. The project team tested a broad range of representative baselines and commonly accepted adjustment approaches using multiple metrics to define the baselines' efficacy. The project team tested nine baseline models with up to 4 variants of each. The variants represent the unadjusted baseline and three standards, same day, adjustments to the baseline, including an additive adjustment, a multiplicative adjustment, and a multiplicative capped adjustment. Accordingly, there were 36 different baseline/variant combinations included in the analysis. Concerning the accuracy, the work shows that one of the top-performing baseline methods (CAISO 10 of 10 with an additive adjustment) did not predict the typical customer's half-hourly load better than 10% of their actual load most of the time. For one out of 10 customers, obtained result will be improved for up to 5%, while for nine out of 10 customers, the prediction will typically be up to 22% of the actual load.

There is no need to segment based on weather sensitivity because the use of the same day adjustment improves both the non-weather sensitive and weather-sensitive segments. The same four baselines, i.e., the two X of Y type baselines, ISONE, and CAISO with same day load-based adjustments, are equally effective across non-weather and weather-sensitive segmentations.

Not entirely unexpected, but it may also be worth noting that the baseline models did not perform well on high variable load customers. The result erodes with increasing load variability.

Some of the recommendations from Phase 2 of DNV Kema's work that is relevant for current study is as follows:

- Utilizing an additive adjustment is recommended. The analysis indicates that the same day additive or multiplicative adjustment has superior performance to an unadjusted Customer Base Load (CBL) or a CBL using weather-sensitive adjustment.
- Highly variable load customers should be segmented for purposes of applying a different customer baseline load.
- Administrative and other factors are essential considerations in the final determination of a CBL or CBLs.
- If multiple baselines are used, then demand response aggregators (DRAs) should be allowed to select the baseline.
- Strategic behaviour in the market to artificially inflate the CBL should not be permitted.

4.1.4 TESTING SELECTED BASELINE METHODOLOGIES TO REAL DATASETS

4.1.4.1 PREVIOUS WORK BY BALTIC TSOs AST, ELERING AND LITGRID

A pilot that enables aggregators and other providers to offer DR to balancing markets as the Estonian TSO Elering has initiated an mFRR product. Latvian (AST) and Lithuanian (Litgrid) partners have been involved in the work in order to try to develop a unified Baltic model, and they delivered in 2017 a report called ‘Demand response through aggregation – a harmonized approach in the Baltic region’. In which four basic baseline-models have been tested against 40 data points has been reported. In which four basic baseline-models have been tested against 40 data points has been reported. The four models were EnerNOC, UK Model, Average, and Daily Profile. All of these models meet the requirements of simplicity, integrity, and transparency. The characteristics are summarized in Table 4.2.

TABLE 4.2 THE CHARACTERISTICS OF THE FOUR MODELS IN THE BALTIC TSO TEST

Method	Advantages	Disadvantages
EnerNOC	It is simple to use and it is difficult for an aggregator to exploit the model	The model produces high errors used in forecasting, and the design of the model (asymmetric) creates overestimation of the baseline.
UK model	This mode utilizes symmetric corrections (no over- or underestimation), which gives the highest forecast accuracy	The weekends are still problematic with respect to forecasting
Average	This is a very simple model, also with respect to implementation. The model applies the hour before in an average, and the hour after, and thus the accuracy is very good.	The model cannot be used for subsequent hours, as the ‘hour before’ is already an estimate. The model also fails during peak hours since it is an average. But so do also most of the models.
Daily profile	Since the model is relies on previous profiles for the same site, the forecast accuracy is high. It also applies symmetric corrections and thus remains unbiased.	The method will fail when the pattern of the day differs significantly from previous patterns

Two models use only data from the days before the activation, while the other two uses data from before and after the event. According to the analysis of 40 random consumption patterns, the study concludes that the most accurate baseline method is the “UK model” with an average error of only 2.5%. Second best is the “Average model” with 3.1%, followed by “Daily Profile” with 5.2% and finally “EnerNOC” with as much as 9.6% average error.

The EnerNOC model showed low forecast accuracy and regular baseline overestimation, which always puts the aggregator in a favourable condition. “Average” and “Daily profile” methods showed high accuracy results but did not achieve the performance of the “UK model.” Moreover, the “Average” method showed high forecast errors in peak/off-peak hours, which are considered to be the most demand response intensive incidents (in

theory) as well as it cannot be used for a calculus of baseline in case of 2 or more subsequent DR hours or predictions.

The “UK model” has a symmetrical forecast error, which is preferable to peak/off-peak forecast errors. In the long-run symmetric forecast, an error will bring aggregators in equilibrium, as in some DR events, its baseline will be underestimated. At the same time, in other hours, it will be overestimated. *Ceteris paribus*, symmetric under- and overestimation of the baseline will not allow any DR party to malfunction the system. As a result, fair economic conditions will be created.

It was concluded that the best choice for the Baltic States is following the baseline method used in the UK. According to the results, the UK method produces the lowest baseline forecast error comparing to other methods. It does not require complex calculations, as well as is simple to use and thus communicate.

4.1.4.2 BASELINE CASE STUDIES # 1

Datasets used

The baseline case studies #1 consider the same baseline methods that were used in the aforementioned work by the Baltic TSOs. However, the datasets used in this work are different and are as follows:

- Scenario 1: A single large industry player with around 10 GWh annual consumption. The consumption is driven by production planning and is not affected by weather conditions or local energy pricing.
- Scenario 2: Three city regions with around 20-25 000 citizens, including households and local industries of a different kind. The consumption monitoring has been performed at the transformer level in the distribution grid and constitutes *the net complete electricity consumption within the region of interest*.

Evaluation metrics used

Concerning evaluating the performance of the different baseline models, several evaluation metrics are available. Qualification tests during this report will be performed by MAPE (Mean Absolute Percentage Error).

Scenario 1: A major industry player with 24/7 production.

Scenario 1 is a major industry player with power consumption from 900-1600 kW—accumulated yearly consumption of around 10 GWh. Figure 4.2.A shows the hour by hour power consumption, and the area that is not shaded has been selected for evaluation. Figure 4.2.B shows a 24h by 90 days plot of the same area. What looks like a data error is a real but low hourly value.

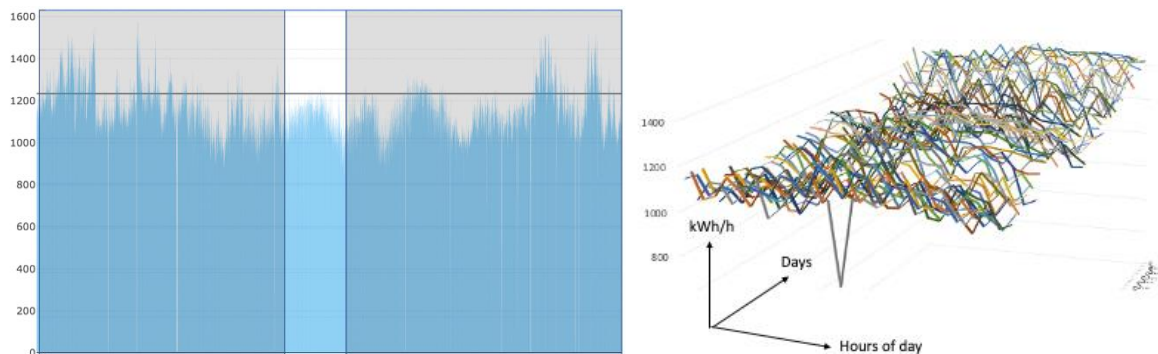
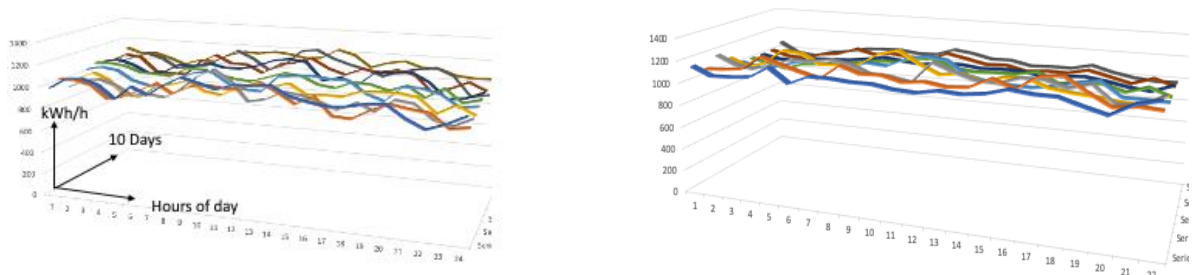


FIGURE 4.2 A) POWER CONSUMPTION HOUR-BY-HOUR B) HOURLY CONSUMPTION 24H, 90 DAYS

The player is a production site with a 24/7/365 production pattern. However, the consumption during the daytime is higher than during the night. Out of this period, ten days for testing the different models have been selected.

Applying EnerNOC model on Scenario 1:

The actual consumption hour by hour for these ten days, and the baseline produced by EnerNOC method is shown in Figure 4.3, and the absolute% deviation is shown in Figure 4.4.



a) Real consumption

b) Baseline by EnerNOC

FIGURE 4.3 CONSUMPTION PATTERNS FOR SELECTED DAYS OF FIGURE 4.1

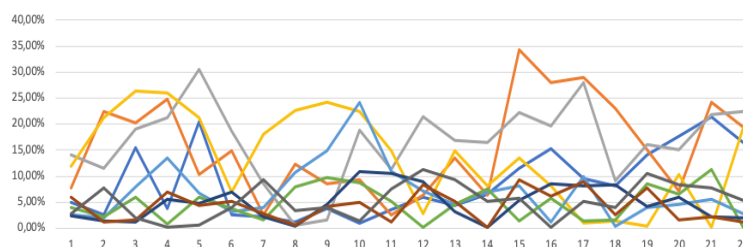


FIGURE 4.4 ABSOLUTE PERCENT DEVIATION FROM THE ESTIMATED BASELINE AND THE ACTUAL CONSUMPTION

As the EnerNOC Method instructs, the baseline is only corrected if the deviation gives a positive contribution. The absolute value of the average deviation is 8.9%, but for single hours it can surpass 30%.

Applying the rest of the models on the same data-sets give the results shown in Table 4.3.

TABLE 4.3 THE RESULTS APPLYING REGULAR BASELINE MODELS ON THE CONSUMPTION OF AN INDUSTRY PLAYER

Model	MAPE	Single % error observed
EnerNOC	8.9%	30%
UK model	7.4%	25%
Average	4%	12%
Daily profile	7%	30%

The consumption pattern at a player like this is not random. It is a result of several different individuals and independent production processes. Even if the consumption is a result of a planned process, no systematic pattern will be observed.

As learned from DNV KEMAs report:

- Baseline models do not perform well on high variable load customers.
- Highly variable load customers should be segmented for purposes of applying a different customer baseline load.

The best way to operate and reward consumers like this is probably that they regularly forward their predicted consumption (baseline) based on their production plan to the system operator. The accuracy of a single baseline like this can be calculated when there is no DR event. In case of a DR request, meter data can be used in combination with the baseline supplied in front of the DR event. The accuracy will be known, and possible gaming will be avoided. Unexpected operational issues that can affect the ex-ante baseline can also be extracted from their Production Management Systems, thus being able to use or accept an ex-post baseline as-well.

Scenario 2: A city-region with 20-25 000 citizens.

Here, the application of the models is presented on a real dataset from a city region, including both households and industrial buildings. The hourly consumption pattern is as shown in Figure 4.5.A. Figure 4.4.B shows weekdays from the selected period arranged as 24 h x 10 days. Weekends are quite different from workdays and are removed (EX rule) from the dataset in Figure 4.5.B) is plotted as a continuous line, but the dataset is based on hourly measurements.

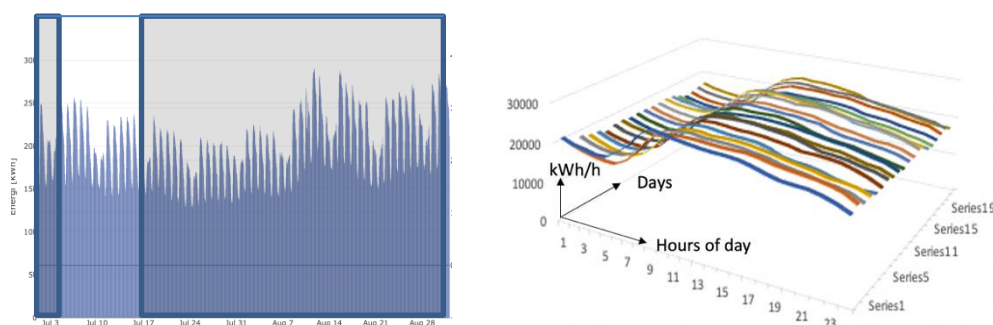


FIGURE 4.5 A) HOUR-BY-HOUR CONSUMPTION B) 24H X 10 DAYS

In Figure 4.6, A) the calculated baseline after the EnerNOC model is plotted, and in B) the absolute value of the percentage deviation.

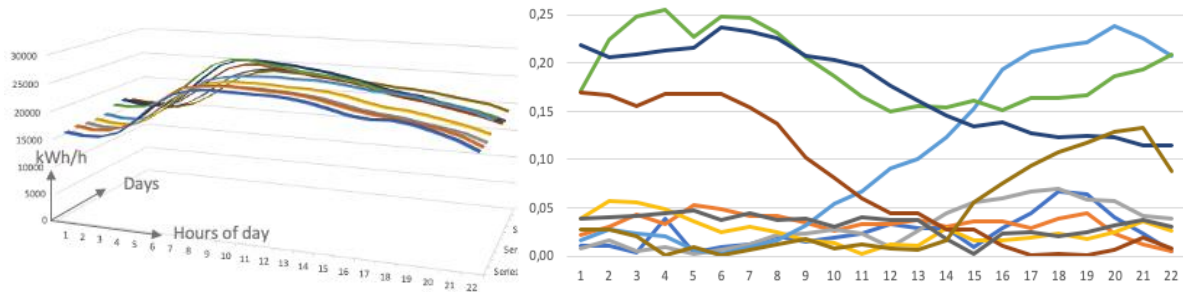


FIGURE 4.6 A) ENERNOC BASELINE B) ABSOLUTE VALUE OF THE DEVIATION

The average of the absolute value of the deviation is 8.3%. The graph shows that it is a few days that generate this large average. They are generated as a result of an actual and real reduction in the consumption pattern.

Figure 4.7 shows the results using the **UK model**, Figure 4.8 shows the **Average** model, and Figure 4.9 the **Daily profile** model.

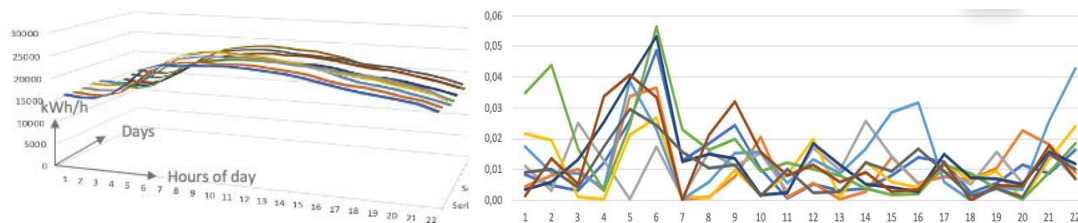


FIGURE 4.7 RESULTS OF APPLYING UK MODEL

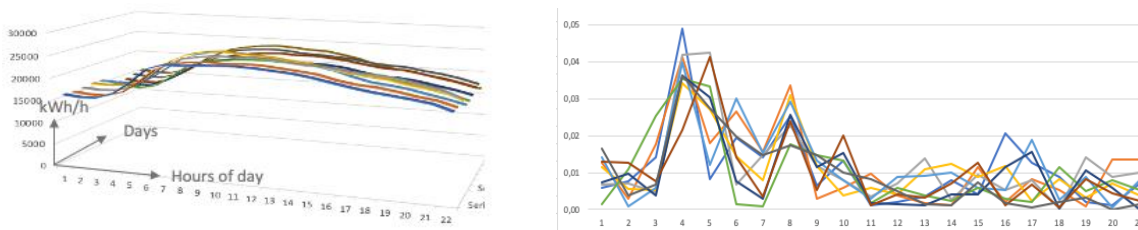


FIGURE 4.8 RESULTS OF APPLYING THE AVERAGE MODEL

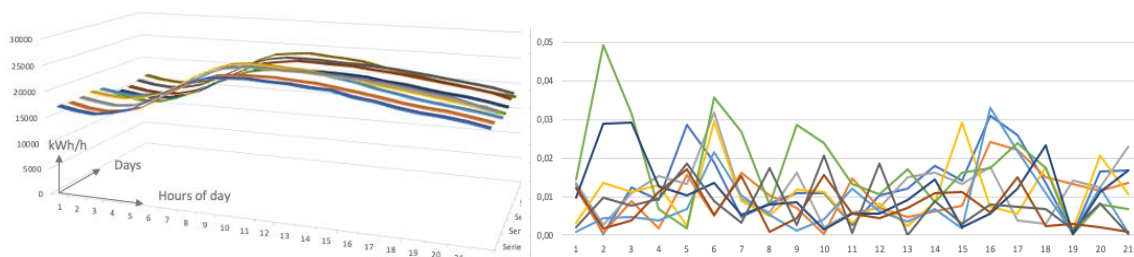


FIGURE 4.9 RESULTS OF APPLYING THE DAILY PROFILE MODEL

TABLE 4.4 THE RESULTS OF APPLYING REGULAR BASELINE MODELS ON THE CONSUMPTION OF AN INDUSTRY PLAYER

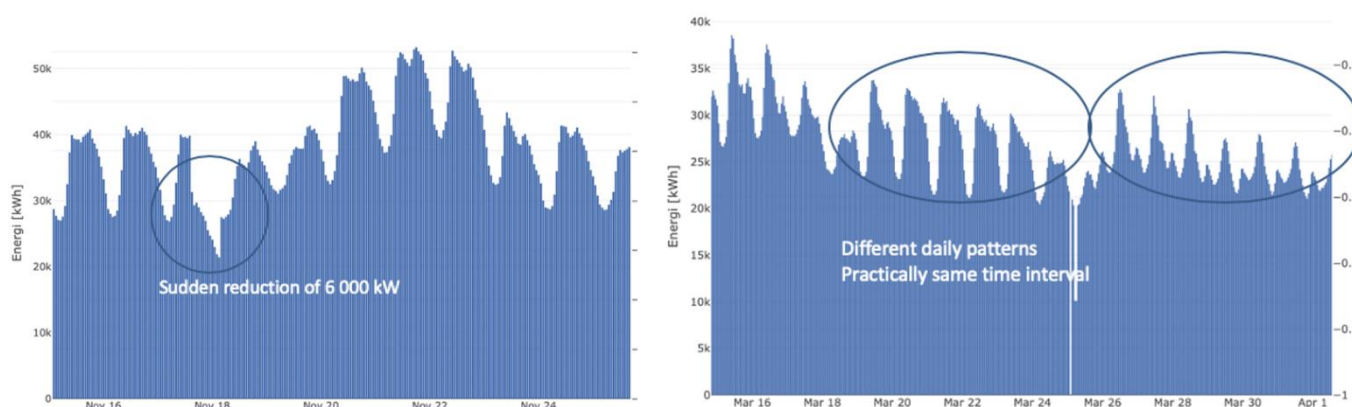
Model	MAPE	Single % error observed
EnerNOC	8.3%	25%
UK model	1.2%	6%
Average	1%	5%
Daily profile	1%	5%

The same exercise has been performed on the other two city regions with practically the same results. As can be seen from Table 4.4, the Average and the Daily profile Model is better suited to follow a consumption pattern like this. In this case, it fits better than the UK model. The standard method seems to have its largest deviations during more massive curvature change, which is natural. The method is not dependent on daily change in baseload caused by variations in temperature. It follows the actual curve of the day. The method can also be **improved** since the direction of the curvature will always be known. It can be compensated with a fraction indicated by the radius of the curvature.

Figure 4.8 shows that the most significant deviation is when the consumption curves upwards. The sharpest curvature is also observed at the same place on the plot. The method overpredicts during these hours. When the curve direction again changes, the method underpredicts, and the same happens during the evening.

Improvements: Actually, the deviation caused by this method can be reduced by introducing a compensation based on the radius of the curvature, and especially its direction.

The data sets in this study contain many individual meters. One would then expect that the total consumption curve was averaged to almost a smooth line, but this is not the case. Figure 4.10 shows that although viewing accumulated metering datasets, strange and unexpected consumption patterns can be found. These results indicate that sophisticated methods will lead to large manual operations and increasing the administrative burden.


FIGURE 4.10 VARIATIONS IN DAILY CONSUMPTION PATTERNS

4.1.4.3 BASELINE CASE STUDIES # 2

The baseline case studies #2 address the same datasets (city regions) which were considered in a study at the University of Tartu (Kurylenko, 2020). The University of Tartu study was more extensive concerning the test window (W) and evaluation metrics. In addition to traditional baseline models (where transparency and simplicity are essential), this study has evaluated the performance of Deep Learning Models like CNN and LSTM. The study has also focused on prediction up to 24 hours ahead. Baseline models using hour before and hour after cannot be applied in such predictions.

Before showing the results of this study, it can be useful to list abbreviations from the study (Table 4.5).

TABLE 4.5 INTRODUCTION OF APPLIED METHODS AND USED METRICS

Method	Short description
CNN	<i>Deep Learning Model:</i> Convolutional Neural Networks consist of convolution layers, pooling layers, and fully connected layers. The CNNs are capable of capturing the local trend and scale-invariant features when the nearby data points have a strong relationship with each other. The CNNs typically combines three critical properties: sparse connectivity, parameter sharing, and equivariant representations.
LSTM	<i>Deep Learning Model:</i> Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feed-forward neural networks, LSTM has feedback connections. It can process not only single data points (such as images) but also entire sequences of data (such as speech or video)
ARIMA	<i>Baseline Model:</i> ARIMA is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to understand the data better or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the integrated part of the model) can be applied one or more times to eliminate the non-stationarity. (Venali Sonone, 2019)
Naïve model	<i>Baseline Model:</i> Naïve Model , also known as the persistence model, is a conventional benchmark in forecasting.
Asymmetric HFoT	<i>Baseline Model:</i> Asymmetric High Five of Ten. In the same group as EnerNOC and UK Model. Asymmetric indicates that the deviation is corrected asymmetrically.
Symmetric HFoT	<i>Baseline Model:</i> Asymmetric High Five of Ten. In the same group as EnerNOC and UK Model. Symmetric indicates that the deviation is corrected symmetrically.
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error

The results of the study made are presented in Figure 4.11. Here MAPE is used to qualify the accuracy of the models. It can be observed that the simplest models (average and daily profile) are competitive.

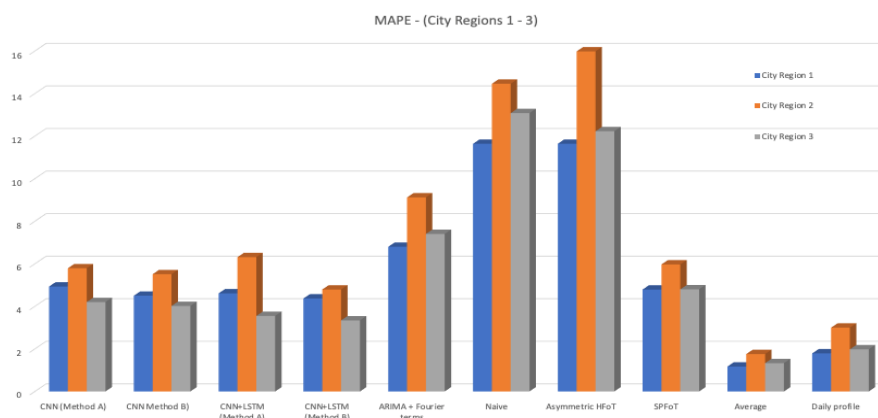


FIGURE 4.11 VALUES OF MAPE ON ALL DATASETS IN THE STUDY (KURYLENKO, 2020).

4.1.5 CONCLUSIONS

In principle:

The baseline models base their entire estimate on the history of the preceding days. *During normal consumer behaviour, these models can estimate the baseline with perfect results, but if the present hour is irregular, the model will fail.* The same will be the case if the preceding days contain irregularities. However, in such cases, the models contain some filtering as they apply a selected subset of the preceding period. A new problem arises if the preceding days contain several conducted DR events.

Single large consumers:

This study shows that none of the baseline models can estimate correctly on single large consumers with irregular production patterns. The DNV KEMA report concluded the likewise: highly variable load customers should be segmented for purposes of applying a different customer baseline load. More information is needed - the models need input on production planning or external factors that will affect or have affected the consumption pattern. For such customers, the pre-produced baseline should be regularly transmitted to the system operator. In case of a DR request, these baselines can be used together with metering data to decide if DR has been delivered as contracted.

Multiple consumer baseline:

All four case studies reviewed in this work show that simple models outperform the more complicated ones. The DNV KEMA study concludes that administrative and other factors are essential considerations in the determination of a CBL. The more complex the baseline method, the less likely the demand response mechanism will attract resources to register and participate. The administrative burden of the System Operator to implement

the baseline methodology (or methodologies) should be taken into consideration during the baseline selection process.

Monitoring in combination with baseline estimates:

The accuracy of the estimate increases as it gets closer to the point of consumption. Meter before / meter after is a different methodology that should probably be used both for single large DR transactions and for the collection of smaller contributors. All of these cases can be handled by a real-time energy monitoring system in combination with transparent and straightforward baseline models. Real-time aggregated info from an energy monitoring system can contribute with valuable adjustments of the baseline models.

4.2 PREDICTION OF AVAILABILITIES AND QUANTITIES OF FLEXIBILITY SERVICES

Main section author: Nick Good (Upside), Mitchell Curtis (Upside)

4.2.1 ABSTRACT

The purpose of this chapter is to explore the requirements of power systems for prediction of availabilities and quantities of flexibility services (i.e., services to balance supply and demand and maintain secure operation of the electricity network), over the various relevant timeframes. Specifically, this chapter estimates the data requirements (in terms of the number of records) for predicting availabilities and quantities of flexibility services.

After describing the different timescales over which prediction of availabilities and quantities of flexibility services are conducted, estimates of such quantities are presented through case studies, which demonstrate how these predictions are made in practice and the volumes and types of data associated with those predictions. Based on estimations of existing data requirements and forecasts of increased flexibility requirements, the future data requirements to predict availabilities and quantities of flexibility services are shown to be significant. Specifically, the number of individual data records required for prediction of availabilities and quantities of flexibility services in real-time for the case study with the highest requirements (Great Britain) was estimated to be 11,038 million/year.

Besides the indication of the scale of the data requirements required for prediction of availabilities and quantities of flexibility services, a major finding was the need for clarity and transparency on the methodologies for prediction. Particularly at the investment and operational planning timescales the methodologies (and hence data requirements) were unclear. Clarity on these methodologies could encourage potential flexibility providers (especially those with long lead times, or for those whose primary purpose is not provision of flexibility services) to make their equipment suitable for providing flexibility.

4.2.2 INTRODUCTION

The purpose of this chapter is to explore the requirements of power systems for prediction of availabilities and quantities of flexibility services (i.e., services to balance supply and demand and maintain secure operation of the electricity network), over the various relevant timeframes. Specifically, this chapter estimates the data requirements (in terms of the number of records) for predicting availabilities and quantities of flexibility services.

After describing the different timescales over which prediction of availabilities and quantities of flexibility services are conducted, estimates of such quantities are presented through case studies, which demonstrate how these predictions are made in practice and the volumes and types of data associated with those predictions. Based on estimations of existing data requirements and forecasts of increased flexibility requirements, the future data requirements to predict availabilities and quantities of flexibility services are shown to be significant. Specifically, the number of individual data records required for prediction of availabilities and quantities of flexibility services in real-time for the case study with the highest requirements (Great Britain) was estimated to be 11,038 million/year.

Besides the indication of the scale of the data requirements required for prediction of availabilities and quantities of flexibility services, a major finding was the need for clarity and transparency on the methodologies for prediction. Particularly at the investment and operational planning timescales the methodologies (and hence data requirements) were unclear. Clarity on these methodologies could encourage potential flexibility providers (especially those with long lead times, or for those whose primary purpose is not provision of flexibility services) to make their equipment suitable for providing flexibility.

4.2.2.1 AIM

This chapter aims to provide some insight into the data requirements for predicting availabilities and quantities of the flexibility services which are required to maintain the secure operation of electricity systems. These services can be provided by generation, demand and storage which can modulate their operation in response to system requirements. This insight is necessary to inform the design of the data exchange, storage and processing capabilities of new Data Exchange Platforms, which can mediate exchange of data between flexibility service providers and procurers.

4.2.2.2 CONTEXT

This work aim is to review the potential data requirements needed when predicting the availabilities of assets able and willing to provide flexibility services and the quantities of flexibility services that they might provide.

4.2.2.3 OVERVIEW

Section 4.2.3.1 provides introduces definition of flexibility services, how they relate to availabilities and quantities in this context. As prediction is a method for understanding the future, the timeframes must be defined. In line

with the ‘Flexibility prediction’ system use case definitions, this review uses three prediction timeframes: Investment (three or more years), Operational (days to three years), and Real-time (intraday), as outlined in section 4.2.3.2. To understand the levels of potential data required for the prediction timeframes a case study approach is undertaken in section 4.2.4.1 by reviewing three different countries, two in the EU (Great Britain and Ireland) and one outside of the EU (New Zealand), for contrast. The case study looks at how each country predicts the requirement (i.e., availabilities and quantities) for existing flexibility services, and readiness of those volumes (i.e., the reliability of delivery) over the three timeframes, and the level of data required for these predictions. Section 4.2.4.2 discusses the outcomes of the case studies and section 4.2.4.3 provides the overall insights gained from this review.

4.2.3 METHODOLOGY AND APPROACH

4.2.3.1 BACKGROUND

The term ‘flexibility services’ in the context of this chapter refers broadly to ancillary services/balancing services/congestion management. The North American Electric Reliability Corporation defines ancillary services as “Those services that are necessary to support the transmission of capacity and energy from resources to loads while maintaining reliable operation of the Transmission Service Provider's transmission system in accordance with good utility practice.” (North American Electric Reliability Corporation, 2008), while the GB Transmission System Operator (TSO), the National Grid, defines balancing services as being used for balancing demand and supply to ensure the security and quality of electricity supply across Britain's transmission system (National Grid ESO, 2019). As noted in the workstream D3.1 report ‘Product Definition for Innovative System Services’ (Nolan *et al.*, 2019), TSOs are the traditional users of ancillary services and future innovations might require the creation of new service types that are used by electricity system actors other than the TSO. Therefore, this chapter uses the term flexibility services to primarily refer to ancillary services for providing flexibility to the grid. However, as Distribution System Operators (DSO), are increasingly looking to overcome localised congestion management issues through the usage of flexibility type services, the findings can be relevant to DSO-procured flexibility services.

Flexibility services are an essential tool for the management and stability of electricity grids. It means that operators of electrical grids need to have a clear understanding of when the flexibility services will be available for usage and the quantity that will be required. To enable planning of grid requirements, they usually operate on set timeframes as a means of determining electricity demand, supply, and payment. For example, in the GB each day is split into 48 half-hourly periods for managing the grid system to ensure enough generation is online to meet anticipated demand. This planning also includes understanding the level of flexibility services needed to handle unexpected events (e.g., a generator failure).

To ensure the reliability of the system is delivered at the lowest cost, the TSO has to assess the quantity required for each half-hourly period carefully. It is done to balance the risk of not having enough, for example, suffering from a blackout, versus spending too much on procuring services that are not used. An additional complication

for the TSO is the need to determine the availability of the flexibility services they have procured. While providers are generally contracted to have flexibility ready in case it is needed, there is always a risk that when called on the provider might not be able to deliver. Non-delivery can be the result of many reasons (for example, a backup generator not being maintained correctly or running out of fuel) and will result in the provider paying penalties. However, the penalty is often minor in comparison to the impact if the TSO experiences a black-out due to not having enough flexibility services available. While the TSO could increase penalties to match the effect of the blackout, this would likely have to be so high that it would mean flexibility providers would be unwilling to offer their services. Therefore, the TSO also needs to understand the expected availability of the contracted flexibility services for each period. For example, they might require 500MW of flexibility services and know that on average, they have a 10% non-delivery rate from providers and so will procure 550MW of services to ensure they have sufficient quantities.

4.2.3.2 PREDICTION TIMEFRAMES

The importance of accurately knowing the quantity and expected availability of flexibility services means that the TSO has to continually predict, over multiple timeframes, what will be required to ensure they have enough capacity when needed. While TSOs will use different timeframes for predicting flexibility requirements, they can be broadly categorized into three planning time ranges - Investment, Operational, and Real-Time. Investment planning (three or more years ahead) aims to understand future flexibility requirements with the prediction being used to ensure long-term measures are in place to provide enough capacity will be ready for the operational timeframe. Operation planning (days to years ahead) aims to predict the quantities and availability of flexibility from existing and new providers in three sub-timeframes (short, medium and long term). Real-Time Planning (Intraday operation) aims to predict the current availability of flexible products for balancing and congestion management requirements for that day. Figure 4.12 illustrates the described planning timescales.

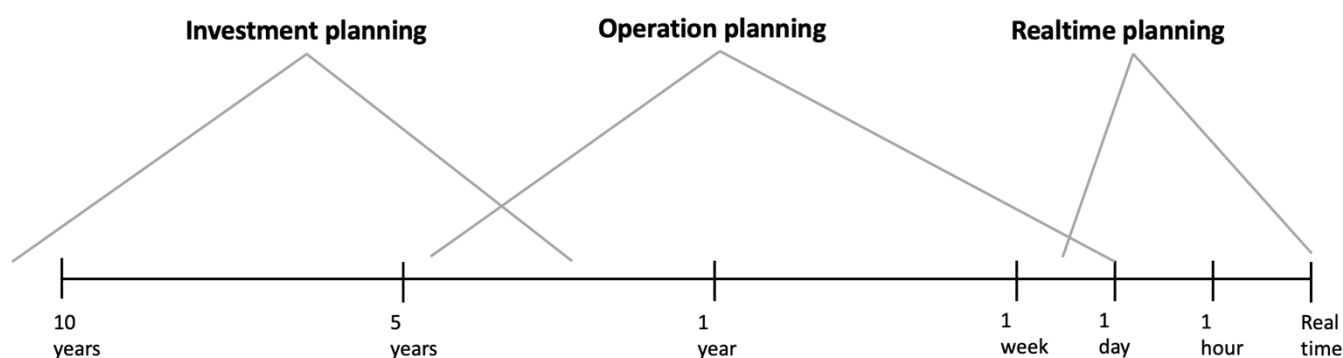


FIGURE 4.12 PLANNING TIMESCALES

Each of these timeframes utilises different types and levels of information and data in the prediction process. This section defines each timeline in detail and provides a generic approach to how prediction could be undertaken.

TIMEFRAME 1 – INVESTMENT PLANNING

The first prediction frame is referred to as ‘Investment’ planning. The investment term is used as this timeframe aims to investigate the future beyond current flexibility capacity that is available for usage, or is in the process of being created, and instead will require investment decisions to be made before the commissioning of new equipment. Generally, it means three or more years ahead and could extend into potentially decades depending on the TSO’s vision. This prediction level aims to understand the future need for flexibility services and if existing and planned services will meet the foreseen requirement. If not, then it will enable the TSO to provide incentives, as required, for flexibility providers to commission more capacity.

A generic approach for investment timeframe prediction would start with the TSO obtaining data on future (more than three years) electricity demand and supply scenarios for the country and individual areas. They would then predict the expected electricity supply over an extended future period based on firm (e.g. generators that can run on demand) and intermittent (e.g. renewables that are subject to external uncontrollable factors) sources, as well as the expected electricity demand and level of inflexible (e.g. demand sources that cannot be temporarily altered like lights) and flexible (e.g. electric vehicle charging) usages over the same period. Using the predicted supply and demand data, they would assess if there is enough flexibility to maintain agreed levels of services. If there is insufficient flexibility, then the TSO signals to the market to procure the required national, and possibly local, flexibility products.

TIMEFRAME 2 – OPERATIONAL PLANNING

The second prediction frame is referred to as ‘Operational’ planning. The operational term is used as this timeframe covers the period required for the TSO to have sufficient flexibility capacity in place to ensure the correct operation of the grid. Generally, it includes a timeframe of days to years ahead, with crossover into the investment plan at around three years, and into the ‘Real Time’ planning phase just before actual usage of flexibility services. Prediction in this timeframe aims to ensure that the correct amount of flexibility services is provisioned for real-time usage. As outlined in the previous background section, the TSO needs to provide enough flexibility capacity is ready for usage without having an excess capacity that will increase operational costs.

A generic approach for operational timeframe flexibility services prediction would involve the TSO obtaining the amount of flexibility required for three sub-timeframes: long-term (years ahead), medium-term (months ahead), and short-term (days/weeks ahead). The long-term period predicts the ‘bulk’ flexibility capacity requirements for the next few years based on overall trends. It allows for the long-term acquisition of this capacity through appropriate mechanisms (e.g. auctions, tenders). The medium-term period aims to refine the flexibility prediction requirements at a detailed level for the next twelve months based on the latest predictors. It allows acquire additional capacity as needed to fill in any gaps. The final short-term period predicts the actual flexibility availability over the next week/month based on the capacity that has been awarded to providers which are then adjusted using forecasting models of actual delivery by the providers based on historical performance data. If any shortfalls in capacity are predicted, then this can be corrected through obtaining additional capacity.

TIMEFRAME 3 – REAL-TIME PLANNING

The third prediction timeframe is referred to as ‘Real-time’ planning. The Real-time term is used as this timeframe covers actual usage of the flexibility services procured via the operational planning period. In the context of flexibility services, the period for Real-time planning usually refers to the current period of grid operation (e.g., next 15 or 30 minutes). However, it can extend to include prediction over the next 24 hours of operation. This timeframe’s forecast aims to ensure that the expected level of flexibility services is available for usage based on real-time data of actual usage and current system/country conditions.

A generic approach for a TSO undertaking Real-time Planning prediction of flexibility services would entail firstly receiving real-time signals/data from capacity providers about their current and near-term ability to provide flexibility. For capacity providers that cannot offer real-time signals, the TSO forecasts their current and short-term ability to provide flexibility based on historical information and prediction parameters (e.g. weather). The TSO then uses the real-time data and forecasts to offer a rolling prediction of available quantities of flexibility services over the next 24 hours, procuring more volume if required.

4.2.4 RESULTS AND CONCLUSIONS

4.2.4.1 FLEXIBILITY CASE STUDIES

The data required to predict the quantities and availabilities of flexibility services varies across each timeframe. To gain an understanding of the data requirements’ scale, a case study review was undertaken. Three countries were selected: Great Britain and Ireland due to being based in the EU, and one outside of the EU, New Zealand, for contrast. For each country, the primary frequency response-based flexibility service is reviewed due to all countries requiring this service for regular grid operation. That service has the highest data requirements for usage. Each country’s case study follows the format of first describing the flexibility service and then reviewing each of the three timeframes to understand how the country undertakes prediction. Where applicable, an assessment of potential data requirements is conducted. As the actual level of data usage by the TSO’s is not public knowledge. If available would require an in-depth understanding of the TSO’s protocols that is beyond the scope of this chapter, the data review element of the case study aims only to provide a scale of the data required.

FLEXIBILITY CASE STUDY 1 – GREAT BRITAIN

Service Description: The GB TSO, The National Grid, is obligated under its license to control the system frequency at 50Hz plus or minus 1% (National Grid, 2020). The flexibility services used for this are either dynamic or non-dynamic frequency response. Dynamic frequency response is a continuously provided service that handles the standard second by second changes on the system (the provider must give a proportional response within 2 seconds to frequency changes between 50.3Hz and 49.7Hz). Non-dynamic answer is a discrete service triggered at a defined frequency deviation (if the frequency reaches 49.7Hz then the response must be provided within 30 seconds and last 30 minutes).

Investment Planning: The National Grid sets out its future requirements in its 'Our Forward Plan' report (National Grid, 2019). However, the future requirements are based on what types of services are going to be required and the way existing services can be improved. The actual amount of each service is then derived via the Operational Planning timeframe.

Operational Planning: The GB TSO acquires frequency response capacity via a month ahead tendering process (though they are looking at moving to weekly) (National Grid, 2017). They publish a report outlining the required amount of capacity to be procured in the forthcoming monthly auction based on expected grid system conditions. Information on how the prediction is undertaken is not provided. However, it is likely to be found on the updated load and variable renewable energy sources generation forecasts and any updated information on generator/interconnection availability. They also indicate requirements up to six months ahead. Assuming that the capacity is predicted for each half-hourly period then each month's prediction will likely be based on the multiple timeframes that include the current year's historical usage and previous year's usage for the months being predicted. It would require providing predictions for up to 180 days (six months) which translates to 8,640 half-hourly periods of data. If they are using a rolling forecast method based on historical data and the previous ten years of information is utilised, then this would require 172,800 historical usage records for predicting the next month's flexibility requirements. So operational planning would create 103,680 data records (assuming the six-monthly forecast is calculated every month), and use 172,800 historical data records, per year.

Real-time Planning: The National Grid does not publish the method and systems used for Real-time prediction of frequency response. However, conversations with National Grid employees imply that the availability of frequency response is predicted with high accuracy for each half-hour period of the day based on the volume of contracted services from providers. The amount available is then adjusted based on an expected non-delivery rate according to historical performance data for the sites and types of assets providing service. An understanding of the data scale required for this prediction can be ascertained based on the reported usage levels. In 2019 the National Grid required between 180-350 MW of capacity depending on the time of day and day of the year (National Grid, 2017). Capacity providers can only contract flexibility services in four-hour blocks with a minimum of 1 MW, though the blocks can be an aggregate of sub-1MW assets. The number of providers varies by month, though as an example, in November 2019 there were 126 contracts. When providing capacity, the flexibility assets are required to record usage every second during the contracted period. It means there are two data requirements for prediction, the data needed to understand contracted flexibility for the current day and the performance data required to understand what was delivered, which will be used to enhance future understanding of prediction vs actual delivery.

The current day's frequency response services quantity and availability prediction data are based on the contracted amounts and historical performance index for the providers. Using the November 2019 records, the highest number of services provided was 350 MW (National Grid, 2017). If provided in 1 MW blocks then this would potentially require 350 data lines per half-hour period that define: who is providing the MW, type of provision (e.g. battery, generator), provisioning factors (response time) and contractual factors (e.g. payment per

MW). The 350 data lines are then multiplied by the 48 half-hourly periods to having a daily requirement of 16,800 lines of data.

The performance data consists of the actual contracted assets usage data. As usage data is recorded every second, this means that if there were 350 MW of capacity provided in 1 MW blocks, then each 1 MW will generate 1,800 measurement values (consisting of a date and time stamp and MW usage value) per half hour, which would be 86,400 values per day. Multiplying by 350 MW works out to be 30,240,000 values per day. Options for collecting, storing and communicating this data are explored in Chapter 2 of this report.

FLEXIBILITY CASE STUDY 2 – IRELAND

Service Description: In Ireland, the TSO, Eirgrid, procures a range of frequency-based reserves with a number of new ones being introduced to manage increased usage of renewable generation (Eirgrid, 2017). The existing method for controlling the grid is through the usage of Primary and Secondary Operating Reserves that need to respond within 5 and 15 seconds (EirGrid/SONI, 2019). To manage more significant levels of variance in the grid, they are now introducing new services, including Fast Frequency Response that needs to respond within two seconds (EirGrid, SEMO and SONI, 2014).

Investment Planning: Eirgrid undertakes a strategic grid review that includes forecasting electricity demand for up to ten years in the future (Eirgrid, 2017). While details of the forecasting method are not provided, they do review how past forecasts compare to actual demand to reflect on variances and changing demand levels. In their strategic report, they mention the usage of flexibility services to help manage the grid, but in this report, they do not directly address what is required. However, the ‘Delivering a Secure, Sustainable Electricity System’ (DS3) programme aims to ‘meet Ireland’s 2020 electricity targets by increasing the amount of renewable energy on the Irish power system safely and securely’. This programme has considered how the grid can handle high levels of renewable generation, which has resulted in new flexibility services being proposed and implemented as reviewed in the following Operational Planning section.

Operational Planning: Eirgrid procures services flexibly in two categories, Volume Capped and Volume Uncapped⁵ (EIRGRID/SONI, 2017). Due to recent changes in the procurement process based on the DS3 programme, the latest capped capacity is being procured two years before being required and with a maximum of six-year fixed terms. The DS3 programme is changing how flexibility services are predicted due to the need to increase the capacity to manage high levels of renewables. While the new predictions are still to be determined a review of the 2017 requirements showed that the traditional primary and secondary operating reserve required is at least 135MW (Eirgrid, 2018). The uncapped services, however, are procured on a six-monthly basis and allow for more significant adjustment to requirements based on medium-term (one to two year) predictions. However, the changing nature of Ireland’s grid and new flexibility services like Fast Frequency Response being launched to

⁵ The ‘Volume Capped’ service fixes the volume procured, with providers competing on price. The ‘Volume Uncapped’ service fixes the price but not the volume procured.

manage higher variability meant the prediction requirements and required data will increase. This is shown in an Eirgrid report on the Facilitation-of-Renewables (Eirgrid, 2010), which modelled 63 different dispatch scenarios to understand the various potential future scenarios and requirements for flexibility services.

Real-time Planning: Eirgrid manages the grid using data inputs from forecasting models and their real-time Energy Management System (EMS) (Eirgrid, SEMO and SONI, 2017). They use a five-day rolling half-hourly demand forecast based on historical usage data with the date, time and weather conditions being the primary parameters. The prediction of availability of flexibility services is undertaken based on three factors, the 'must run' power units, relatively static minimum system inertia level requirements, and dynamic requirements for operating reserves which are a percentage of the Largest System Infeed (LSI) on the system, with the primary operating reserve needing to cover 75% of the LSI.

The data required to predict the quantity of needed flexibility service will vary due to the LSI amount changing across the day. Working on the bases of the 135 MW minimum for each of the primary and secondary operating reserves and an estimated similar amount of the future fast frequency response brings a base total of 405 MW. Reference information on the input calculations for capacity calculation is not provided. Therefore, using the same metrics as per the previous National Grid case study, it can be ascertained that 405 potential data input records (based on one asset per 1 MW) will be required. Multiplying these input records by the 48 half-hourly periods results in a daily requirement of 19,440 data records. Performance data is also needed at one-second intervals, which translates into 34,992,000 records each day ($405\text{MW} * 60 \text{ recordings a minute} * 60 \text{ minutes} * 24 \text{ hours}$) (EirGrid/SONI, 2019).

FLEXIBILITY CASE STUDY 3 – NEW ZEALAND

Service Description: In New Zealand, the TSO, Transpower, procures frequency keeping reserve to maintain the average frequency band of $50 \text{ Hz} \pm 0.2 \text{ Hz}$. The primary flexibility service used for this is a fast reserve (reserve that must act within six-seconds of an off-frequency event and then maintain its post-event output for 60 seconds).

Investment Planning: Transpower sets out its future requirements in its 'Transmission Planning Report' (Transpower, no date b). The report is produced every two years and provides a 15-year forecast of expected demand and generation using both top-down and bottom-up forecasting methods. The details of the modelling methods are not shared. Therefore, data inputs are not known. They also do not directly state the amount of frequency reserve that will be required in the future. They do, however, note where flexibility services could be an option to overcome specific areas of significant transmission constraint.

Operational Planning: Transpower acquires flexibility services, including frequency response, via two methods. Firstly, long term capacity is acquired through yearly tenders that procure capacity on a two-year term (Transpower, no date a). Secondly, short term capacity requirements are calculated from the need to contain an under/over-frequency excursion and return the frequency to 50 Hz following the loss of the largest generating

unit or a pole in the High Voltage Direct Current connection between the two main islands (Transpower, 2019b). While no specific information is provided on the flexibility service capacity required, an estimate can be obtained based on the most significant single generator unit being 500 MW (Energy Market Information, 2015) which sets maximum required size at 500 MW. The data required for operational planning can be calculated based on the yearly tender covering capacity for two years ahead, which means that Transpower predicts capacity requirements for two years being tendered. With half-hourly timeframes, the output of Transpower's prediction for two years ahead translates into 35,040 capacity requirement values (48 half-hourly periods * 365 days * 2 years). To generate these predictions, it is assumed they use historical records, with at least the last ten years of records being used and therefore requiring 172,800 historical half-hourly usage records. Additionally, other input factors are likely to be used, including weather conditions and hydro lake reserves (due to the majority of generation in New Zealand being from hydroelectricity).

Real-time Planning: Transpower uses their Reserve Management Tool and Scheduling, Pricing and Dispatch application to calculate the amount and type of reserve required to maintain quality standards (frequency, voltage, and time deviation) (Transpower, 2019b). The flexibility service requirements are predicted based on covering the loss of the most tremendous generation asset, which will change throughout the day as different generators are started and stopped. The data required to predict the number of flexibility services will vary throughout the day due to the varying levels of a generation that need covering. Working on a worst-case scenario that they need to cover the failure of the largest generation asset means they need to predict the availability of up to 500 MW of flexibility services. The size of flexibility services providers will vary as the primary source is through reserving a proportion of other generation units' capacity to cover the 500 MW required (e.g. all generation units are paid to run at part load in order to have enough spare ability to quickly come on online to cover the failure of another generation unit). Therefore, it is unlikely there will be 500 different providers, each with 1 MW of reserve; however, to understand the maximum required data, it will be assumed that 500 various providers are being used. While Transpower does not specify the inputs for its management and prediction tools it can be expected that at a minimum it requires similar information as used by the National Grid namely: who is providing the MW, type of provision (e.g. battery, generator), provisioning factors (e.g. response time) and contractual factors (e.g. payment per MW). This will result in 500 potential data input records that are then multiplied by the 48 half-hourly periods to having a daily requirement of 24,000 data records. Performance data is also required to monitor availability and usage of the reserves (see Cost of data exchange for energy service providers). Transpower requires that providers measure response in intervals no more significant than six to ten seconds (depending on the type of service being provided) (Transpower, 2019a). Based on 500 MW of the capacity supplied in 1 MW blocks then this will generate at six-second intervals 7,200,000 records each day (500MW * 10 recordings a minute * 60 minutes * 24 hours).

FLEXIBILITY CASE STUDY SUMMARY

Table 4.6 presents a summary of the number of records estimated to be used in the prediction of quantities and availabilities of flexibility services, where these were able to be estimated.

TABLE 4.6 ESTIMATION OF ANNUAL NUMBER OF RECORDS NEEDED FOR PREDICTION OF QUANTITIES AND AVAILABILITIES OF FLEXIBILITY SERVICES

		Case study		
		Great Britain	Ireland	New Zealand
Planning timescale	Investment	Unknown	Unknown	Unknown
	Operational	276,480	Unknown	207,840
	Realtime	11,038 million	12,772 million	2,628 million

4.2.4.2 DISCUSSION

The case study review of flexibility service prediction in three different countries shows several similarities in the approaches taken. For the investment planning timeframe (three or more years), all three countries provided future strategy reports on what is expected to happen and be required in 10 to 15 years. Though predictions of electricity demand and generation were provided, the methods used to determine the values were only described at a high level. At the same time, all countries noted flexibility services as being required, none provided any predictions on the quantities needed. There was also no prediction of the expected availability of future services. Instead, it appears that the actual determining of required flexibility is not addressed until the operational timeframe. The closest a country came to discussing the future need of flexibility services was Ireland with the DS3 programme. It outlines the need to meet renewable energy targets but does not directly specify quantities needed, only different types of services that will be required. The lack of future prediction of amount and availability of flexibility services means that only technologies that can be implemented within one or two years can be utilised. It creates limit on the broader adoption of flexibility services from assets that cannot be easily enabled to provide those services. For example, domestic air conditioning units can be utilised for flexibility services. However, retrofitting the capability to provide flexibility services would be prohibitively expensive, and such devices would have to be enabled to offer flexibility services during installation. Such devices could only be used to provide flexibility services if the requirement was identified before installation. It requires long-term vision and support for flexibility services to provide the necessary incentives and to encourage providers to utilise these types of assets.

At the operational planning timeframe (days to three years), the three countries use a similar multistage approach for procuring flexibility capacity. Firstly, they procure long-term capacity (up to two years ahead) that provides a base quantity of flexibility service. Secondly, they acquire short-term capacity (weeks to months ahead) to meet any additional flexibility service requirements based on the latest usage and predictions. Unlike the investment planning timeframe, the operation planning timeframe provided high-level values of the quantities of flexibility services required. The amounts varied by country, as would be expected due to different mixes of generation and population sizes. Numbers could be determined at this timeframe, the methods and information on how the prediction is undertaken were not provided. While the methods are not provided, it can

be assumed that they conduct a form of historical data analysis based on the electricity demand and supply following recurring yearly patterns. Like the investment planning timeframe, no information is provided on how they determine the availability of the capacity procured. The availability of the flexibility services obtained is an crucial factor in ensuring sufficient capacity is ready for utilisation. They likely have a common non-delivery percentage factor calculated for each type of flexibility service that would be calculated based on historical performance analysis. For example, if they analyse the usage of gas turbine generators flexibility services and determine that over the last five years those assets have failed to provide agreed capacity 20% of the time, then they would use this to be their non-delivery factor. This would then be used when deciding on the amount to be procured to ensure that enough extra capacity for covering expected non-delivery is available.

At the real-time planning timeframe (intraday), the three countries use software tools to manage flexibility service requirements for the current half-hourly operational period and near term (the next few hours to days). The tools assist the grid management teams by helping with the scheduling, pricing and dispatching of services. Approaches for determining the number of flexibility services required throughout the day varied by country, though Ireland and New Zealand had a similar approach of needing to have enough capacity to cover an unexpected loss of the largest generation source. No information was provided on how they determine the availability of flexibility services beyond an understanding gained from conversations with the GB National Grid that they do adjust services based on historical performance. Information on the data used to undertake the predictions of real-time flexibility service quantities and availability was limited. It is estimated that all three countries would use similar parameters, potentially including the type of provision (e.g. battery, generator), provisioning factors (e.g. response time) and contractual factors (e.g. payment per MW) due to the lack of information on predicting real-time quantities. The quantities would at a minimum be predicted up to a day ahead at half-hourly intervals for each MW of flexibility services being required. This prediction would initially be undertaken using the contractual procurement data, which for half-hour intervals translates to 48 lines of data per MW per day (if an MW is the minimum procurement level).

Determining the quantities of flexibility services required for each half-hour period is, at first glance, straightforward due to being based on procured capacity. The complexity increases due to the need to understand their availability – just because a provider has been contracted to deliver flexibility do not mean it is going to. Service providers aim to provide the agreed capacity. However, there is always a risk to experience issues that cause service providers failing to deliver some or all of the contracted service. Predicting when services will not be offered is not straightforward due to the causes of failure being unexpected. It means that the prediction will be based on historical behaviour of different flexibility service types to build a profile of the risk associated with each one. To support the risk profiling, all countries have access to detailed performance monitoring records of the flexibility services when operating. In GB and Ireland, the flexibility service providers are required to monitor the flexibility assets usage levels every second, in New Zealand it is every six seconds. It generates massive amounts of data that can be utilised for analysing the assets performance and ability to delivery against contractual obligations. How this data is analysed to determine availability is not shared by the

countries, likely due to commercial sensitivity as it would enable potential manipulation to make competitor services appear more available than it is.

The level of performance data collected is already significant and will increase as more flexibility is required in the future. Based on the current approach of collecting usage data every second per MW of flexibility results in 31,536,000 records per year (60 seconds * 60 minutes * 24 hours * 365 days). The amount of data rapidly increases based on countries already having over 500 MW of flexibility - 15,768,000,000 records per year. With increasing flexibility requirements, this could result in tens of GW being required across the EU. Additionally, the MW minimum level of recording may drop, going to 100 kW or lower. It is not inconceivable that flexibility services may generate trillions of records each year if domestic level services of 10 kW are recorded every second (10 GW of services being provided by a million residential houses would result in 31 trillion records). While data storage systems can handle this level of data, the availability to use it for prediction purposes becomes complicated and computationally heavy. Therefore, it is likely that in the future, the data requirements will be adapted as needed to provide the required balance between accuracy and ease of data collection/processing.

4.2.4.3 CONCLUSION

This study presented different timeframes for prediction of availabilities and quantities of flexibility services with a general description for the procedure for predicting availabilities and amounts of flexibility services for each timeframe. Then, three case studies, each on a different system, have been presented to explore the process of predicting availabilities and quantities of flexibility services in practice. As part of that exploration, the data requirements for the methods of prediction in the different timeframes were estimated. Finally, the outcomes of the case studies were discussed, and some overall insights from this review were formulated.

The key result of this work which meets the aim of the chapter (to provide insight into the data requirements for predicting availabilities and quantities of the flexibility services), is the quantification of the likely scale of the data requirements for predicting availabilities and quantities of flexibility. The key metrics (in terms of the amount of data to be collected, stored, analysed and communicated) will be those related to the performance of a flexibility provider. In two of the three systems studied, relevant data is collected at one-second intervals, generating a large amount of data. A second critical insight is that the data requirements for monitoring performance of flexibility providers are likely to increase due to trends for increased demand for flexibility services and provision of flexibility services by smaller units.

Another critical insight is that the lack of transparency on methodologies for calculating/ estimating the availability of flexibility providers complicates the assessment of the data requirements for predicting the availabilities and quantities of flexibility services. Publishing of formal methods that system operators' use for assessing availabilities, which should reveal how various attributes (e.g., technology type, a historical record of performance, location, declared availability) affect the evaluation of unit availabilities, would offer clarity to potential flexibility providers, and clarify the associated data requirements. However, safeguards would have to

be put in place to avoid ‘gaming’, as flexibility providers try to improve their scoring against formal methods for assessing availability, if that produced undesirable behaviour (i.e., incentives are not correctly designed).

Clarity on the methodologies for calculating or estimating the availability of flexibility providers, and long-term commitment to those methodologies, would also enable and encourage new sources of flexibility services to be developed. Especially those with long lead times, or for those whose primary purpose is not provision of flexibility services such as air-conditioning units, which require enablement at installation. Many of these sources of flexibility could be non-traditional sources, which may result in new data requirements concerning predicting the availabilities and quantities of flexibility services.

4.3 NEAR REAL-TIME RESIDUAL LOAD FORECASTING AT GRID POINTS

Main section authors: Katharina Brauns (IEE), Nicolas Kuhaupt (IEE)

4.3.1 ABSTRACT

One of the aims of the EU-SysFlex project is to provide solutions for the integration of a large proportion of renewable energy, which is increasingly variable, while maintaining at the same time the safety and reliability of the European electricity system. Especially for the security and reliability of the system it is essential to have a good knowledge about the grid states. Therefore, forecasts for the next few hours of infeed and load are needed. In order to be able to forecast future grid states, these forecast values of generation and consumption are required. These forecasts are then processed together with other grid data in the German demonstrator of WP6 to create schedules for active and reactive power feeds from generators and loads needed for congestion management and voltage regulation in the transmission and distribution grid to optimize power flow.

The focus in this chapter is the residual load forecast calculated in the forecast system of the German demonstrator and its timely provision to the DSO. It includes the delivering of the residual load forecast of a large number of around 1500 transformer stations in time under at least 15 minutes for the active and reactive power. The forecasts are generated using a Long Short-Term Memory (LSTM) machine learning approach. This chapter examines three different approaches to measuring the processing time for the timely provision of all forecasts to the DSO in a near real time system which means that the forecasts are continuously delivered every 15 minutes to the DSO. It includes a big data approach using a Hadoop cluster which is compared to the usage of a standalone server by using at first up to 32 CPUs and in a second evaluation phase 2 GPUs. The main challenge is that the focus is on evaluation in a near real time system rather than on the probably more widely used variant by training a variety of forecasting models. In this case, the forecast model is only used to calculate the forecast for one time step that only includes a small input data set instead of large data sets with the purpose for training a deep neural network.

Nevertheless, for the Hadoop cluster and the GPU approach, there is still a certain amount of traffic that needs to be taken into account, which is time consuming compared to the fast calculation of the forecast itself. Finally, it was demonstrated that these forecasts can be generated with all three approaches.

Regarding the comparison between the different approaches the evaluation on the Hadoop cluster and the GPU did finally not outperform the usage of CPUs. For the delivery of about 3000 forecasts (including active and reactive power) under 15 minutes the usage of a stand-alone server with 5 CPUs is still sufficient. As a conclusion, much experience was gained in evaluating the forecast in a near real time scenario using a Hadoop cluster, a stand-alone CPU and GPU server.

4.3.2 INTRODUCTION

The focus of this work relies on the guarantee of a residual load forecast value, which needs to be delivered to the DSO in time under 15 minutes. The basis of this work is built on the available and required data for the development of the forecast system of the German demonstrator described in D6.2 “Forecast: Data, Methods and Processing. A common description”. The German demonstrator is developed in Work Package 6 (WP6) where three demonstrators are set up including also Italy and Finland. The main objective of the demonstrators is to provide system services from the DSO to TSO which should result from the optimized usage of distributed flexibility resources connected to the distribution grids. Forecast values are necessary to predict possible future grid states. They are used to create schedules for active and reactive power injections from generators and loads, which in turn are needed for congestion management, and voltage regulation for the transmission and distribution network. The German demonstration processes these forecasts together with other grid data to optimise power flow and determine P and Q flexibility bands in future grid states. The residual load forecast results from the vertical power flow subtracted by wind and PV generation at grid connections in HV and aggregated at HV/MV substations.

In the context of massive data flows the delivering of forecasts “in time” gets more and more complicated as the number of prediction points for which a forecast is calculated increases. Therefore, three different approaches for the investigation of the processing time with an increasing number of prediction points is evaluated and compared in this chapter. This includes a big data approach using a Hadoop cluster which is compared to the usage of a standalone server with 32 CPUs and 2 GPUs. First, only the CPUs are used to evaluate the processing time and in a second evaluation phase the 2 GPUs are added. The big data approach with a Hadoop cluster is mostly used for dealing with outages. Here a redundancy strategy is automatically included, if one node in the cluster fails, another takes over the work, so that the delivering of a forecast value can be guaranteed.

A typical use case where a Hadoop cluster or GPUs are used is generally speaking to accelerate parallel computation with large data sets, such as when training a large number of forecast models. However, since the focus of this thesis is more on real-time operation, where mainly smaller data sets are used, the question arises to what extent an acceleration on these two systems can still be achieved compared to the simple use of a CPU

server. The issue here is that the data traffic takes up a large part of the data, which is time-consuming compared to the fast calculation of the forecast itself.

Additionally, for using a high number of forecast models (e.g. up to 3 000) it is crucial to have a generic robust model architecture, which gives high quality forecasts for each forecast point. In order to avoid having to develop a separate, suitable architecture for each model, the master thesis, Abdullayeva (2019), investigated whether there could be a robust architecture that would provide good results for all models. As a result, a 'Univariate Stack' LSTM model architecture was evaluated as a robust model architecture. This master thesis was supervised in a co-working process of the Fraunhofer IEE and the University of Tartu in the framework of the EU-SysFlex project and is described in *“Application and Evaluation of LSTM Architectures for Energy Time-Series Forecasting”* (Abdullayeva, 2019); see section 4.5 of this document, “Forecasting in integrated energy systems”. In the following subchapter the aim of the different approaches, the context with constraints due the forecasting system of WP6 and an overview of the experiments are given.

4.3.2.1 AIM

Forecast values are necessary to predict future grid states. The German demonstrator processes these forecasts together with other grid data to optimise power flow in future grid states. One of the produced forecasts is the residual load forecast. The residual load forecast results here from the vertical power flow at grid connections in the high-voltage grid and aggregated at high- and medium-voltage substations, from which wind and PV generation is subtracted.

The aim of the analysis is to compare different approaches used to measure the processing time for delivering this residual load forecast. This includes the identification of the best, practical solution, which has on the one hand the shortest processing time, but also on the other hand an easy to use application.

4.3.2.2 CONTEXT

For a better understanding of the processes that are considered for the delivery of the residual load forecast, the idea of the forecasting system is shown in Figure 4.13. As there is its own deliverable for the different demonstrators and their including forecasting systems of WP6 (D6.2 “Forecast: Data, Methods and Processing. A common description”), only a small overview is given.

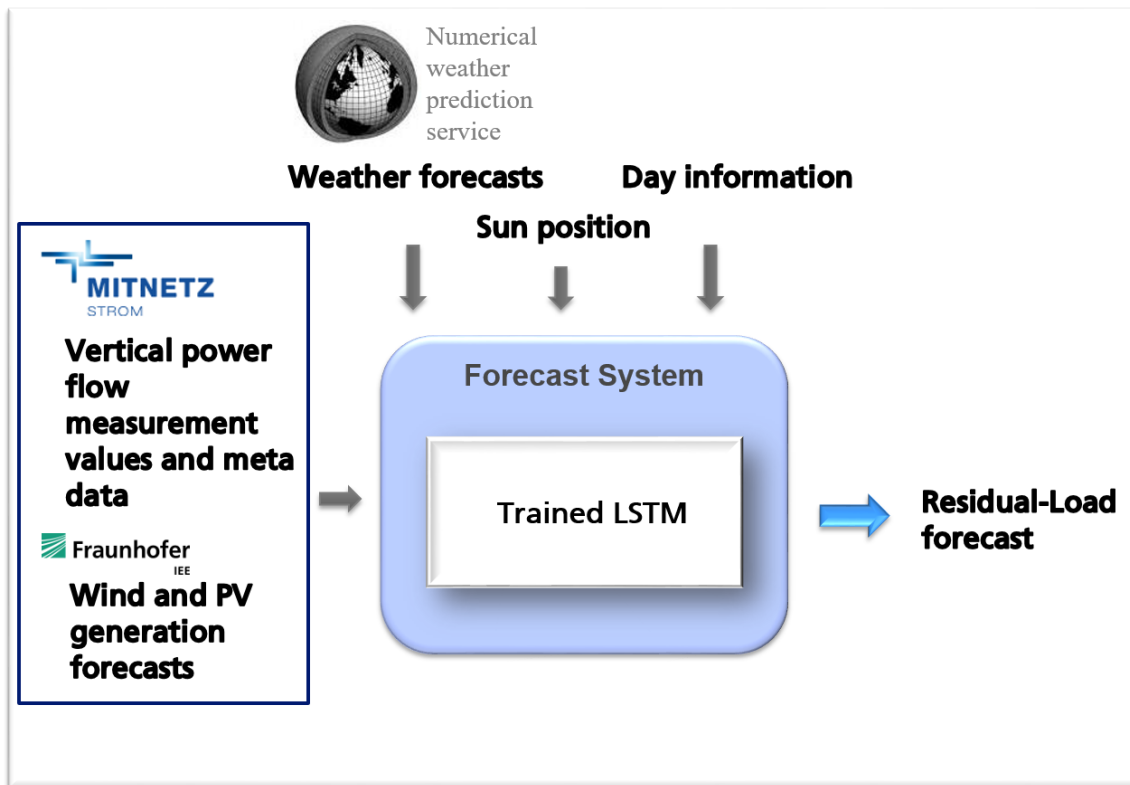


FIGURE 4.13 FORECAST SYSTEM FOR CALCULATING THE RESIDUAL LOAD FORECAST USED IN THE GERMAN DEMONSTRATOR IN WP6

The forecast for the residual load is calculated for each transformer station in the medium voltage connected to the high voltage grid. The residual load is the vertical power flow where wind and photovoltaic generation is subtracted at each transformer station. The input data for the forecast consist on the one hand of the vertical power flow measurement values from the DSO (Mitnetz) and the wind and photovoltaic generation forecasts used here with the shortest forecast horizon, but on the other hand also of numerical weather predictions, the sun position and day information data. The time resolution of these input values are 15 minutes for the historical as well as the online data. The numerical weather has a lower resolution and is therefore interpolated. The calculated residual load forecast has a maximum lead time of about 48 hours. The forecast is generated and delivered every 15 minutes with a time resolution of also 15 minutes. Instead of a shallow learning architecture used in the preceding project “SysDL2.0 - Systemdienstleistungen aus Flächenverteilnetzen”⁶, a deep learning approach with Long Short-Term Memory (LSTM) layers is used here.

The challenges of the forecasting processes involve the loading of input data, calculating forecasts up to 3 000 models and saving of the results within less than 15 minutes so that a delivering of a forecast value can be guaranteed. It leads to the main tasks of delivering an up-to-date forecast in time.

⁶ <https://www.sysdl20.de/>

Due to the massive data flow aspect, several scalability studies are investigated where e.g. the number of models is increased. Therefore, big data tools like Apache Spark⁷ and Apache Hadoop⁸ are analysed. A technical architecture and a way of working is developed to address the specific issues associated with the forecast models. Issues under consideration include, among others, the data transfer from MongoDB to Apache HDFS⁹ and the consistency between the models and the used data. The detailed requirements for evaluating the forecast models in this specific cluster are the following: The forecast models are created with Python using Scikit-learn¹⁰, Keras¹¹ and TensorFlow¹². The time constraints imply a delivery of the forecast within 15 minutes. However, due to additional processes like subtraction of the wind and PV forecasts, the model evaluation time is minimized to only a few minutes. Due to already existing forecast modules which are used in the forecast system, the general applicability of the existing modules is essential. It should also be guaranteed that a forecast value for each time step is available which could either be solved through redundancy or failover in case of failure.

4.3.2.3 OVERVIEW

Three different approaches evaluate the processing time to deliver the residual load forecast. At the one hand is the big data Hadoop cluster with several different tools which needed to be researched and analysed for their applicability for the forecast tasks. The investigations on the Hadoop cluster are not only extended by additional analyses using the standalone server for comparison purposes, but also two other reasons. Parallelizing on one server has the advantage that the data distribution is not needed which simplifies the management of workload, since there is no network traffic. The additional parallelization via GPUs fits also better the purpose of using neural networks. Therefore, comparisons are made with the same experiments on server with 32 CPUs (Central Processing Unit) and two GPUs (Graphics Processing Unit; NVIDIA Tesla P100) server. The primary purpose of investigating the experiments on the GPUs was to evaluate how the additional usage eventually speed up the process. Since for neural networks the main computational part contains multiple matrix/tensor multiplications, GPUs are usually the optimal choice for training a deep neural network with a huge dataset, which can efficiently parallelize these kinds of operations. In this context we consider the question if this is still valid in a near real time scenario where we just apply the already trained model to online data.

The forecasting process involves the loading of the input data and the corresponding model, the calculation of the forecast and finally saving the result. The experiments mainly serve to investigate how long the forecasting process takes for a different number of underlying models. For the online process in WP6, 1415 transformers are included and a separate model was trained for each transformer. As not only a forecast of the active power is needed but also a forecast of the reactive power the number of the models will be doubled. At the moment

⁷ <https://spark.apache.org/>

⁸ <https://hadoop.apache.org/>

⁹ https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

¹⁰ <https://scikit-learn.org/stable/>

¹¹ <https://keras.io/>

¹² <https://www.tensorflow.org/>

models for 585 transformers are available and therefore the following counts (i.e. number of assumed transformers) for the experiments are chosen: 1, 10, 100, 585, 1 000, 3 000 and 10 000.

4.3.3 METHODOLOGY AND APPROACH

In the following, an overview of the methodology and approach of the different experiments is given and the innovation of this approach is discussed.

4.3.3.1 OVERVIEW

BIG DATA CLUSTER SETTINGS

As a prerequisite for evaluating the options for big data computations, the tools had to be researched and evaluated if they can be used within existing IT infrastructure. One of the obvious choices for big data Tasks is Hadoop and its ecosystem. This is Hadoop Distributed File System (HDFS), Hadoop YARN (Yet Another Resource Negotiator) and supplemented by Apache Spark. All tools are developed as big data tools and have certain properties engrained in their implementation. This is amongst others scalability and fault tolerance. Scalability means in this case to distribute the workload (computation or storage) over many servers. Thereby, increasing the amount of computation or storage needed by an application should not lead to an increase in time needed but the workload being distributed over more nodes (server) within the big data cluster. As every server can (and will) fail at some point, fault tolerance is needed in the big data cluster. Therefore, all the tools offer fault tolerance and will still work if one server fails.

HDFS is the tool for storing data. It has one master node, which is responsible for managing all the other slave nodes. The slave nodes in turn are responsible for storing the data. There is just one slave instance per server. If a file is saved in HDFS, the master node decides on which slave node it is saved. The files are split in partitions and each partition may reside on another node. To offer fault tolerance, each partition is saved on at least three nodes. Therefore, if one node fails, there are still two copies of the data left. As soon as a fail of a server is detected, the master node also initiates to copy the data again, so that three copies are again available. The master node also has an extra node which takes over in case of failure. All of this is abstracted away by HDFS and one can communicate with HDFS the same way as the standard file system on Linux. HDFS is in this project used to save both the data and the LSTM model. Through special Python libraries (Pydoop¹³) both can be loaded within Spark.

Yarn is responsible for scheduling the computation resources. It works with containers on each node of the server and can schedule those containers with a certain amount of RAM and CPU to execute computations. Again, there is a master service, which takes care of restarting failed containers. Yarn offers also a feature called “locality awareness” which makes sure, that computations based on data in HDFS is carried out on the server, where the

¹³ <https://pypi.org/project/pydoop/>

data resides. This makes sense as sending the data to the computations is far more expensive (in terms of network traffic and time) than sending computations to the corresponding data. Yarn is here used to schedule all the computations on the cluster.

Apache Spark is the tool for implementing the computations and offers a library for distributed computations. It is also called a general purpose framework as through its broad library it is used for many different use cases. It can also work with different programming languages. Python is used for the machine learning part of the forecasting system. Spark is used for loading the data and LSTM models from HDFS, transforming the data, evaluating the LSTM models and saving the results back to HDFS. To let the computations of Spark run on the cluster, there is a special scheduler from Spark included. However, it also works on top of Yarn, which is used here. The data is saved in HDFS. Additionally, the existing Keras Models are also saved and loaded from HDFS.

In Figure 4.14 is an overview shown of the available big data cluster and its resources. It is important to notice that this cluster is a small cluster and only a sandbox for experimentation. Both the number of servers available and the resources per server is limited. In Table 4.7 also an overview of the big data resources used on the Hadoop cluster is given.

<input type="checkbox"/>	Name	IP Address	Rack	Cores	RAM	Disk Usage	Load Avg	Versions	Components
<input type="checkbox"/>	[redacted] iee.fraunhofer.de	[redacted]	/default-rack	2 (2)	3.86GB	<div><div></div></div>	0.21	HDP-2.5.3.0	18 Components
<input type="checkbox"/>	[redacted] iee.fraunhofer.de 2	[redacted]	/default-rack	4 (4)	7.80GB	<div><div></div></div>	0.06	HDP-2.5.3.0	7 Components
<input type="checkbox"/>	[redacted] iee.fraunhofer.de 2	[redacted]	/default-rack	4 (4)	7.80GB	<div><div></div></div>	0.20	HDP-2.5.3.0	7 Components
<input type="checkbox"/>	[redacted] iee.fraunhofer.de 2	[redacted]	/default-rack	4 (4)	7.80GB	<div><div></div></div>	0.00	HDP-2.5.3.0	6 Components
<input type="checkbox"/>	[redacted] iee.fraunhofer... 1	[redacted]	/default-rack	4 (4)	15.67GB	<div><div></div></div>	0.16	HDP-2.5.3.0	27 Components

FIGURE 4.14 OVERVIEW BIG DATA CLUSTER AND THE RESOURCES

TABLE 4.7 OVERVIEW OF BIG DATA RESOURCES ON THE HADOOP CLUSTER

Number of servers in the cluster	5
Edge node	4 GB RAM and 2 cores
Worker Nodes	8 GB RAM and 4 Cores, respectively one server with 16 GB RAM
Resources of GPU cluster	2 Tesla P100

In the following the procedure and implementation for the evaluation will be described. It is essential that by distributing both models and data over the whole cluster, the data and model which belong together are also saved on the same node. Otherwise at execution time data or models would have to be sent over the network, which is a costly overhead in terms of time and resources. Therefore, both data and models were saved on the cluster with the maximum replication count.

In Spark the PySpark API¹⁴ was used to load the files. Spark works with an abstraction called Resilient Distributed Datasets (RDDs). Those Datasets are distributed in the sense that the data is split into partitions on different nodes (generally on the node the partition was saved on) and resilient in the way that failed computations can be restarted. On top of the RDDs several data transformations can be implemented. Firstly, every partition is transformed with the python package numpy from a comma-separated string of numbers to a list of floats. Afterwards, the RDDs are again transformed such that they fit the input dimensions of the LSTM model. The last transformation is the inference functions, which loads the corresponding model and evaluates the data on basis of the model. In Spark there is a distinction between so called transformations and actions. Transformations are not evaluated, unless an action is called upon the pipeline of transformations. In this case, the corresponding action is saving the results to a file. This action triggers the whole pipeline. The results are again saved back to HDFS.

CPU AND GPU SERVER SETTINGS

For the experiments on the server using CPUs and GPUs a python environment needed to be implemented. The Linux server has 32 CPUs with 384 GB of RAM and also two GPUs with 16 GB of RAM. For the training and loading of the models the Keras library using the TensorFlow backend is used. TensorFlow needs the CUDA Toolkit¹⁵ for the usage on GPUs. Instead of the standard Keras LSTM layer a CuDNNLSTM layer is used. Each transformer model weights are stored in its directory and are loaded to the LSTM model architecture. In order to load the model weights, a LSTM model architecture used for each transformer is generated as can be seen in Figure 4.15. At the beginning, there are two input layers which describes the actual measurement values (input_1) on the one side and the other side all the other features, mainly the weather forecasts (input_2). These both input time series are input for two LSTM layer which use as activation function the Leaky Rectified Linear Unit 'LeakyReLU'. The output hidden states from these LSTMs are concatenated and are input to a Dense Layer. Another Dense and a Dropout Layer follow this. The Dropout Layer and the recurrent dropout are used for regularization and to prevent overfitting in the LSTMs. Moreover, at last a fully connected output layer is used.

This architecture differs from the result of the master thesis, Abdullayeva (2019), which was a 'Univariate Stacked' LSTM. This is mainly due to the different data that could be used as input for the model. On the one hand the real vertical load measurements could be used and on the other hand weather forecast were available. Thus, a multivariate LSTM was used instead of a univariate approach and first tests with the real data showed promising results. It is also assumed that the use of weather forecasts instead of weather measurements is of great benefit in terms of reducing the overall error of the vertical power flow forecast.

After loading the model weights of each transformer, the corresponding input data of each transformer are loaded for one time step. The forecast results of 48 hours are finally saved in a CSV file back to each transformer

¹⁴ <https://spark.apache.org/docs/latest/api/python/pyspark.html>

¹⁵ <https://developer.nvidia.com/cuda-toolkit>

directory. A general applicability of the CUDA and CuDNN¹⁶ libraries to parallel processes is no trivial process which should be considered when planning to use it. This is especially true for the usage in a near real time scenario. Therefore, some additional research was done in order to find a solution for the usage of GPUs in general in a near real time operation. The idea is that it might not be efficient to use GPUs for the online evaluation of the forecast models due to the overhead of loading and saving the data to the GPU. The usual more efficient way of using GPUs is for training the models, which takes much time and can be accelerated by using GPUs. In the works of (Maceina, 2017), (Elliot, 2011) and (Yang, 2018), it can be seen that parallel processing on a GPU in a near real time operation is not trivial and probably needs further investigation for the best and efficient handling. The main issues are due to scheduling problems as well as GPU concurrency issues which may block other processes once one process was started (Elliot, 2011 and Yang, 2018). However, these issues need to be solved with the cuDNN library. Thus, in the end there need to be done further investigation on the proper usage of the cuDNN library for the parallel processing of forecast tasks using LSTM DNN models. One approach is the usage of the TensorRT Software Development Kit¹⁷, which is specialised in the optimization and acceleration of deep learning inference.

¹⁶ <https://developer.nvidia.com/cudnn>

¹⁷ <https://developer.nvidia.com/tensorrt>

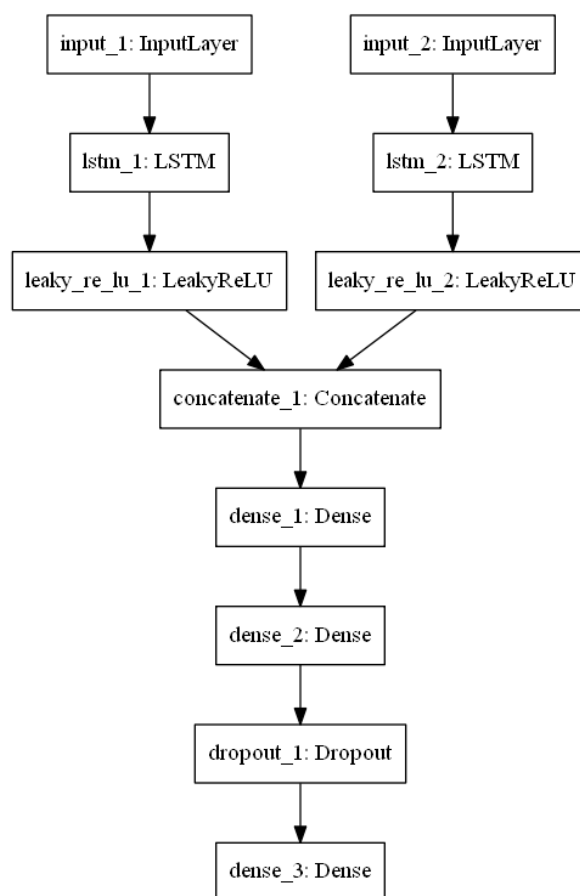


FIGURE 4.15 DEEP NEURAL NETWORK ARCHITECTURE WITH LSTM LAYERS

4.3.3.2 DISCUSSION ON INNOVATION

In this project, there is the chance to evaluate data and processes from a near real time system. As the constraints for such systems are, among others, to deliver results in time and to have stable systems, big data tools combined with Artificial Intelligence (AI) are investigated. big data in the Energy Domain is not well established and the aim is to gather experience in this area as well as for the parallelization processing for loading and calculating a high number of Keras models used in an online system for time series forecasting.

4.3.4 RESULTS AND CONCLUSIONS

In the following results for the three different approaches are presented. At first, results for the Hadoop cluster for the preliminary experiments with an open source data set are shown. Results for the evaluation of the processing time for delivering forecast values for a high number of transformer stations calculated on the Hadoop cluster are shown and compared to the same evaluation on a server with up to 32 CPUs which is then supplemented by adding GPUs. Finally, a conclusion for the results completes this paragraph.

4.3.4.1 RESULTS

PRELIMINARY EXPERIMENT

In order to understand the scaling properties of the cluster, preliminary experiments were carried out, since the forecast models from WP6 that should be used were not yet available. Thereby, a simple LSTM model is evaluated on test data at hand at this moment (in this case the MNIST Dataset¹⁸). As a baseline served the evaluation on a single server, i.e. without parallelization and big data tools. The hypothesis is that with increasing data size, the implementation with the help of big data tools gains an increasing advantage in time for computations. However, as can be seen in Figure 4.16, this was not the case. A more detailed explanation of this behaviour is discussed below.

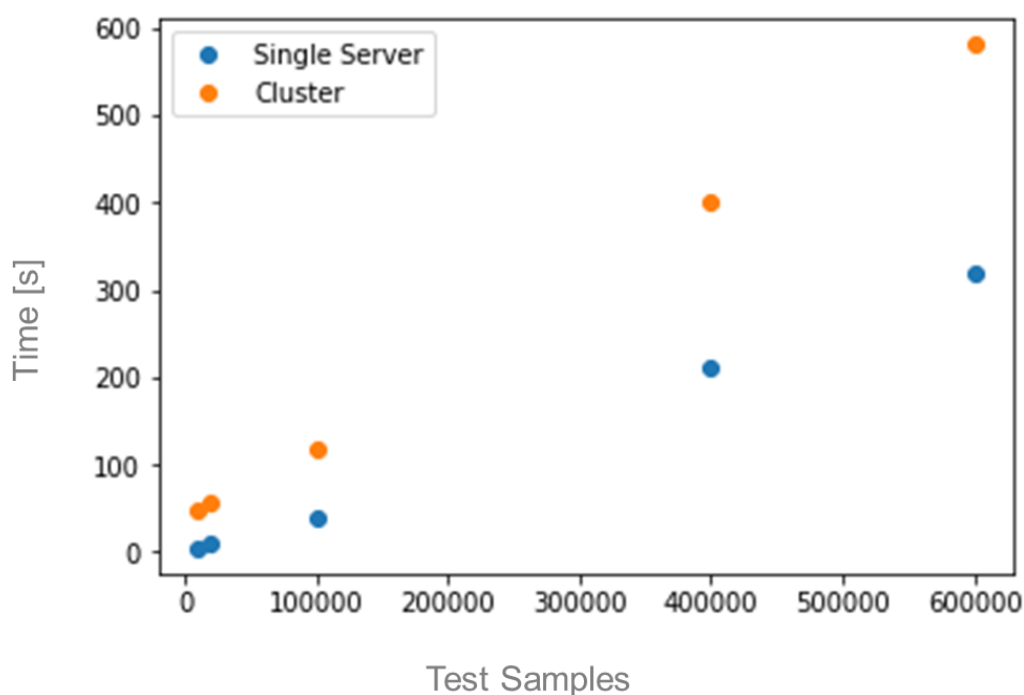


FIGURE 4.16 SCALING RESULTS OF EVALUATING A LSTM MODEL ON THE MNIST TEST DATA¹⁹

As can be seen, the data size increases up to 600 000 data samples. This is ten times the original size of the MNIST dataset and in turn means evaluating the LSTM model on 600 000 data samples. The single server outperforms the cluster in each case. One also has to have in mind the already mentioned feature called “locality awareness” of Yarn, which makes sure that the computations are shipped to the data and not vice versa. Therefore, as the

¹⁸ <https://keras.io/datasets/>

¹⁹ The processing time for evaluation the model is plotted over the increasing test samples as input to the model

test data here is already saved in HDFS, there is no network traffic of data involved and the computations are carried out on the node, where the data resides.

To understand the results, one has to look at the resources available on the big data cluster. There are only five nodes available, which in turn have limited resources. On each node, the services from the big data tools have to run. This is for example the master node for HDFS and YARN, but also from other tools, which are also installed on the cluster (to be precise: this a Hortonworks²⁰ cluster). In the right column of Figure 4.14 the number of components (services) which are installed on the corresponding node can be seen. Therefore, not many resources are left for the actual computations. Additionally, setting up all services needed for the evaluation takes time. This means that there is much overhead involved in evaluating just a single data sample in the big data cluster, as both Spark, Yarn and HDFS have to communicate and Yarn has to schedule resources in the cluster, thereby waiting for other nodes to verify that the resources are available.

FINAL EXPERIMENTS

For comparing the results of the three different approaches, the setting for the experiment is comparable. The setting is as follows: each transformer station has LSTM model. Each model needs to be evaluated on one data point. Therefore, scaling in this experiment involves not only the data size, but also the models loaded. Time is taken for the following transformer station counts: 1, 10, 100, 585, 1'000, 3'000 and 10'000. The count of 585 comes from the current available models and around 3,000 will be included at a later stage of the overall project. In order to make the timing results more reliable for each count number, the evaluation was repeated ten times on the CPU and GPU server and a mean value was calculated.

Furthermore, the number of used CPUs was set to: 1, 2, 5, 10, 20 and 32 and the duration for the computation on these different numbers is evaluated. Once the right setting of the parallel processes for the forecasting was achieved, the parallel computation could obtain much better results compared to the Hadoop cluster. In Figure 4.17 these results are summarized.

²⁰ <https://de.cloudera.com/products/hdp.html>

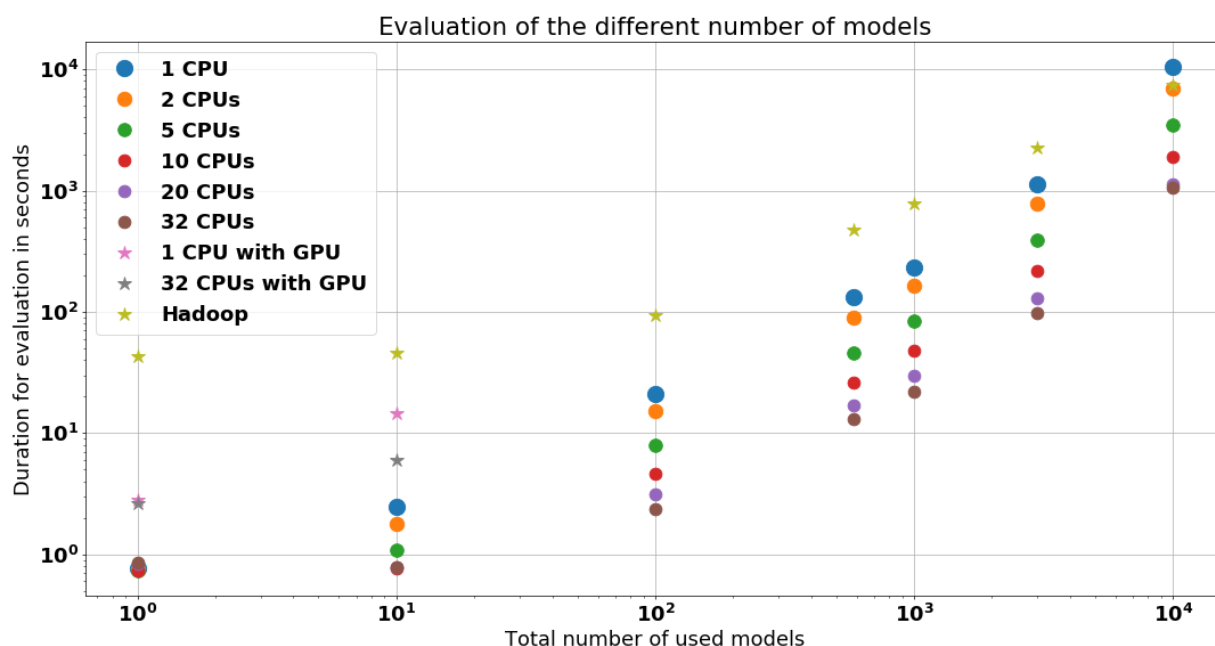


FIGURE 4.17 EVALUATION OF THE MEAN PROCESSING TIME FOR 1 TO 10,000 MODELS FOR 48H FORECAST

Compared to the Hadoop cluster the results for one and two CPUs are similar. However, with an increasing number of CPUs the processing time for the evaluation of the different number of models surpasses the results for the Hadoop cluster. Especially for a number of models below 1,000, the computation time is already lower on one CPU. The results for 1,000 and 3,000 models are particularly interesting because the number of models corresponds to the actual number of about 1,415 available transformers. Since not only the forecast of the active power but also the forecast of the reactive power is required, twice the number of models also needs to be considered. For the delivery of the residual load forecast the deployment will take place on a server of about five CPUs. With the focus on these five CPUs a processing time for 1,000 models will be around 1.4 minutes and for 3,000 models around 6.6 minutes, which again can be seen in Figure 4.17. These periods are less than the necessary 10 minutes. For the evaluation on at least 20 CPUs the results for a model number below 3,000 are in a similar range. For a model number below 585 models, it seems that it does not matter how many CPUs are chosen, all results are below half a minute and for the model number lower equal 100 even below 0.25 minutes. Nevertheless, 10,000 models still require a processing time of about 58 minutes for 5 CPUs and this only goes down to 18 minutes for 32 CPUs. This concludes that with an increasing number of models, the results show that other solutions are necessary.

Figure 4.17 shows comparison with the addition of one GPU is made for one and 10 models. It also shows that the results are not as expected, as the processing time takes even longer with an additional GPU. Due to software incompatibilities between the software versions of TensorFlow, CUDA and CuDNN, the other models could not be calculated. The solution of this kind of incompatibility issues is not part of this work and must be done elsewhere.

4.3.4.2 CONCLUSIONS

CHALLENGES

Big data cluster: Although the results evaluated on the Hadoop cluster are not as good as expected, it is not surprising due to the limited resources available on the cluster. With the big data cluster come some services that run on the nodes and need to use resources (e.g. master node for HDFS). In general, there indeed needs to be a certain amount of data such that it is worth to distribute the workload on a cluster. The results for the evaluation time for one data point is about 42 seconds and about 45 seconds for 10 data points. It shows how many overhead is involved in just starting a job on the cluster. For evaluating a single data point, it is not worth to initiate all this overhead. Tensorflow was installed in a virtual environment and shipped with every job, in a way to ensure same requirements are met on each node. Therefore, plain Tensorflow and Keras was available on every node and could be used within Spark. The model was loaded from HDFS and the predict method called from within a Spark transformation. This environment has to be shipped with each computation and is therefore costly.

GPU: The computation on GPUs with Keras LSTM models is not that straight forward as assumed. Due to software incompatibilities between the software versions of TensorFlow, CUDA and CuDNN, some experiments could not be evaluated. For the experiments which took place, the evaluation time was slower than by using CPUs. The main issue here, is the evaluation in a near real time scenario. GPUs are accelerating the training process of deep neural networks with massive datasets, since the process involving multiple matrix/tensor multiplication can efficiently be parallelized. For the near real time scenario, a trained model was applied to online data which is compared to a training data set only a small data set.

FINAL CONCLUSION

This work demonstrated the residual load forecasts calculated in the forecast system of the German demonstrator (WP6) can be generated with all three approaches. The scaling up to 10000 forecast models could take place for the Hadoop cluster and the standalone CPU server. For the GPU server software incompatibilities limited the number of scaling experiments. Regarding the comparison between the different approaches the evaluation on the Hadoop cluster and the GPU did finally not outperform the usage of CPUs in the near real time scenario. For the German demonstrator the objective is the delivery of about 3000 forecasts (including active and reactive power) under 15 minutes. It could only be achieved using the stand-alone server with 5 CPUs. For the delivery of the residual load forecast the deployment will take place on a server of about 5 CPUs. On such a server, the time needed for the evaluation of 3000 models is around 6.6 minutes. When using more than 5 CPUs, the time is less than 5 minutes for a model number not exceeding 3,000. However, with an increasing number of models, the results show that other solutions are necessary. For 10,000 models, a processing time of about 58 minutes is still required for 5 CPUs and 18 minutes for 32 CPUs.

Much experience has been gained in handling and using the Hadoop cluster to produce many forecasts in a real-time environment. In general, a larger cluster is needed and it also needs to be evaluated whether the data is

large enough to require big-data tools, or, as shown here, whether a regular standalone CPU server with a sufficient number of CPUs is more practical. As already mentioned, the time required for implementations is relatively large and requires additional knowledge and experience.

In summary, we could gain much experience for the processing in a near real time scenario and a procedure has also been defined as to how the large number of forecasts can best be delivered in sufficient time. Furthermore, there are several possibilities for dealing with the challenges as mentioned above in the future. On the one hand, the usage of a Kubernetes cluster instead of the Hadoop cluster for scaling tasks could be considered. Another possibility seems to be the Spark Streaming API²¹, which can help to speed up the processes. For further studies it would also be interesting to analyse the comparison between YARN and MESOS²², which can schedule the computation not only with a Hadoop-like cluster but also a cluster of GPUs. Additionally, a comparison of the big data and ML tool BigDL²³ (distributed deep learning on Apache Spark) with deep learning libraries like Keras, Tensorflow and PyTorch would be interesting. Finally, a comparison using the architecture of B1 could be an interesting next step.

On the other hand, the bottleneck of the loading and evaluation of the Keras LSTM models could be done more efficient, e.g. with caching in the future. For the GPU approach, the transferring of data needs to be separated from the evaluation process, so that only the computation of the deep neural network can be done on the GPU. Another possible way to investigate this further, is the usage of the TensorRT Software Development Kit, which is specialised in the optimization and acceleration of deep learning inference.

4.4 DATA EXCHANGE BETWEEN DSO AND TSO

Main section authors: Kalle Kukk (Elering), Wiebke Albers (innogy), Jan Budke (innogy), Camen Calpe (innogy), Maik Staudt (Mitnetz Strom)

4.4.1 ABSTRACT

The rising need for system flexibility creates new requirements for data exchanges. Besides flexibility providers it mostly concerns DSOs and TSOs, as more and more flexible resources are connected to the distribution grid.

Several clauses can be identified in both pre and post-Clean Energy Package (CEP) EU regulations, which concern DSO-TSO data exchange for flexibility usage. Regulations approved already before CEP include several network codes. CEP itself has resulted in amended electricity market directive and electricity market regulation, potentially followed by new network codes and implementing acts still to be established.

²¹ <https://spark.apache.org/streaming/>

²² <http://mesos.apache.org/>

²³ <https://software.intel.com/content/www/us/en/develop/articles/bigdl-distributed-deep-learning-on-apache-spark.html>

This chapter provides two approaches based on EU-SysFlex demonstrators in the context of EU regulation in terms of data exchange for flexibility usage – German demonstrator and Flexibility Platform demonstrator.

The increasing share of renewable energy sources (RES) creates a rising demand for active and reactive power flexibilities in Germany, especially for congestion management and voltage control. The need for the evolution of DSO-TSO data exchange is recognized by multiple players in the German electricity sector and goes beyond the EU-SysFlex project. Hence, multiple German TSOs and DSOs work jointly on a TSO/DSO data exchange approach within a project called “connect+”, with which the German demonstrator’s principles are in line. In this approach, each system operator selects needed flexibilities to solve congestions in its grid (subsidiarity principle) and determines the maximum flexibility potential for the upper system operator. Only the grid data relevant for re-dispatch, including costs, sensitivities, and flexibility limitations, is exchanged with the upstream system operator. The connecting system operator initiates the flexibility activation in his grid based on his own need and the received request by the upstream system operator.

The starting point of Flexibility Platform demonstrator is to maximize the liquidity of and to allow easy access to the flexibility market through a single flexibility market concept. Such concept implies massive flows of data: in terms of a number of stakeholders and services/products as well as in terms of data granularity and up to very-near-real-time exchanges. In a single market, several market places or market platforms can coexist and even compete with each other, and therefore, it is essential to ensure interoperability. Also, the single market concept requires the involvement of all stakeholders, obviously including TSOs and DSOs. TSO-DSO data exchanges result from the need to ensure that flexibilities are procured and activated most efficiently, including a case of joint procurement and that all flexibilities have access to the market regardless of where they are physically connected.

4.4.2 INTRODUCTION

4.4.2.1 AIM

The chapter aims to provide two approaches to TSO-DSO data exchange based on EU-SysFlex demonstrators in the context of EU regulation in terms of data exchange for flexibility usage.

4.4.2.2 OVERVIEW

First, the overview of several EU legal acts is given. The focus is on specific articles which concern DSO-TSO data exchange for flexibility usage. The interpretation of these articles in the context of the EU-SysFlex is presented.

Second, two approaches from EU-SysFlex are explained:

- WP6 German demonstrator (The need for an evolution of TSO-DSO data exchange is explained from the German demonstrator perspective. The TSO/DSO data exchange approach for congestion management and voltage control presented in this chapter goes beyond the EU-SysFlex demonstrator and explains

principles of a TSO-DSO data exchanged worked on by multiple TSOs and DSO jointly within a project called “connect+”, with which the German demonstrators' principles are in line.

- WP9 Flexibility Platform demonstrator

These demonstrators were selected because they are different. Both are in line with the current regulation and future vision. The need for the evolution of DSO-TSO data exchange for flexibility usage is elaborated from the perspective of both cases. Then the concept of data exchange of each is explained.

4.4.3 CONTEXT AND APPROACH

4.4.3.1 CURRENT REGULATION REGARDING DSO-TSO DATA EXCHANGE FOR FLEXIBILITY USAGE

Several clauses can be identified in both pre and post-Clean Energy Package (CEP) EU regulations, which concern DSO-TSO data exchange for flexibility usage more or less explicitly (see Annex IV – Data exchange between DSO and TSO: SUPPLEMENTARY INFORMATION ABOUT EU REGULATIONS). Regulations approved already before CEP include several network codes. CEP itself has resulted in amended electricity market directive and electricity market regulation, potentially followed by new network codes and implementing acts still to be established.

Guideline on System Operation (SOGL) and Key Organisational Requirements, Roles, and Responsibilities (KORRR) established according to article 40 of SOGL regulate data exchanges between TSOs, DSOs, and grid users. Data addressed in these legal texts concerns structural, scheduled, and real-time data. This data address a wide variety of data exchanges relevant to system operation, including flexibility management. However, this report focuses mainly on flexibility data exchanges between TSOs and DSOs. Therefore, articles regulating flexibilities connected to the DSO grid are presented in detail as only these may require explicit TSO-DSO data exchange.

To mention few “generic” articles of SOGL and KORRR which involve some elements of flexibility data exchanges it should be referred to the ones explaining structural data exchange between TSOs and DSOs, real-time data exchange between TSOs and DSOs, real-time data to be provided by DSO-connected generators, scheduled data exchange between TSOs, DSOs and DSO-connected generators, notification of schedules.

Other network codes considered include Guideline on Electricity Balancing (EB GL), Network Code on Demand Connection (DCC), and Network Code on Requirements for Grid Connection of Generators (RfG NC).

As a result of CEP, the electricity market directive sets new rules for interoperability and for access to and management of end-customer related meter and other data, which is necessary for the functioning of markets, including flexibility market. Electricity market regulation encourages cooperation between TSO and DSO, which would facilitate data exchanges useful for different flexibility processes.

A detailed overview of concerned articles is presented in Annex IV – Data exchange between DSO and TSO: SUPPLEMENTARY INFORMATION ABOUT EU REGULATIONS. More comprehensive insight into EU regulations in

terms of data management, in general, has been investigated in Task 5.1 report European Level Legal Requirements to Energy Data Exchange.

The following listing provides interpretations of the articles in the Annex where the named EU regulations address TSO-DSO data exchange for flexibility usage:

Guideline on System Operation (SOGL) and All TSOs' proposal for the Key Organisational Requirements, Roles and Responsibilities (KORRR)

- Structural data (e.g., capacity of substations) may be required for flexibility optimization, prediction, and prequalification (SOGL articles 48, 53; KORRR article 3 and others).
- Real-time data (e.g., aggregated generation/demand data in the DSO area, active and reactive power flows of DSO-connected generators) may be required for flexibility activation and optimization (SOGL article 53; KORRR article 3 and others).
- Scheduled data (e.g., forecasted scheduled active power output) may be required for flexibility optimization, prediction, prequalification, baseline calculation (SOGL article 53; KORRR article 3 and others).
- FCR, FRR, and RR technical minimum requirements may concern DSO-connected flexibilities. Technical requirements are needed for flexibility activation and flexibility baseline calculation (SOGL articles 154, 158, 161, 182).
- FCR, FRR, and RR prequalification processes may involve the prequalification of DSO-connected flexibilities (SOGL articles 155, 156, 159, 162, 182).

Guideline on Electricity Balancing (EB GL)

- DSO provides all necessary information relevant for imbalance settlement to the connecting TSO (article 15). This could include data for flexibility baseline calculation and flexibility verification.
- Information relevant for flexibility prequalification, bidding, activation (SOGL refers to 'operation'), baseline calculation and verification (SOGL refers to 'imbalance calculation' and 'evaluation') must reach TSO and reserve connecting DSO (articles 16, 18).

Network Code on Demand Connection (DCC)

- Demand units and aggregated demand, including the ones connected to DSO, should be able to receive instructions from relevant DSO and TSO to modify their demand and to transfer the necessary information (article 28).

Network Code on Requirements for Grid Connection of Generators (RfG NC)

- Type B and type C power-generating modules²⁴ should be able to exchange information useful for different flexibility processes with relevant DSO and TSO. Where applicable, TSO and DSO should sign an agreement for data exchange required (articles 14, 15).

Directive on Common Rules for the Internal Market in Electricity

- TSO and DSO can ensure easy access to demand response data relevant for different flexibility processes, where they have the role of distributing meter data to third parties operator (and where this role is not assigned to other roles like Metering Data Operator or Smart Meter Gateway Operator or Data Hub Operator or Data Exchange Platform Operator²⁵) (article 17).
- Management of smart meter data relevant for different flexibility processes may involve both TSO and DSO, where they have the role of distributing meter data to third parties (article 20).
- Management of final customer data relevant for different flexibility processes may involve both TSO and DSO, where they have the role of distributing meter data to third parties (article 23).
- Interoperability is relevant for TSO-DSO data exchanges in all flexibility processes, including the retail market, where they have the role of distributing meter data to third parties (article 24).

Electricity market regulation

- TSO and DSO agreement for data exchange is useful to facilitate the cooperation for mutual information sharing relevant for different flexibility processes (article 57).

4.4.3.2 NEED FOR EVOLUTION OF DSO-TSO DATA EXCHANGE FOR FLEXIBILITY USAGE

WP6 German demonstrator view

The German demonstrator calculates the possible flexibility range of active and reactive power at grid nodes at the DSO-TSO²⁶ interface so that the TSO can access these flexibilities as ancillary services for congestion management and voltage control in the transmission grid without harming the distribution grid. For that purpose, renewable energy sources (RES) connected to the distribution grid are integrated into the schedule-based process for congestion management and voltage control in the transmission grid while considering the interdependencies of active and reactive power flexibilities.

The increasing share of RES and the additional delays in Germany's planned grid expansion projects make TSOs face challenges in day-to-day grid operation and operational planning. It has become increasingly common that

²⁴ According to article 5 of RfG NC the type B and C power-generating modules' connection point is below 110 kV and maximum capacity at or above a threshold proposed by each relevant TSO and approved by the relevant regulatory authority or Member State.

²⁵ Data Hub Operator and Data Exchange Platform Operator are roles proposed by EU-SysFlex WP5.

²⁶ In Germany the interface DSO-TSO is at the HV/EHV transformer between the 110kV and 220kV or 380kV level.

TSOs are forced to re-dispatch in order to relieve grid congestions. In situations with high feed-in of RES, a sufficient number of power plants suitable for re-dispatching in operation is not always available. Instead, RES feed-in, mostly connected to the distribution grid, is curtailed as emergency measures, resulting in compensational payments and imbalances in the system. For that reason, the demonstrator tests how the curtailment can happen on a planned basis avoiding imbalances and resulting in improved TSO-DSO coordination and thus enhanced TSO-DSO data exchange.

Also, the usage of existing tools for voltage control depends on the availability of a sufficient amount of reactive power flexibilities in extra-high voltage (EHV). The decrease of generation at EHV level due to a higher share of RES at the distribution level also decreases the reactive power potential located in the transmission grid. Additionally, the limited coordination between TSOs and DSOs regarding reactive power management also leads to a non-efficient use of existing potentials for voltage control.

Since the provision of frequency reserve by generators at transmission-level is getting substituted by flexibilities at the distribution level. Since distribution grids reach their thermal limits more often due to the increasing simultaneity of RES and new loads (such as electric cars, heat pumps), the use of frequency reserve can cause congestions at the distribution level. Thus, DSOs are forced to carry out counteractions, which in return can reduce or even eliminate the effect of the frequency reserve provision. For that reason, DSO-TSO coordination is needed to keep the DSO grid congestion-free when TSOs need to activate flexibilities to solve system imbalances.

Summarized, the main drivers considered in the German Demonstrator are external, namely the increased share of RES, especially volatile RES like wind, not located close to the demand sites and increasingly connected to the distribution grid, which leads to a structural change in the power system. This effect leads to a higher demand for congestion management and, at the same time, to a shortage of re-dispatch and reactive power potential in the transmission grid. Uncoordinated measures of TSOs and DSOs can lead to counteracting measures and system imbalances. From the previously named drivers, increasing demand for active and reactive power flexibilities for congestion management and voltage control, as well as the improvement of TSO-DSO coordination arises. Such coordination must include a cost-efficient process of flexibility selection, considering all constraints of TSOs and DSOs and the sensitivities of the flexibilities towards the congestion or voltage problem.

WP9 Flexibility Platform demonstrator view

The WP9 starting point is to maximize the liquidity of and easy access to the flexibility market. It means ideally pan-European flexibility market with lots of both service exchanges (i.e., flexibility) and data exchanges. Intermediate steps towards that can take the form of regional or national markets. However, more than one flexibility buyer and more than one product should be present (not to mention several flexibility sellers/providers).

A single market is not necessarily a single market place or single market platform. Several platforms can coexist and even compete with each other. It is merely essential to ensure the interoperability of them because:

- Same buyers can be active in several market places
- Same providers can be active in several market places by offering same flexibilities
- Same products in same timeframe can be traded in several market places
- However, the system is one.

Indeed, such complexity implies massive flows of data – in terms of a number of stakeholders and services/products as well as in terms of data granularity and up to very-near-real-time exchanges.

Single market concept requires the involvement of all stakeholders, obviously including TSOs and DSOs. TSO-DSO aspects need consideration in several aspects:

- 1) Both as flexibility buyers are kind of competing with each other.
- 2) They should pursue socio-economic efficiency – thus, optimization is needed.
- 3) Joint procurement of flexibilities should be allowed where a single bid is used for both TSO and DSO needs.
- 4) All flexibilities should have free access to the market regardless of which network they are physically connected.

Customers (i.e., flexibility service provider), not TSO or DSO, should have the control and freedom in providing flexibility. They usually rely on economic justifications and does not undermine system security to be ensured by system operators (SOs).

Efficient tools are needed for both market operations (e.g., clearing, settlement, activation orders) and data exchanges.

4.4.4 RESULTS AND CONCLUSIONS

4.4.4.1 JOINT APPROACH FROM GERMAN TSO AND DSO AND PRINCIPLES OF GERMAN EU-SYSFLEX DEMONSTRATOR OF TSO-DSO DATA EXCHANGE CONCEPT FOR FLEXIBILITY USE

The need for evolution of DSO-TSO data exchange is recognized by multiple payers in the German electricity sector. Therefore not only the German EU-SysFlex demonstrator is designing and testing a TSO/DSO data exchange approach for congestion management as well as voltage control²⁷. Hence, multiple German TSOs and DSO work jointly on a TSO/DSO data exchange approach within a system operator's project called "connect+"²⁸, with which the German demonstrators principles are in line. The designed data exchange approach relates to a future prove re-dispatch process for the use of flexibilities at distribution level to solve distribution and

²⁷ Approach is designed for congestion management and voltage control. Focus is not on frequency products. Requirements for frequency products might be different from the presented approach.

²⁸ www.netz-connectplus.de

transmission congestions. The in the following presented approach, therefore, goes beyond the EU-SysFlex project and explains principles of a TSO-DSO data exchanged worked on by many German TSOs and DSOs jointly within the “connect+” project. The general approach for this process is visualized in the following figure. More information regarding the process as well as the connect+ project can be found in the connect+ report²⁹.

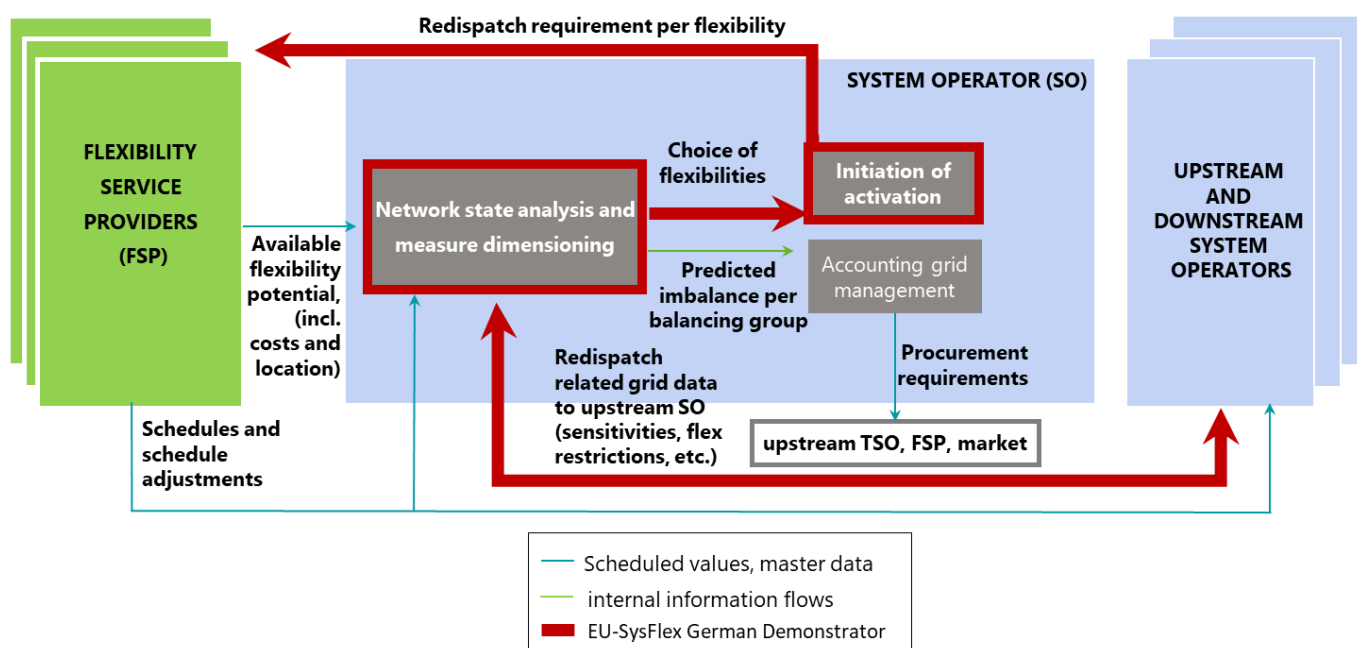


FIGURE 4.18 REDISPATCH PROCESS APPROACH BASED ON CONNECT+ APPROACH AND IN LINE WITH GERMAN EU-SYSFLEX DEMONSTRATOR

Flexibility service providers continuously or cyclically send the schedules or schedule adjustments to the system operators. Furthermore, they send the available flexibilities from their resources to the system operators, which includes the cost and the location of the flexibilities. Based on this, the system operator executes his network state analysis and measure dimensions. The latter includes the selection of appropriate flexibilities to solve congestions in its grid and calculate the possible flexibility potential for the upper system operator.

The re-dispatch relevant grid data is exchanged with the upstream system operator, which includes sensitivities and flexibility limitations. Based on the results of the own measure dimensioning of needed flexibilities and the requests of flexibilities from external system operators, the connecting system operator initiates the activation of the needed re-dispatch flexibility by sending the re-dispatch requirements per flexibility to the respective flexibility service providers.

²⁹ <http://netz-connectplus.de/wp-content/uploads/2020/04/BerichtInitialeAnforderungen.pdf>

In addition to the activation, the predicted imbalance within the balancing group is passed to the accounting grid management that takes care of any needed procurement requirements. Furthermore, the approach includes the settlement process, which is not further described in this document.

The sub-process of network state analysis and measure dimensioning is a core part of the German demonstrator within EU-SysFlex. Furthermore, the sub-process of the initiation of flexibility activation is also considered within the demonstration.

This approach of a coordination mechanism between TSOs and DSOs for the re-dispatch process has the advantage that it builds upon existing TSO-DSO data exchange (such as for forecasted power exchange) along with the natural structure of the grid, being resistant against blackouts and ensuring adequate responsibility allocation.

4.4.4.2 WP9 FLEXIBILITY PLATFORM CONCEPT OF DATA EXCHANGE FOR FLEXIBILITY USE

WP9 demonstrates data exchanges of flexibility marketplace ('Flexibility Platform') while making some necessary assumptions in terms of business processes and market design. This demonstrator relies on data system use cases (SUCs) described in Task 5.2.

The concept for the use cases stands on two legs:

- 'market platform' for managing flexibility business processes.
- 'data exchange platform' for managing any data exchanges

Overview of relevant SUCs for flexibility use to be demonstrated in WP9:

1. Flexibility prediction

Flexibility products are described as either slow (e.g., manual frequency restoration reserve (mFRR) and the UK short term operating reserve (STOR)) or semi-fast (e.g., automatic frequency restoration reserve (aFRR)) or fast (e.g., frequency containment reserves (FCR) and fast frequency response (FFR)) and can provide services for balancing and congestion management at local and national levels for TSOs and DSOs. The assessment of flexibility availability is split into three timeframes: investment planning (3+ years ahead) aims to understand future availability and if the predictions highlight insufficient capacity that needs addressing; operation planning (days to years ahead) aims to predict the short, medium and long term availability of flexible products that have committed to providing service; real-time planning (intraday operation) aims to predict the current availability of flexible products for balancing and congestion management requirements for that day.

2. Flexibility prequalification³⁰

³⁰ In WP5 two alternatives of this SUC are developed with different underlying business process assumptions. WP9 demonstrator focused on testing one of these alternatives.

The use case describes the process of pre-qualification of the flexibility service providers (aggregators and individual consumption, generation, and storage units) before they can make bids to the market and be activated. Prequalification involves both ‘product’ and ‘grid’ prequalification. Thereby the Flexibility Platform acts as a uniting element, which gathers flexibility needs provided by system operators as well as flexibility potentials provided by flexibility service providers (FSPs). For grid prequalification, coordinated actions with optimization operator or system operators are required for ‘grid validation’ process.

3. Flexibility bidding³¹

The use case describes the bidding process ending with a ranking of flexibility bids, which will then be activated by the system operator (see separate system use case for flexibility activation). Thereby the Flexibility Platform acts as a uniting element, which gathers and registers flexibility bids provided by FSPs. Before including bids in the merit order list, coordinated actions with the optimization operator or the system operator are required for the ‘grid validation’ process.

4. Flexibility activation³²

This use case describes data exchanges needed for the initiation of activation of flexibilities bids that previously have been sent to the Flexibility Platform. Delivery of notification of activation requests to the FSPs, in a reliable and timely manner according to the relevant terms and conditions applicable to FSPs. Right before activation, coordinated actions with optimization operator or system operator are required for ‘grid validation’ process. This use case does not apply to high-speed products in which the flexible units must react automatically to prescribed events in the system (like FCR product applied to immediate frequency deviations).

5. Flexibility baseline calculation

If a market participant bids flexibility in the flexibility market, the baseline consumption/generation of such market participant needs to be identified for the verification and settlement processes (see SUC ‘Verify and settle activated flexibilities’). There are two options for this:

- a) A market participant has to declare its power schedule (baseline) ex-ante in such a way to permit the system operator to implement the settlement processes. Such player (FSP) usually declares the baseline directly, but the system operator could provide specific tools to help market participants in the baseline definition, promoting market participation.
- b) Market operator (TSO or DSO or Flexibility Platform operator) itself calculates the baseline ex-post based on meter data. The methodology to calculate the baseline is transparent and public.

³¹ In WP5 two alternatives of this SUC are developed with different underlying business process assumptions. WP9 demonstrator focused on testing one of these alternatives.

³² In WP5 two alternatives of this SUC are developed with different underlying business process assumptions. WP9 demonstrator focused on testing one of these alternatives.

6. Verification of activated flexibilities

The actual flexibility delivered is calculated as the difference between baseline and metered consumption/generation of that FSP. The verification takes place by comparing the delivered flexibility and flexibility requested by the system operator. A settlement means that an FSP is asked for a penalty if delivered flexibility is less than the requested flexibility. The imbalance settlement process follows but is out of the scope of this use case.

These use cases will be tested in Flexibility Platform demonstrator.

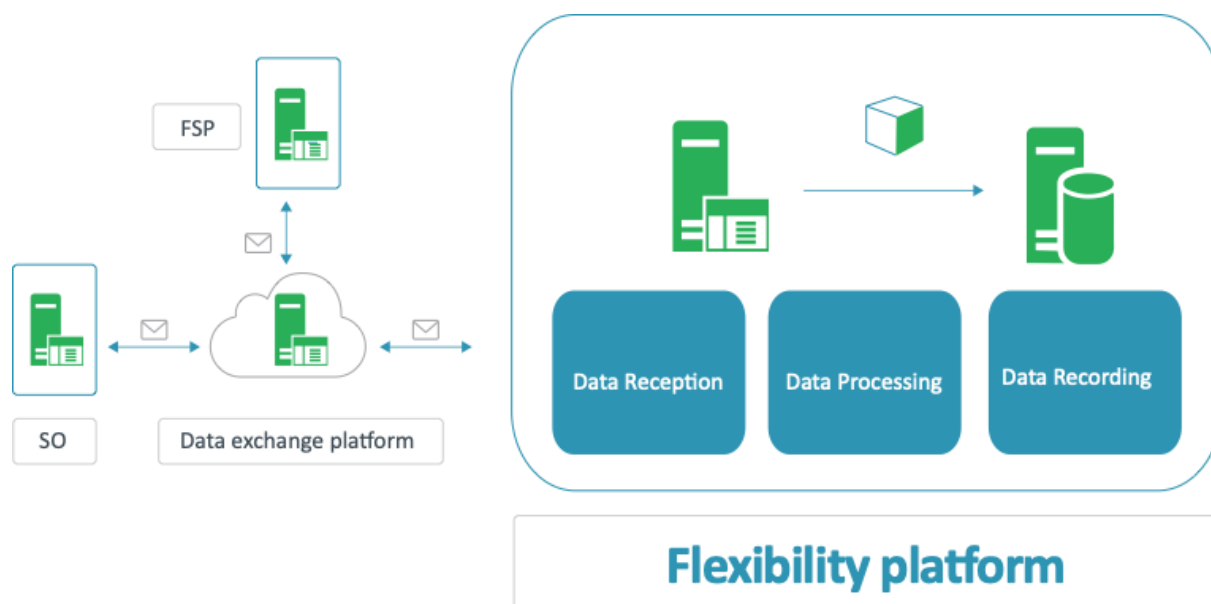


FIGURE 4.19 HIGH-LEVEL ARCHITECTURE OF FLEXIBILITY PLATFORM CONCEPT

A list of flexibility related processes that are conducted via market platform ('Flexibility Platform'). These processes involve:

- 1) Register flexibility need
- 2) Register flexibility potential
- 3) Send necessary information for grid impact assessment to system operator/optimization operator – grid constraint check-in prequalification phase
- 4) Collect the results of grid impact assessment in the prequalification phase
- 5) Prequalify FSP based on information provided by FSP in its 'flexibility potential' and based on results of grid impact assessment
- 6) Set 'long-term restrictions' to the FSPs not passing constraint check-in prequalification phase
- 7) Register 'long-term restrictions'
- 8) Publish prequalification results, incl. information about 'long-term restrictions'

- 9) Register flexibility call for tender opening
- 10) Receive flexibility bid
- 11) Send necessary information for grid impact assessment to system operator/optimization operator – grid constraint check-in bidding phase
- 12) Collect the results of grid impact assessment in the bidding phase
- 13) Set 'short-term restrictions' to the FSPs not passing constraint check-in bidding phase
- 14) Rank bids based on merit order principle (taking into account 'short-term restrictions')
- 15) Register flexibility bid
- 16) Register flexibility call for tender closure
- 17) Receive request for flexibility activation
- 18) Send necessary information for grid impact assessment to secondary system operator / optimization operator – grid constraint check in activation phase
- 19) Collect the results of grid impact assessment in activation phase
- 20) In case activation would cause grid constraint to select next set of bids for activation
- 21) In case activation would cause imbalance to send information for counteraction to an appropriate role (primary system operator assumed in a use case)
- 22) Forward request for activation to FSP
- 23) Register activation request
- 24) Register activation confirmation received from FSP
- 25) Receive data (meter data, sub-meter data, external data) for baseline calculation
- 26) Calculate baseline
- 27) Record baseline
- 28) Receive meter data for verification
- 29) Calculate delivered flexibility
- 30) Verify delivered flexibility
- 31) Send information for settlement

As a result of the discussions, the process 'grid impact assessment' / 'optimisation' in different phases (prequalification, bidding, activation) was considered to leave outside platform.

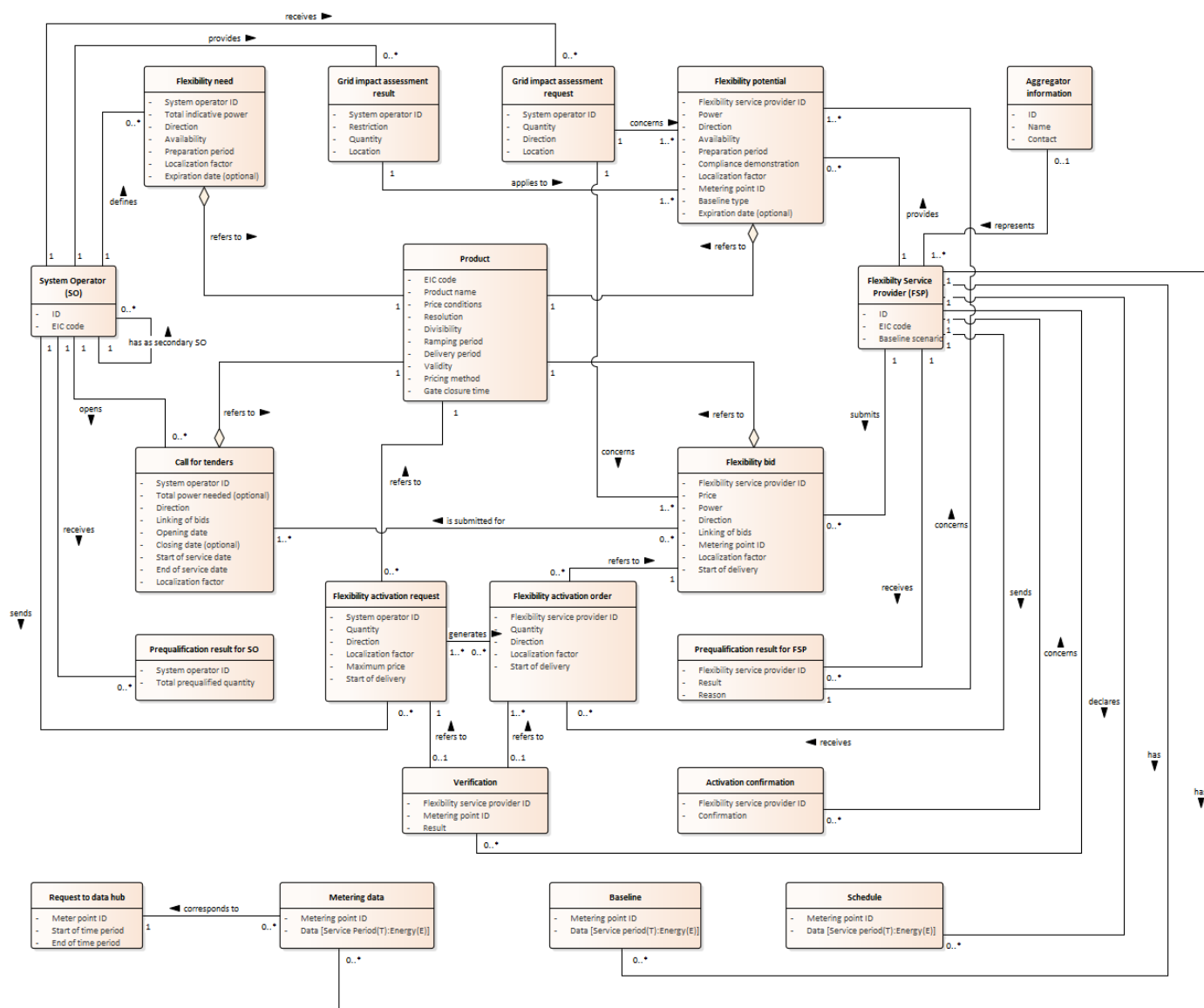


FIGURE 4.20 DATA MODEL OF FLEXIBILITY PLATFORM CONCEPT

Data exchange platform Estfeed will be applied to demonstrate the concept from data exchange perspective. Any data exchange will happen via Estfeed.

4.4.4.3 CONCLUSIONS

More and more flexible resources are connected to the distribution grid, and both DSOs and TSOs are in a rising need for system flexibility. This in turn implies new requirements for data exchanges. As analysed in this chapter, various EU regulations set a regulatory framework for flexibility data exchange already. However, at least these two examples of EU-SysFlex demonstrations indicate that there is a clear need for the evolution of DSO-TSO data exchange for flexibility usage. These two following demonstrator approaches define and test such DSO-TSO data exchanges:

WP6 German demonstrator

The increasing share of RES creates a rising demand for active and reactive power flexibilities in Germany, especially for congestion management and voltage control. Also it leads to an increasing need for enhanced TSO-DSO coordination, as uncoordinated measures of TSOs and DSOs can lead to counteracting measures and imbalances within the electricity system. For this coordination and the respective TSO-DSO data exchange, a cost-efficient process of flexibility selection, considering all constraints of TSOs and DSOs and the sensitivities of the flexibilities towards the congestion or voltage problem, is essential. These circumstances motivate the German EU-SysFlex demonstrator.

The need for evolution of DSO-TSO data exchange is recognized by multiple payers in the German electricity sector and goes beyond the EU-SysFlex project. Therefore, German TSOs and DSOs jointly work on designing the re-dispatch process and the therefore needed TSO/DSO data exchange for the use of flexibilities at distribution level to solve distribution and transmission congestions within the project called “connect+.” Such a process is also set up and tested within the German EU-SysFlex demonstrator.

The process and data exchange is based on the following principles:

- Data about availability of flexibilities from flexibility resources is continuously sent to the system operators.
- Each system operator selects needed flexibilities to solve congestions in its grid (subsidiarity principle).
- Each system operator determines the maximum flexibility potential for the upper system operator.
- Only the grid data relevant for re-dispatch, including costs, sensitivities, and flexibility limitations, is exchanged with the upstream system operator.
- The connecting system operator initiates the flexibility activation in his grid based on his own need and the received request by the upstream system operator.

WP9 Flexibility Platform demonstrator

The single flexibility market implies massive flows of data – in terms of a number of stakeholders and services/products as well as in terms of data granularity and up to very-near-real-time exchanges. In a single market, several market places or market platforms can coexist and even compete with each other. Therefore it is essential to ensure interoperability. Also, the single market concept requires the involvement of all stakeholders, obviously including TSOs and DSOs. TSO-DSO data exchanges result from the needs to ensure that flexibilities are procured and activated most efficiently, including a case of joint procurement (in this context a single bid to be used – simultaneously if needed – for more than one service by more than one system operator). Every flexibilities have access to the market regardless where they are physically connected.

‘Flexibility Platform’ demonstrator of WP9 of EU-SysFlex demonstrates elements of six core flexibility use cases, which were initially described in Task 5.2: flexibility prediction, flexibility prequalification, flexibility bidding, flexibility activation, flexibility baseline calculation and verification of activated flexibilities. As part of these use

cases, 31 flexibility related processes have been identified, which can all be realized through the same system – a market platform called ‘Flexibility Platform.’

Such a market platform should be able to satisfy quite complex actions in terms of quantity and quality. Many stakeholders (business roles) and software tools (system roles) need to exchange data with each other. In reality, some processes could be split into different systems, but this could even add complexity as these systems should be able to exchange data between themselves additionally. Furthermore, more than one market platform can exist in parallel, which means that the interoperability of data flows between platforms has to be ensured on top. In the WP9 Estfeed data exchange platform is used for secure data exchanges between any parties.

4.5 FORECASTING IN INTEGRATED ENERGY SYSTEMS

Main section authors: Alan Tkaczyk (UTartu), Stanislav Sochytskyi (UTartu), Gunay Abdullayeva (UTartu), Oleksandr Kurylenko (UTartu)

4.5.1 ABSTRACT

Demand Response (DR) mechanisms facilitate balancing the demand-supply ratio and provide greater flexibility within the electric grid. When the demand needs to be displaced, a DR event is activated on the market, and the amount of reduced electricity consumption is measured to assess the DR performance. The crucial part of the DR performance assessment is the evaluation of the business as usual, or baseline, load.

To investigate the stable neural networks, two use cases are studied for short-term multi-step ahead energy forecasting over multiple time series: 1) Long short-term memory (LSTM) based recurrent neural networks (RNN) were implemented by considering three types of LSTM based architectures. 2) Convolutional neural networks (CNN) were developed with various configurations. In both use cases, the results of the networks are compared to the Naïve and ARIMA benchmark models. Additional to these benchmark models, CNN-based models are also compared to the industry-standard baseline models (Asymmetric HFoT, SPFoT, Average, Daily Profile). RMSE, MAE, and MAPE evaluation metrics are used for the comparison of the models. To evaluate the robustness of the LSTM models more precisely, correlation and bias metrics are also calculated.

The conducted experiments have shown both LSTM and CNN based models outperform the baseline models in most time series. Stack LSTM and CNN-LSTM models show more stable results over all-time series.

These projects are two separate master's thesis works. In the research of the first use case, it has been collaborated with Fraunhofer IEE Institute for Energy Economics and Energy System Technology (Kassel, Germany) and the second project partners are AKKA Technologies and Enoco AS.

4.5.2 INTRODUCTION

Short-term load forecasting (STLF) is crucial for everyday operations in the energy sector. Accurate energy forecasting enables efficient operation of power systems, preservation of the balance between supply and demand, reduction of production cost, and management of future capacity planning. One of the cases considered in this chapter is DR mechanisms, aimed to balance the demand-supply ratio and provide greater flexibility within the grid. When there is a necessity to decreasing the demand, a DR event is activated, following energy consumption being decreased, and consequently, reducing the load on the energy grid.

An accurate electricity consumption forecasting is crucial for the evaluation of business as usual, or baseline, load, which, in turn, is required for the assessment of the amount of the energy consumption reduction during the DR event. To assess the amount of reduced energy consumption, the difference between the baseline load and actual load is taken. Accurate assessment of the consumption being reduced would prevent significant losses by customers and Demand Response Service Providers (DRSPs) due to under- or overestimation of the actual energy consumption reduction efforts. Moreover, STLF allows estimation of the availability of flexibility products shortly.

As STLF tasks usually involve vast amounts of data with erratic patterns, intrinsic to energy data, deep learning methods are of particular interest in this chapter. Deep learning methods are not only capable of capturing long-term dependencies, inherent non-linearity, and volatility in the data but also allow higher flexibility while working with big data frameworks, meanwhile demonstrating promising results in STLF as compared to various classical and industry-standard approaches.

In this chapter, the usage of deep learning methods is demonstrated in two use cases. The first use case focuses on the exploration of LSTM-based models on open-source data. In contrast, the second use case focuses on the exploration of CNN-based models on the data provided by a Norwegian EMS operator Enoco.

4.5.3 AIM

The main objective of this chapter is to investigate and develop various LSTM- and CNN-based architectures for short-term load forecasting, which would offer more robust forecasts as compared to the considered baseline models. Deep learning methods are capable of capturing long-term dependencies, inherent non-linearity, and volatility present in the energy data. Therefore, the aim is to demonstrate that LSTM- and CNN-based models are relevant and flexible enough to be used in the process of electricity consumption forecasting in real energy systems.

Three variations of LSTM (standard LSTM, stacked LSTM, and LSTM-based sequence-to-sequence architectures) are explored on six open-source datasets. The goal is to estimate the performance of the LSTM-based models using five evaluation metrics (RMSE, MAE, SMAPE, bias, and correlation coefficient) and to perform the comparison of the models' performance to baseline Naïve and Autoregressive Integrated Moving Average (ARIMA) models.

Two variations of CNN-based models (CNN and CNN+LSTM) are explored on electricity consumption data from three regions in Norway. The goal is to estimate the performance of the CNN-based models using three evaluation metrics: RMSE, MAE, and MAPE. It also required to perform the comparison of the models' performance to baseline Naïve and ARIMA + Fourier terms models, as well as to the industry-standard baseline models, such as Daily Profile, Asymmetric High Five of Ten (Asymmetric HFoT), Average, Similar Profile Five of Ten (SPFoT).

These projects are two separate master's thesis works. In the research of the first use case, it has been collaborated with Fraunhofer IEE Institute for Energy Economics and Energy System Technology (Kassel, Germany) and the second project partners are AKKA Technologies and Enoco AS.

4.5.4 USE CASE #1: INVESTIGATING ROBUST LSTM ARCHITECTURE FOR ENERGY TIME SERIES FORECASTING

4.5.4.1 METHODOLOGY AND APPROACH

For this study, both univariate and multivariate forecasting techniques for energy forecasting are considered. If the forecasting problem consists of one single series, it is called a univariate forecasting problem. The multivariate forecasting model is an extended version of the univariate forecasting model where future data points not only depend on the preceding values of the same series but also the values of other time series.

The Persistence forecast and the ARIMA statistical model are used as benchmark models that provide a point of comparison with LSTM architectures. As it is not possible to discover the multivariate forecasting technique with benchmark models, the only univariate forecasting problem is explored.

The Persistence forecast provides a computationally inexpensive forecast as complicated calculations do not happen in the learning procedure. Persistence introduces the concept of "memory." The algorithm utilizes the value at the previous time step t to forecast the outcome at the next time step $t + 1$.

ARIMA is the acronym for Auto-Regressive Integrated Moving Average where each component has a crucial characteristic: AR (Autoregression), relying on a dependent relationship between an observation and some number of lagged observations; I (Integrated), the number of differences of actual observations, needed to make the time series stationarity; and MA (Moving Average), the number of lagged forecast errors in the prediction equation.

In this work, three various LSTM architectures are investigated for the univariate and multivariate time series forecasting. LSTM networks are specially designed to learn long term dependencies in sequences. Three kinds of LSTM architectures are investigated: i) Standard LSTM, ii) Stack LSTM, and iii) Sequence to Sequence (S2S) LSTM. Both univariate and multivariate forecasting problems are studied for each architecture.

4.5.4.2 DATA OVERVIEW

The datasets for this use case were selected from three various data sources: UCI Machine Learning Repository, Driven Data, and Open Power System Data. These datasets were chosen as they cover electricity and weather data, had appropriate time resolution and multiple time series to consider for multivariate forecasting problem. In total, the work was done with four different datasets. These datasets have different sampling rates (ex: one-minute, ten-minute, one-hour). For simplicity, the time series with small frequency were downsampled to fifteen-minute. As a result, the work was done with fifteen minute and hourly sampled datasets.

4.5.4.3 DISCUSSION ON INNOVATION

Investigating both univariate and multivariate forecasting problems with LSTM models is an innovative approach as most of the traditional forecasting problems are solved by applying univariate forecasting approach. However, in this work, the impact of the other factors with multivariate forecasting is also analysed. The robustness of the models is measured with five different evaluation metrics to discover the errors from different aspects.

4.5.4.4 USE CASE #1 RESULTS

The experiments have been done on six different time series. Six LSTM models were trained for each time series considering both univariate and multivariate problems. The Persistence and ARIMA benchmark models were applied for the univariate forecasting problem. The full history was used for training of the ARIMA model. For the LSTM models training, three different window sizes were experimented: 24 hours, 36 hours, and 48 hours. The optimal window size changes depending on the time-series and the LSTM models. In the following figure, the average scaled relative RMSE results of each model are shown for each time series.

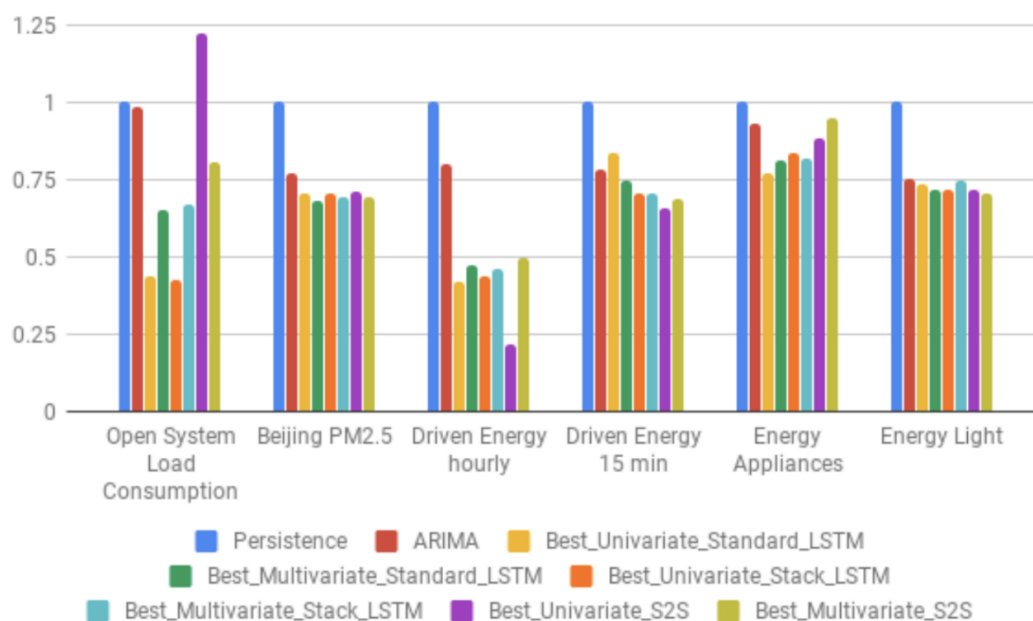


FIGURE 4.21 THE AVERAGE SCALED RMSE ERRORS ARE RELATIVE TO PERSISTENCE³³

The experiments show that the performance of the multivariate and univariate LSTM models is dependent on the LSTM architecture and time series. Before, it was expected that the multivariate LSTM models would beat the univariate LSTM models over all-time series. However, the results showed that the performance of the univariate and multivariate LSTM models depend on the LSTM architecture, the time series, and hyperparameters. The Multivariate Standard LSTM, Univariate Stack LSTM, and Multivariate Stack LSTM models are more stable than Univariate Standard LSTM, Univariate S2S LSTM, and Multivariate S2S LSTM models.

³³ Gunay 2019, used with permission

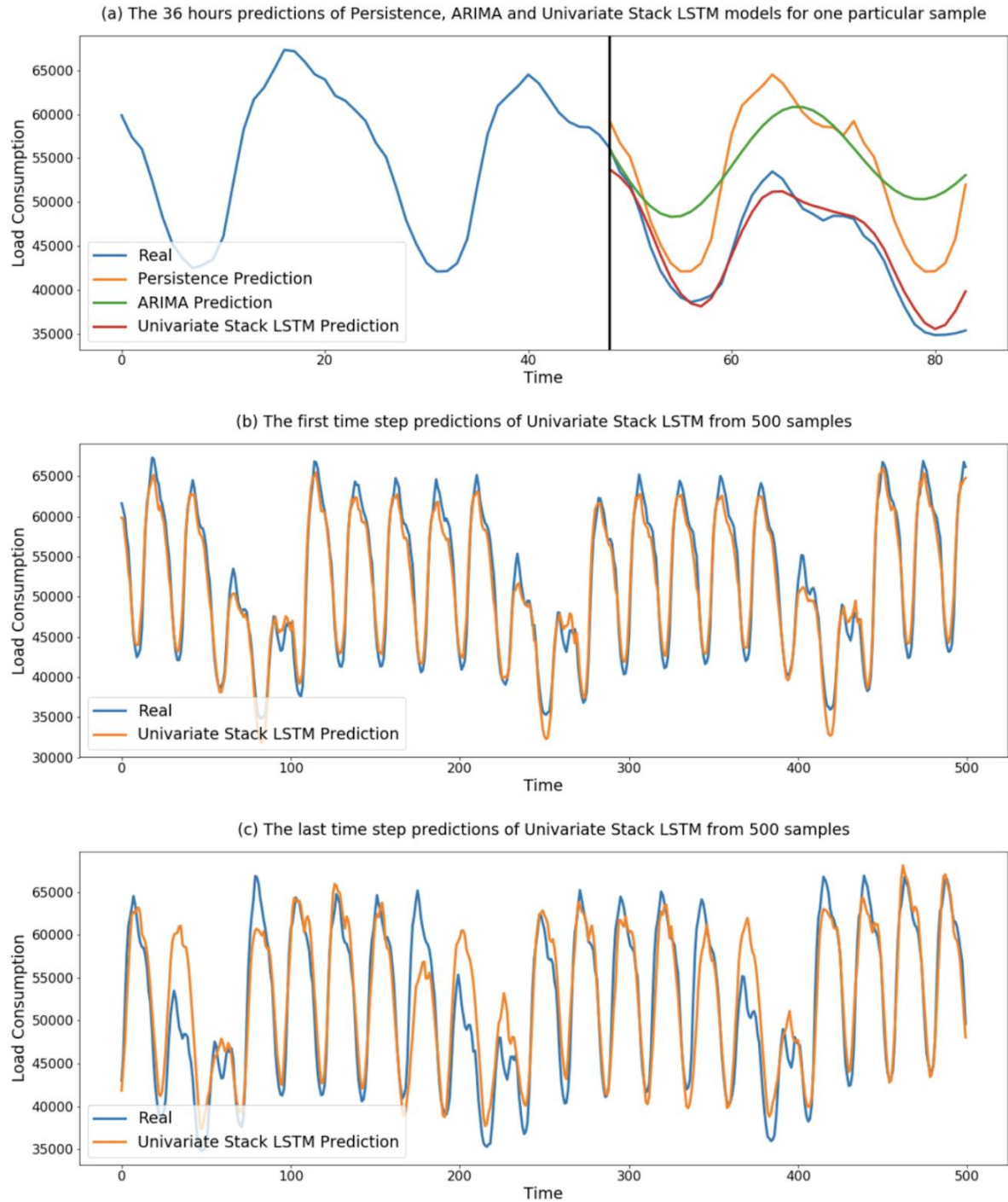


FIGURE 4.22 THE 36 HOURS TIME STEPS PREDICTIONS OF THE PERSISTENCE, ARIMA, AND UNIVARIATE STACK LSTM MODELS ³⁴

To understand the quality of the predictions, the forecasted time points are presented for one particular example time series from the results of the Univariate Stack LSTM model together with Persistence and ARIMA models in the following figure. In Figure 4.22 (a), the 36 hours time steps forecast of Persistence, ARIMA, and Univariate

³⁴ Gunay, 2019, used with permission

Stack LSTM models are depicted for one particular sample. The time series on the left side of the vertical black line is the historical data of the last 48 hours. The predictions and the actual values of the time series are described with different colours which are labelled in the figure. Generally, it can be seen that the predictions can follow the trends in the time series. As expected, the Univariate Stack LSTM model predictions are more accurate than the baseline methods. In Figure 4.22 (b), and (c), the first time step and the last time step predictions of the Univariate Stack LSTM model are shown for 500 samples. Figure 4.22 (b) proves that the model can understand the patterns in the time series, and has accurate results for the first time step. The predictions for the last time step of the samples (Figure 4.22 (c)) are still meaningful, but they are less accurate, especially for the weekends. This issue could be solved by introducing the weekends as additional input features to the model.

TABLE 4.8 AVERAGE SCALED EVALUATION METRICS RESULTS OF MULTIVARIATE STANDARD LSTM, UNIVARIATE STANDARD LSTM AND MULTIVARIATE STACK LSTM FOR EACH TIME SERIES³⁵

Models	Evaluation metrics	Open System Load Consumption	Beijing PM2.5	Driven Energy hourly	Driven Energy 15 min	Energy Appliances	Energy Light
Multivariate Standard LSTM	Avg. RMSE	0.065650	0.073729	0.002840	0.001314	0.075656	0.077640
	Avg. MAE	0.049925	0.051067	0.002134	0.001210	0.046617	0.047086
	Avg. SMAPE	3.496004	30.086966	10.749476	48.071790	24.891671	95.017037
	Avg. BIAS	-0.012298	0.009619	-0.000091	-0.000509	-0.010367	-0.005836
	Avg. CORR	0.861872	0.587559	0.781927	-0.061754	0.451007	0.032446
Univariate Stack LSTM	Avg. RMSE	0.042978	0.076154	0.002654	0.000550	0.0769506	0.077973
	Avg. MAE	0.031728	0.054697	0.001973	0.000471	0.045606	0.047965
	Avg. SMAPE	2.213380	31.137350	9.704834	27.692919	26.089760	94.801964
	Avg. BIAS	-0.002075	8.170910	-0.000467	0.000138	-0.003392	-0.007594
	Avg. CORR	0.941322	0.538969	0.818458	0.220415	0.349457	0.035428
Multivariate Stack LSTM	Avg. RMSE	0.067629	0.074942	0.002781	0.000586	0.076281	0.080890
	Avg. MAE	0.050789	0.052348	0.002059	0.000484	0.045893	0.047052
	Avg. SMAPE	3.541214	30.089051	10.259448	26.106360	25.723219	95.562714
	Avg. BIAS	0.009303	0.005010	-0.000075	-0.000202	-0.005385	0.001445
	Avg. CORR	0.856286	0.557959	0.770238	0.168665	0.371051	0.103801

4.5.5 USE CASE #2: DEVELOPMENT OF CNN-BASED MODELS FOR SHORT-TERM LOAD FORECASTING

4.5.5.1 METHODOLOGY AND APPROACH

In this use case, the primary attention is placed on STLF for evaluation of baseline load during the DR events. Both univariate and multivariate forecasting models are implemented for 24-hour-ahead forecasting. For univariate

³⁵ Gunay 2019, used with permission

models, only the electricity consumption data is used for making the forecast, while for the multivariate models additional to the electricity consumption data is used. The CNN-based models, ARIMA, and Naïve model are implemented to employ MIMO (Multiple Input Multiple Output) strategy for multi-step forecasting when a single model is used to forecast the values for the whole forecast horizon at once.

There are two CNN-based models developed in this use case: CNN and CNN+LSTM model. The CNN model belongs to the class of univariate forecasting models as only historical electricity consumption data is used to make a forecast. The CNN+LSTM model belongs to the class of multivariate forecasting models as weather (average wind speed at 10 meters above the ground in the last 10 minutes, the general wind direction in the last 10 minutes, the highest wind speed at 10 meters above the ground in the last hour, air temperature at 2 meters above the ground at the given date and hour, relative air humidity at the given date and hour) and day information (date of the measurement, ISO week number, day of the week, hour of the measurement, holiday flag) are used in addition to the electricity consumption data to make a forecast.

The CNN model is based on one-dimensional convolutions followed by a fully-connected layer used to process the features extracted from electricity consumption data and to produce the model's outputs. The CNN+LSTM model is based on two-dimensional convolutions used for processing the electricity consumption data and Long-Short Term Memory (LSTM) for processing weather and day information data. A fully-connected layer is used in the CNN+LSTM to merge the extracted from electricity consumption, weather, and day information features and to produce the model's outputs.

In both CNN and CNN+LSTM models, Leaky ReLU and Parametric ReLU are used as activation functions. Dropout layers are added after the convolution and LSTM layers to avoid overfitting. To train the developed CNN-based models, cleansed and prepared for models' training electricity consumption, weather, and day information time-series data are transformed using the sliding window method. The sliding window method is implemented by placing the window of length W on the input sequence and window of length H on the output sequence. At each iteration, both windows are simultaneously shifted forward by some step s . One week of hourly historical electricity consumption data ($W=168$) is used to make the next calendar day hourly electricity consumption forecasts ($H=24$). Blocked k -fold cross-validation and prequential block method are used for the evaluation of the CNN-based models.

In comparison to the developed CNN-based models, ARIMA and Naïve models are implemented as benchmark models. The ARIMA model is modified by adding Fourier terms to it in order to capture multiple seasonalities present in the electricity consumption data. Both ARIMA + Fourier terms and Naïve models belong to the class of univariate forecasting models. Walk-forward cross-validation is used for the evaluation of ARIMA + Fourier terms performing 24-hour-ahead forecast. Naïve model is implemented to perform forecasting for the next calendar day and is evaluated on the whole data set at once.

Additionally, the industry-standard baseline models are used for comparison as benchmark models. The considered industry-standard baseline models are Asymmetric HFoT, SPFoT, Average, and Daily Profile. It is stated

in the sources, that all four models satisfy the main conditions for the baseline models, i.e., accuracy, integrity, simplicity, and alignment. Accuracy is necessary to avoid considerable errors in baseline load evaluation, and integrity is necessary to avoid “gaming the system” by the market players, simplicity is necessary to make it easy for all market players to understand the calculations, alignment is necessary to avoid underestimation or overestimation of the electricity consumption reduction in case of DR events.

In this use case, the length of the DR events is assumed to be not more than an hour as an assumption for performance estimation of the industry-standard baseline models. The industry-standard baseline models are evaluated on the whole datasets at once.

4.5.5.2 DATA OVERVIEW

For the model’s evaluation and comparison, the hourly electricity consumption data from three regions in Norway from 2nd of January, 2016, to 31st of December 2018, was used. The data was provided by a Norwegian EMS operator Enoco and represented the total electricity consumption in each of the regions. The total consumption is calculated as a sum of load of 2 to 3 transformers, which represent one particular region. Using “transformer data” is the most precise approach to representing the total electricity consumption per particular region. However, additional information on the unusual load curves is needed if using the “transformer data” to correctly understand the issues occurring at the specific time points when such curves occur. Enoco also provided this information upon request.

In addition to the electricity consumption data, weather data was retrieved from the web page of Norwegian Meteorological Institute eKlima. The retrieved and used in the models data represents wind, temperature, and relative air humidity characteristics.

4.5.5.3 DISCUSSION ON INNOVATION

The innovation of this use case lies in the exploration and comparison of the CNN-based models to the industry-standard ones. Such comparison is useful to understand the potential of the developed CNN-based deep learning in the task of baseline load evaluation as compared to the widely used approaches. Three evaluation metrics are used to evaluate the models in order to explore the quality of forecasts from different perspectives.

The results of the experiments conducted in this use case are useful to assess the potential of deep learning models being “plugged-in” into the process of baseline load evaluation in the real-world energy systems.

4.5.5.4 USE CASE #2 RESULTS

The main results of the conducted experiments are visualized in Figure 4.23, Figure 4.24, Figure 4.25 and mean values of evaluation metrics are given in Table 4.10,

Table 4.11, Table 4.12. Each figure and table correspond to one of the calculated evaluation metrics, RMSE, MAE, or MAPE. The green, red, and blue bars in figures represent the mean values of the evaluation metrics for each region data, and the error bars represent standard deviation of the evaluation metrics values. The absence of the error bars means that the evaluation metrics for the corresponding models were calculated on the whole datasets at once.

The considered models can be divided into three groups: predictive models that produce a 24-hour-ahead forecast at once, predictive models that produce a 1-hour ahead forecast, analytic models. Predictive models use the electricity consumption data only from before the DR event activation, and analytic models may use the electricity consumption data from both before and after the DR event activation. Analytic models have an advantage over the predictive models as electricity consumption data following the forecasted time point is used. Within the predictive models, models that produce 1-hour-ahead forecasts have an advantage over the models, which produce 24-hour ahead forecasts as a smaller forecast horizon guarantees usage of more recent electricity consumption data for making the forecast. The separation of models into 3 groups is given in Table 4.9.

The results of the experiments have shown high accuracy of forecasts produced by the developed CNN-based models. The developed CNN+LSTM model showed the best results among the predictive models which perform a 24-hour-ahead forecast at once.

Daily Profile showed the best results among the predictive models overall. Taking into account the assumption of maximum one-hour DR events, Daily Profile can be applied to one-hour-ahead forecasting only. Smaller forecast horizon (1 hour) for Daily Profile as compared to a 24-hour forecast horizon for the CNN-based models can explain the better forecasts being produced by Daily Profile. Verifying the assumption of the developed CNN-based models producing better forecasts if forecasting for one-hour-ahead only, the preliminary results have proven this.

Average showed the best results among all the models; however, it belongs to analytic models and can be used only for the evaluation of baseline load on the historical data, not for forecasting.

TABLE 4.9 HIERARCHY OF THE MODELS

Predictive models		Analytic models
24-hour-ahead forecasts	1-hour-ahead forecasts	Average SPFoT
CNN CNN+LSTM ARIMA + Fourier terms Naïve model	Daily Profile Asymmetric HFoT	

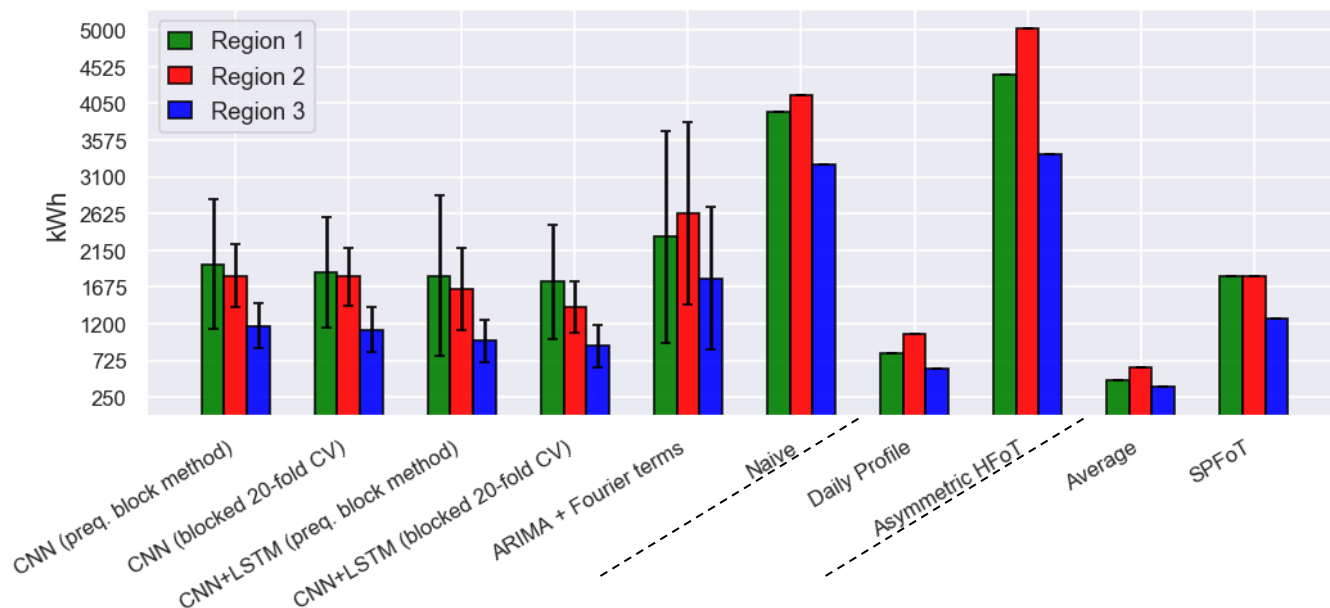


FIGURE 4.23 VALUES OF RMSE (KWH) ON ALL DATA SETS

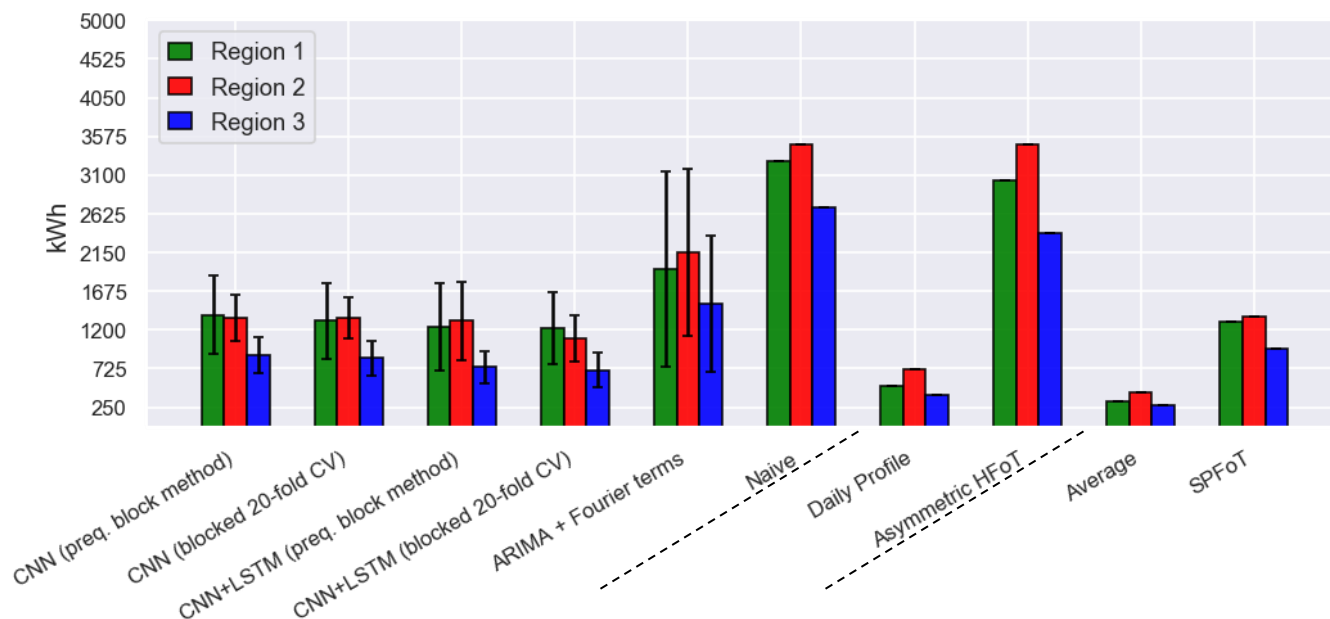


FIGURE 4.24 VALUES OF MAE (KWH) ON ALL DATA SETS

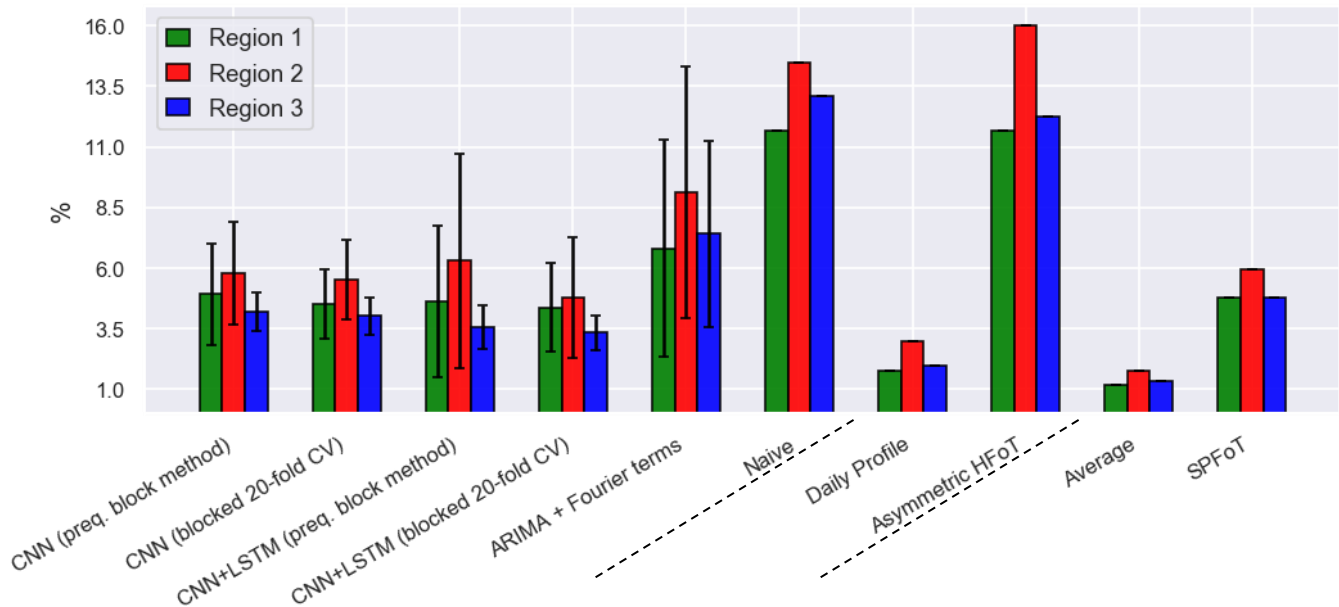


FIGURE 4.25 VALUES OF MAPE (%) OF ALL DATA SETS

TABLE 4.10 MEAN VALUES OF RMSE (KWH) ON ALL DATA SETS

Model	Region 1	Region 2	Region 3
CNN (prequential block method)	1971.47	1820.58	1173.38
CNN (blocked 20-fold cross-validation)	1864.74	1813.34	1125.18
CNN+LSTM (prequential block method)	1819.89	1645.46	981.21
CNN+LSTM (blocked 20-fold cross-validation)	1742.67	1411.83	912.90
ARIMA + Fourier terms	2326.62	2628.48	1786.83
Naïve	3936.76	4152.95	3253.71
Daily Profile	818.92	1067.56	618.8
Asymmetric HFoT	4427.94	5014.45	3387.97
Average	474.45	637.14	394.58
SPFoT	1819.74	1810.66	1274.84

TABLE 4.11 MEAN VALUES OF MAE (KWH) ON ALL DATA SETS

Model	Region 1	Region 2	Region 3
CNN (prequential block method)	1385.02	1344.11	888.67
CNN (blocked 20-fold cross-validation)	1311.84	1344.06	852.76
CNN+LSTM (prequential block method)	1230.75	1309.29	739.37
CNN+LSTM (blocked 20-fold cross-validation)	1217.40	1095.82	702.30
ARIMA + Fourier terms	1945.49	2147.61	1517.23
Naïve	3273.45	3469.20	2701.60
Daily Profile	504.12	720.70	406.01
Asymmetric HFoT	3035.47	3472.48	2382.75
Average	328.83	430.28	274.42
SPFoT	1294.30	1365.44	972.09

TABLE 4.12 MEAN VALUES OF MAPE (%) ON ALL DATA SETS

Model	Region 1	Region 2	Region 3
CNN (prequential block method)	4.93	5.79	4.20
CNN (blocked 20-fold cross-validation)	4.50	5.51	4.02
CNN+LSTM (prequential block method)	4.61	6.31	3.55
CNN+LSTM (blocked 20-fold cross-validation)	4.37	4.79	3.34
ARIMA + Fourier terms	6.80	9.12	7.40
Naïve	11.64	14.46	13.08
Daily Profile	1.79	3.00	1.98
Asymmetric HFoT	11.64	15.98	12.23
Average	1.17	1.76	1.33
SPFoT	4.79	5.97	4.80

4.5.6 GENERAL CONCLUSIONS

The various LSTM- and CNN-based models were developed for multi-step electricity consumption forecasting problems. ARIMA and Naïve baseline models were built for comparison of the forecast results. Additionally, the industry-standard models were used for comparison to CNN and CNN+LSTM. The performance of the models was evaluated on multiple time series through three evaluation metrics, RMSE, MAE, and MAPE. To evaluate the robustness of the LSTM models more precisely, correlation and bias metrics are also calculated for the predictions.

In the case study of the LSTM-based models, the analysis of the errors revealed that the performance of the Univariate Standard LSTM, Univariate S2S LSTM, and Multivariate S2S LSTM models are worse than benchmark methods for some of the time series. In turn, the Multivariate Standard LSTM, Univariate Stack LSTM, and Multivariate Stack LSTM models performed better than the benchmark methods for all-time series. To compare the stability of the models, as presented in Table 4.8, all evaluation metrics were analysed for the Multivariate Standard LSTM, Univariate Stack LSTM, and Multivariate Stack LSTM models. The univariate Stack LSTM model

shows the best results for three-time series overall evaluation metrics and has results near to the best performance for the other time series. In conclusion, the Univariate Stack LSTM model was chosen as a robust model due to the stable results over the all-time series. During the analysis of the predictions, it has been noticed that the days of the weeks are essential to do more accurate forecasts. As future work, the time features as months, days, hours can be added to the feature set to train the Multivariate LSTM models.

In the case study of CNN-based models, it was investigated that the developed CNN+LSTM model showed the best results as given in Table 4.10,

Table 4.11, Table 4.12 among the models from the same level of hierarchy using both 20-fold cross-validation and prequential block method (20 folds, ten folds as an initial training set) for performance estimation. Compared to the developed CNN model, it shows that utilization of more complex architectures along with additional features as weather data or day information yields a better forecasting ability of the CNN-based models. Among the predictive models, Daily Profile showed the best results on all data sets. However, the preliminary results of the CNN-based models being implemented to produce 1-hour-ahead forecasts show the prevalence of both CNN and CNN+LSTM over Daily Profile on all data sets. Therefore, CNN-based models show a strong potential to be included in the real-life energy systems for baseline load evaluation or assessing the availability of flexibility products for the next calendar day.

4.6 PRIVACY-PRESERVING DATA ANALYSIS

Main section authors: Angela Sahk (Cybernetica), Ville Sokk (Cybernetica)

4.6.1 ABSTRACT

The purpose of privacy-preserving data analysis chapter is to introduce better privacy-enhancing technologies (PETs) in the electricity market. The goal is to bring innovation to the sector by introducing privacy technologies and showcasing the benefits of including such technologies. The ultimate goal is to protect customers and lower risks associated with unsuitable data storage and processing.

The approach here is a case study. One limited use case was chosen to be implemented using a suitable PET. Learnings from the piloting are included in this report, but there is also intention an article to be produced to introduce further the innovation PETs can bring to the sector. Additionally, the PoC (proof of concept) implementation is turned into a demonstrator under WP9 to better showcase PETs within the project.

The main findings are:

- Confirmation that PETs can be used to implement various use cases in the electricity/flexibility market;
- Insight that PETs should be included in early (re-)design of systems as changes to approaches and processes may be required to adopt PETs and enable new innovative ways of doing things;
- Laws and regulations will start impacting the electricity sector more and more, as it uses and stores highly sensitive data:
 - Privacy issue was highlighted by The European Consumer Organization (The European Consumer Organisation, 2018) in regards to consumer data availability for aggregators.

As a result of this work we suggest that any new designs / re-designing of existing systems and functionalities would consider the need for PETs as early as possible, where it's necessary to protect consumer data.

4.6.2 INTRODUCTION

4.6.2.1 CONTEXT

The utilities sector in general, of which electricity is a big part, is collecting consumers' consumption and production information in order to provide them with required services. As with all developing areas, the sector is moving towards more flexible approaches and globalisation. It increases the need to think about data privacy and confidentiality. Both to protect customers and in order to lower risks associated with unsuitable data storage and processing.

For example, *Electricity aggregators* are businesses that collect energy metering data from end consumers in order to estimate future energy consumption. They could sell the aggregated demand response capacity on the flexibility market, so that when transmission and distribution system operators need flexibilities to operate the grid reliably and securely they can buy decreased consumption from aggregators.

Privacy problems with sharing consumer data with aggregators have been highlighted in The European Consumer Organization paper "*Electricity Aggregators: Starting off on the Right Foot with Consumers*" (The European Consumer Organisation, 2018):

As they control vital activities of a households' everyday life the remote reading of electricity consumption can provide a detailed insight into households' private sphere. Aggregators requesting consumers' data should provide justification on the necessity of the data and should be able to access it only after the explicit consent of the consumer.

--- The European Consumer Organization

Furthermore, from article 20 "*Functionalities of smart metering systems*" (EU Directive 2019/944, 2019) it is clear that electricity data collection with smart devices must comply with GDPR (and other) regulations:

(c) the privacy of final customers and the protection of their data shall comply with relevant Union data protection and privacy rules

The use of secure computing can, in certain situations, make GDPR compliance easier to attain (Bogdanov et al., 2016).

4.6.2.2 AIM

Privacy-preserving data analysis study aims to improve privacy of consumers and confidentiality of businesses via the use of privacy-enhancing technologies (PETs) such that aggregators still can perform fixed data analytics queries. Consumer data remains protected via cryptographic means even during data processing. Although the task has focused on limited use case it can be expanded to most use cases where consumption or production data is being used and privacy is of concern. This study aims to bring innovation by introducing privacy technologies

and showcasing the benefits of including such technologies. Ultimate goal with introducing privacy technologies is to protect both the private data of consumers and the sensitive data of businesses.

4.6.2.3 OVERVIEW

The privacy-preserving data analysis work carried out relates to this statement of the EU-SysFlex project DoA:

"Implementation and demonstration of some of the above data exchange, data storage and data processing functionalities required for the success of cross-border and cross-sector demonstrations with WP9, adhering to the requirements of volumetry, time, security, privacy."

In particular, this chapter investigates and demonstrates adherence to strict security and privacy requirements in sensitive metering data collection and analytics via the use of PETs.

In general, the approach for privacy-preserving data analysis, is two-fold. A proof of concept (PoC) has been developed that could:

- provide useful maintenance, development such as insight on privacy-preserving development in the electricity sector;
- be turned into a useful privacy-preserving demonstrator as a part of EU-SysFlex data management demonstrators in WP9.

Also, an article will be published in association with EU-SysFlex Task 5.3 to introduce privacy-technologies in the electricity sector. This article will make use of the findings from the PoC whilst relying on the broader aggregators use case described above.

4.6.3 METHODOLOGY AND APPROACH

4.6.3.1 OVERVIEW

Baseline calculation (used to estimate energy consumption from historical meter readings) was selected as an example of analytics to be implemented in a privacy-preserving manner. The ideas presented here can be extended to other more complicated analytics just as well.

System or market operators, assuming they have the responsibility to calculate baselines, may not be willing or able to calculate baselines. To allow baseline calculation to be outsourced to a third party the one must ensure that the computation party does not learn the consumer's metering data and the individual baselines calculated.

To implement baseline calculation in a privacy-preserving manner Sharemind MPC (2018) was used. It is a framework that enables data analytics without leaking individual values. A PoC was developed that calculates a baseline for a customer (or aggregated baseline for a set of customers) without revealing their actual meter

readings or the baseline(s) to any of the computing parties. One can learn more about Sharemind MPC from Sharemind web page (Sharemind MPC, 2018) or the *Sharemind Privacy Ecosystem (2020)* document.

A Sharemind MPC deployment consists of three Sharemind MPC servers which must be hosted by different entities. Distributed control ensures privacy since values in Sharemind MPC are encrypted so that no server host can see the original values even during computation. Distributed control also ensures that only agreed upon computations can be run on the encrypted values. A single host cannot run arbitrary computations in the distributed Sharemind MPC deployment on their own.

The contents of the detailed PoC "*Privacy-preserving baseline calculation proof of concept*" can be found in Annex V - Privacy-preserving data analysis: Proof of concept where the following information is presented:

1. more details about Sharemind MPC (section 2 of Annex);
2. specified high-five-of-ten baseline calculation algorithm chosen to be implemented (section 3);
3. expansion of various assumptions made in PoC design and implementation (section 4);
4. specification of input, output, and message formats (sections 5, 6, 8); and
5. exploration of possible future integration with Flexibility Platform using Estfeed (section 7).

The PoC developed Task 5.3 will be integrated into Flexibility Platform via Estfeed for demonstration as a part of WP9 Task 9.3. Preliminary overview of the components included in baseline calculation can be seen in the Figure 4.26.

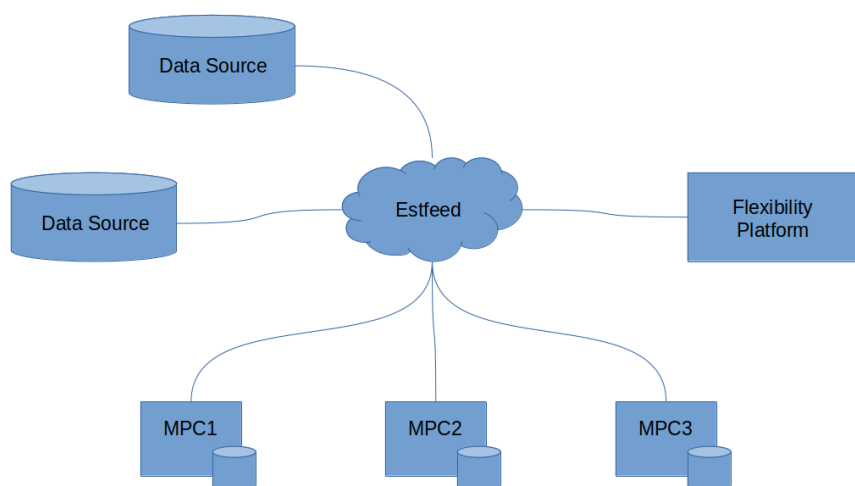


FIGURE 4.26 DIAGRAM OF COMPONENTS PARTICIPATING IN THE BASELINE CALCULATION PROCESS

4.6.3.2 DISCUSSION ON INNOVATION

Privacy-preserving computation is an innovative approach that facilitates better decision making without compromising privacy of individuals involved. Additionally, with increasing regulations, such technologies help to reduce risks associated with data collecting and processing. The PoC has demonstrated privacy-preserving

computation applied to the energy sector for analytics on moderately large data sets. It introduces a better understanding of the usage of PETs at the utilities market.

4.6.4 RESULTS AND CONCLUSIONS

4.6.4.1 RESULTS

To evaluate PoC, Estonian metering data has been received (2017 time period) from transmission system operator Elering aggregated by postcode. Aggregates with a small number of consumers (fewer than 10) and exceptionally low or high consumption were left out. Only historical metering data was used for estimation, and no weather data or other sources were used. The dataset contained hourly metering data from 3836 aggregation points with a total of over 33 million rows.

Developed PoC implementation performed acceptably despite very conservative security choices made. Estimation of 24 hours for a single consumer took 4 seconds. It is expected that by relaxing some security constraints and by applying algorithmic optimizations performance would improve several folds. Currently, all the data columns (user ID, timestamp, measurement) are kept private. For most applications keeping only the measurement value itself private is sufficient. Furthermore, the algorithm is trivially parallelisable (Deng, 2013) and, with appropriate hardware resource, the performance can scale linearly in the number of consumers.

As a result of the work, it is important to highlight that when designing new systems then privacy and security should be considered from the start (Kubo, Sahk, Berendsen & Saluveer, 2019). Considering privacy and security as an afterthought may require much more work and re-design of processes and systems. Moving forward with IoT devices collecting consumption and production information will lead to much more devastating data leaks and (legal) consequences for both the service providers and their customers.

4.6.4.2 CONCLUSIONS

In this research the possibility to use privacy-enhancing technologies in energy sector was explored. During study, a privacy issue was addressed which was highlighted by The European Consumer Organization (The European Consumer Organisation, 2018) in regards to consumer data availability for aggregators. To remedy the situation, it is proposed to use privacy-preserving computation (like secure multi-party computation or fully-homomorphic computation) to protect consumer data while not overly restricting aggregation services in energy market. To demonstrate feasibility, a proof-of-concept baseline calculation algorithm was implemented. Its performance based on moderately sized data sets was evaluated to be reasonable and integration with Flexibility Platform via Estfeed DEP being possible. The results and PoC will be used for the demonstrator to introduce privacy-technologies in the electricity sector.

4.7 DEVELOPMENT OF A BIG DATA SYSTEM FOR THE ELECTRICITY MARKET

Main section authors: Philippe Szczech (AKKA), Riccardo Benedetti (AKKA), Florentin Dam (AKKA)

4.7.1 ABSTRACT

In this case study, a big data system to support the electrical market is developed based on the big data components and architectural design patterns defined in Chapter 1 “Big data framework”. In this intention, two use cases have been selected, developed and deployed over this big data system: the computation of the near real-time electrical consumption prediction based on streaming data and the batch calculation of consumption prediction for a longer time scale. Both are quite representative of the electrical sector which wants to exploit in a near real-time manner massive amount of data issued from smart meters or different sensors and produce various type of predictions such as consumption, flexibility availability. All the technical details of the big data system for the selected and implemented use cases are described.

In parallel, it has been also experimented an improved version of the Seq2Seq prediction algorithm in which is added a residual LSTM network supported by attention mechanisms. This algorithm, initially designed for natural language processing, is not commonly deployed in the electricity domain so far. Conducted experiment revealed that the results obtained are sufficiently accurate for the prediction of electrical consumption in the context of the data used.

4.7.2 INTRODUCTION

4.7.2.1 AIM

The study objective is formally described in the EU-SysFlex Description of Action: *“Implementation and demonstration of some of the above data exchange, data storage and data processing functionalities required for the success of cross-border and cross-sector demonstrations with WP9, adhering to the requirements of volumetry, time, security, privacy.”*

The general objective was refined to a practical approach which consists in the development of a big data system for two business use cases. The first one is related to the prediction of sub metering consumption by exploiting their streaming dataflow and the second one refers to the batch prediction of Estonia or Norwegian area consumptions. The idea behind is to use the big data components and architectural patterns selected in the ‘Big data framework’ dedicated to the elicitation of big data frameworks. The resulting system has been also designed to participate to the demonstrations of the WP9 of EU-SysFlex which is intended to test the ability of the data exchange platforms to connect various data sources, applications or other platforms with each other. The big data system can be considered as one of these applications or platforms.

In this study, it is added a secondary objective consisting in the experimentation of an improved version of the Seq2Seq algorithm to predict electrical consumption.

4.7.2.2 CONTEXT

This chapter reuses the artefacts produced in the Big data framework (see Chapter 1), namely: the big data reference architecture and the list of selected big data components. Also, it is related to the WP9 demos in the way that the developed architecture could serve them by providing an additional business service/application which contributes to the testing of data exchanges between various systems. Finally, Illustration of a big data system for the electricity market is also linked to Cost of data exchange for energy service providers (see chapter 3) which describes a use case to support a business processes of an aggregator through the big data system. The application developed hereafter could represent an example of a real implementation and testing of such aggregator system.

4.7.3 METHODOLOGY AND APPROACH

4.7.3.1 OVERVIEW

The practical approach selected for this study is composed of the following steps:

- The definition of the application use case, in particular the prediction objectives.
- The identification of the data sources to be used to compute the predictions.
- The acquisition of the historical and streaming data.
- The storage of the ingested data into the big data infrastructure to be retrieved by the analytical layer in a later stage.
- The analysis of data to understand structure and quality level. Identification of possible correlations and transformations that can be exploited in a later stage.
- The data preparation, which includes operations like data cleaning, data transformation, feature engineering, normalization and dimensionality reduction.
- The selection of the prediction algorithms compatible with the characteristics of the available data.
- The training of prediction models.
- The validation of models and their optimisation with hyper-parameter tuning techniques.
- The deployment of the trained models in the big data infrastructure.

Those steps have been completed with the preparation of the IT environment:

- Implementation of the big data architecture in a private IaaS solution using the cloud operating system “OpenStack”. The big data components selection is inspired by the reference architecture designed in the Big data framework.
- Interfacing the big data system and the different data sources.
- Interfacing the big data system and applications of WP9 demos to communicate the predictions. A data exchange platform is used to facilitate and secure the data exchange between the different systems.

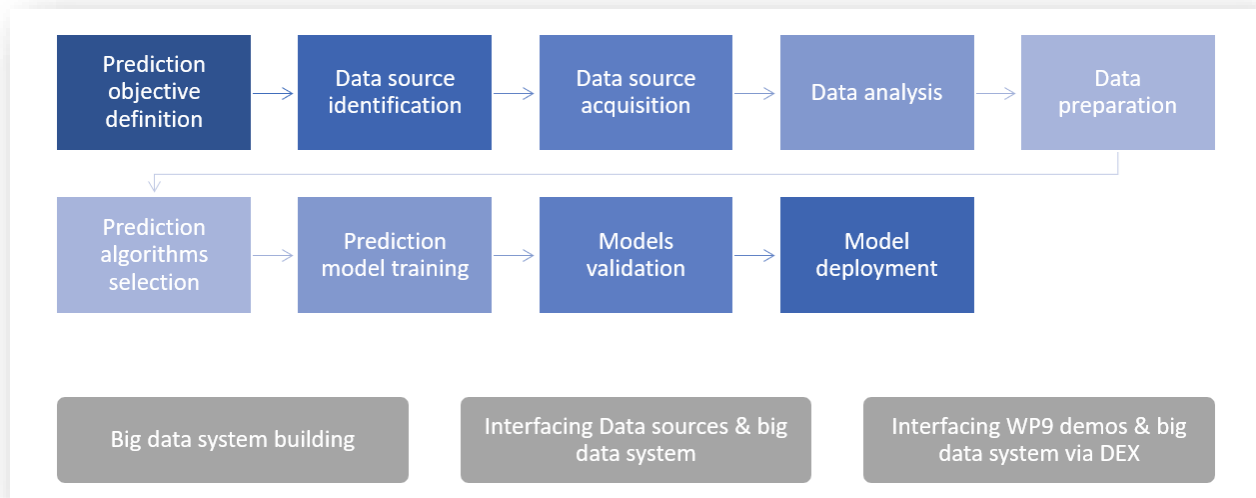


FIGURE 4.27 METHODOLOGY OVERVIEW

4.7.3.2 DISCUSSION ON INNOVATION

The implementation of the two use cases in a big data system built over a private cloud infrastructure and the interfacing with external systems and data sources can be seen as an experimentation of the concept of big data as a Service (BDaaS) applied to the electricity domain with reference to BDaaS by relying on the definition proposed in (International Telecommunication Union Telecommunication Standardization Sector, 2018): *A cloud service category in which the capabilities provided to the cloud service customer are the ability to collect, store, analyse, visualize and manage data using big data.* One of the advantages of this solution is to make accessible big data services to the actors of electricity market which cannot setup this kind of infrastructure on their premises due to its CAPEX impact. For example, it could result being the case of small or medium aggregators.

The innovative aspect of this task is completed with the test and the deployment of an improved Seq2Seq algorithm. The algorithm is quite popular in the natural language processing domain but it not primarily used for prediction of electricity consumption so far.

4.7.4 USE CASE DESCRIPTION, TECHNICAL IMPLEMENTATION AND RESULTS

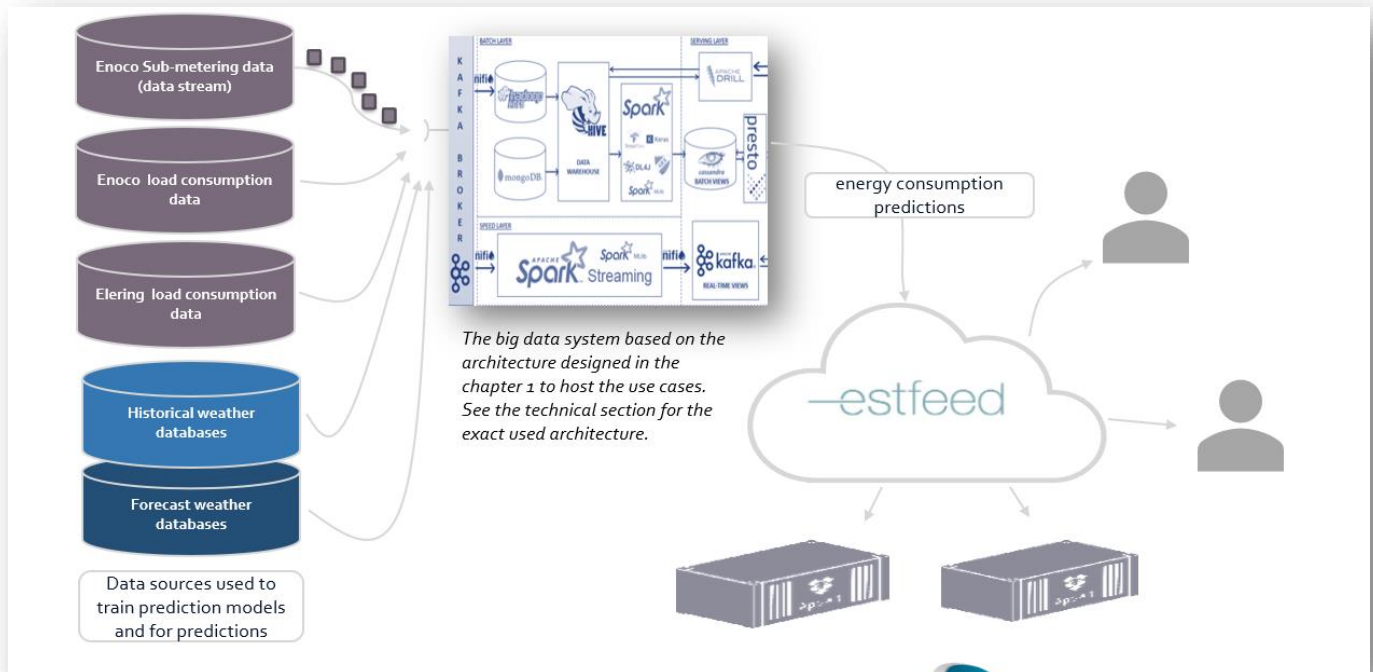


FIGURE 4.28 OVERVIEW OF USE CASE DATA FLOWS

4.7.4.1 USE CASE #1 – CONSUMPTION PREDICTION BASED ON STREAMING DATA

The general objective is the recurrent prediction of the future consumption of a set of household appliances (namely main meter, heat pumps, ventilation systems, boiler and water heater) from a given building. The future predictions are based on the history of the past consumptions. This use case refers to the requirements of the *SUC flexibility bids* and *SUC flexibility prediction* developed in Task 5.2 of EU-SysFlex. Specifically, it attempts to fulfil the requirements by providing a short-term prediction for the definition of the flexibility bids.

The whole scenario has been split in the following series of tasks, each representing a specific step of the data lifecycle:

Data acquisition: the streams of power consumption data are measured and published in near-real-time (with 1 second frequency) from each sub-meter. The stream also includes temperature information about the area outside the building, in order to support the predictions by exploiting possible correlations between weather condition and power usage. Enoco provided the data sources and their primary destination is a data broker into the big data platform.

Data processing: the collected data are retrieved by the processing engine which cleans, parses and transforms the raw data structure. From this stream of power values, it computes the energy consumption for the last 15 minutes, repeating the computation every next 15 minutes.

Data analytics: a machine learning model has been designed and trained to predict the future energy consumption continuously. Precisely, given the current time t_0 , the model must predict the consumption at the time t_0+36h and this task is repeated every 15 minutes.

Data serving: the sequence of the predictions computed in the previous step must be available for the WP9 demos through a service exposed by the big data system. Estfeed is used as the data exchange platform to enable the access to this service.

Data visualization: an external application retrieves the data available in the serving layer in order to show the result of the previous steps in a graphical user-friendly interface, e.g. for monitoring purpose. The information displayed are real-time energy consumption and comparison between predicted values and effective ones.

4.7.4.2 USE CASE #2 – CONSUMPTION PREDICTION BASED ON BATCH COMPUTING

The goal of this second use case is the long-term prediction of the electrical load of some geographical areas, or cities. In opposition with the previous use case which illustrates real-time and continuous treatments, here, the emphasis is on the batch processing of data, meaning an asynchronous task in which the treatment is applied to a massive data volume in one go.

This use case is built based on the datasets obtained from Enoco, Elering and the University of Tartu. Respectively, it has been provided: the load consumption of three regions in Norway, the metering data of 3836 Estonian districts for the year of 2017 and the historical weather data in Norway. These weather data are used only for model training, whereas forecasting weather data from met.no accessible through an API are used to generate the load predictions.

This has been split in a following series of tasks:

- 1) **Batch data analytics:** service that computes the prediction of load consumption on-demand for a given Norwegian city or an Estonian district and for a given timestamp.
- 2) **Serving layer:** hosts the service which delivers the computed prediction to the application or the user. From a web interface, the end-user will be able to select the area and the date of prediction he wants to retrieve. These inputs will be provided to the big data system through the Estfeed data exchange platform which will be also used to transfer the results of the prediction to the user.
- 3) **Data visualization:** similar to the use case #1, it consists of a set of graphs that will display the collected data (historical and forecast for weather data) and the computed predictions.

DATA ANALYSIS AND PREPARATION OVERVIEW

The two datasets provided for the prediction model training include the hourly consumption of some Norwegian regions and some Estonian districts. These data have been analysed and transformed to prepare the following prediction model development phase.

For Estonian data, the dataset received was composed of 33.5 million of values generated by the measurement points of cities. The data analysis revealed that there was no missing data. However, it was also found that the number of values is hugely different depending on the number of measurement points in the different cities. Indeed, the smallest city has 10 measurement points, the biggest one has 2289 while on average there are 145. As the initial data set was too big for the training and testing operations, 3.3% of data from the initial dataset were extracted by keeping the same data distribution among the different cities.

During the data transformation, the *weekday* information which were initially coded as textual values, have been transformed in numeric values with the one-hot encoder technique. Each day is replaced by a binary 7-value vector where the position of a single '1' value indicates the day while all other values are set '0'. This step is necessary because the neural network cannot handle raw textual values. Besides, in order to speed up the learning process and improve the results, the data were normalized to work only with consumption values ranging between '0' and '1'.

As a next step, data is split into two groups: 75% of the data for training model phase and 25% of the data for testing.

For Norwegian data, the dataset includes the hourly load consumption of 3 Norwegian regions since beginning of 2016 to end of 2018 which represent 26304 measurements for each region. The weather conditions (temperature, precipitation, humidity, etc.) of each measurement were provided as well.

The data analysis concludes that there are only 10 missing values, managed by removing the relative lines. Data have been normalized as done for Estonian data. The study of possible correlations between the features is summarized in the matrix below. The *wind speed* was removed ('FF' in the correlation matrix) based on its high correlation with the *highest wind speed* of the last hour ('FG_1' in the correlation matrix).

At first, just one region is used for the model training and the other two for testing. It could have also been possible to use cross-validation or prequential blocked methods to provide different trade-offs between training and test sets in order to increase the model accuracy possibly.

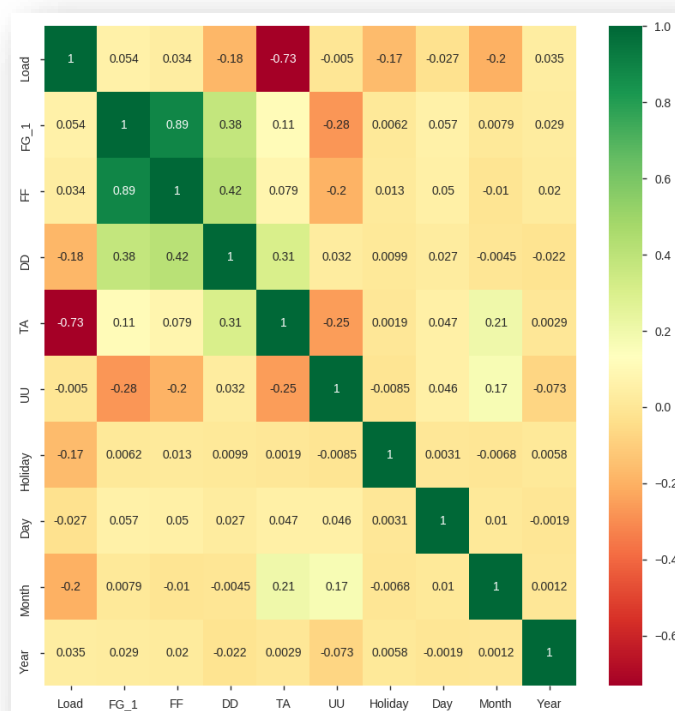


FIGURE 4.29 CORRELATION MATRIX

THE IMPROVED SE2SEQ ALGORITHM

During the prediction algorithm selection phase, the first tests of some algorithms such as RNN and LSTM did not deliver the expected accuracy. For example, the prediction score R2 (coefficient of determination) obtained with the RNN algorithm was 0,54 and the LSTM one was -0,19 (see the figure hereafter).

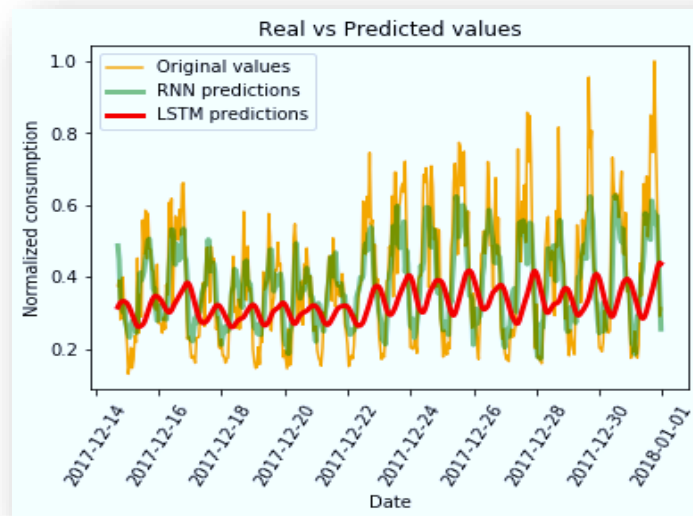


FIGURE 4.30 ENERGY CONSUMPTION IN ESTONIA

For this reason, it has been decided to investigate a new algorithm. The Seq2Seq with Residual LSTM and Attention mechanisms architecture has been finally selected as an innovative technique which achieves excellent results in predicting the energy consumption. This algorithm is fully described in (Gong, An, Mahato, 2019) where the authors have proposed this architecture to predict the energy consumption in New York. In this case, with a mean squared error obtained of 0,000083, this algorithm has outperformed popular deep learning algorithms such as RNN, LSTM and GRU.

The algorithm is composed of several parts:

A Seq2Seq core: a neural network composed of two components: an encoder and a decoder. The encoder is used to compress the data to preserve the most critical information which will serve to the decoder to realize predictions. This Seq2Seq represents the central core that performs the predictions and contains the Residual LSTM and Attention Mechanisms.

A Residual LSTM: The Seq2Seq core component is completed with some mechanisms to learn patterns such as seasonality, consumption habits. For this reason, Recurrent Neural Networks like LSTM are traditionally used inside the encoding and decoding blocks. Nevertheless, in LSTM, when the volume of training data is too high, the phenomenon of gradient explosion tends to occur. It means that during the training phase the error accumulated by the neural network has become quite significant so that the network becomes unstable and it cannot consequently learn from the training data. To address this issue, an architecture called Residual LSTM has been proposed by (Kim, El-Khamy, Lee, 2017).) which consists in adding a connection between two consecutive layers of a LSTM network to deactivate non-linear activation functions which tend to provoke the explosion gradient

(the cumulated error becomes so important that the learning becomes instable) or vanishing gradient (the cumulated error becomes so little that no improvement happens in the learning phase).

Attention Mechanisms: The previous architecture works well with fixed sequence of data but not with long and variable sequences of data. In order to reduce the length of the sequences to be processed, an algorithm called Attention Mechanism is used to focus the learning only on the most essential features. It helps the encoder to know which set of features to prioritize during the learning phase. More concretely, the attention mechanisms will be placed just before the Residual LSTM layer to perform a series of transformations on the data that generate a matrix of weights for the Residual LSTM.

THE IMPROVED SEQ2SEQ MODEL TRAINING, EVALUATION AND OBTAINED RESULTS

In order to train the algorithm, the model has been designed in PyTorch, a deep learning framework for Python programs and developed by Facebook. It has been also proved that it is a good fit for professional real-world application.

It has been used 70% of the Norway dataset for training which consists of about 18400 records. The remaining data is used for testing in order to evaluate the model by measuring the error rate as the 'Mean Squared Error' (MSE) between predictions and real consumptions.

In a first step, the algorithm was trained and evaluated on one region, and it has obtained a MSE of 0,13. It demonstrates that the improved Seq2Seq algorithm is effective for the load consumption prediction in the context of used data.

The graphic hereafter shows the evolution of the error rate for each epoch. The training stops after the 7th epoch as the error began to re-increase.

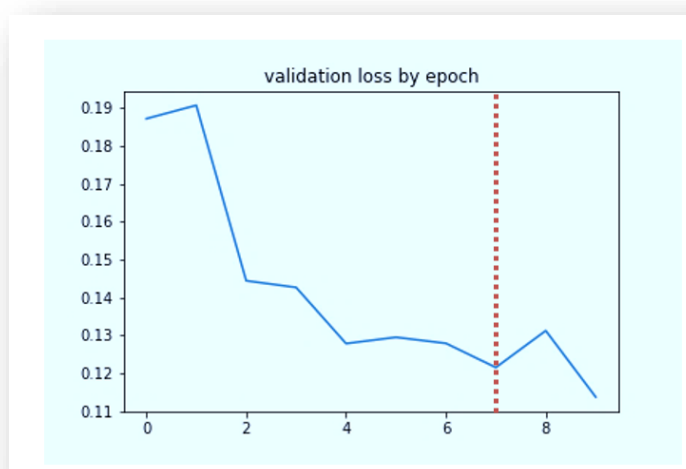


FIGURE 4.31 MSE BY EPOCH ON THE VALIDATION SET

4.7.4.3 TECHNICAL IMPLEMENTATION OF THE USE CASES

The big data system has been deployed in a private cloud IaaS solution, the OpenStack cloud solution. The big data components are distributed over the nodes of the cluster and managed through the Cloudera Hortonworks Data Platform and the Apache Ambari administration tool.

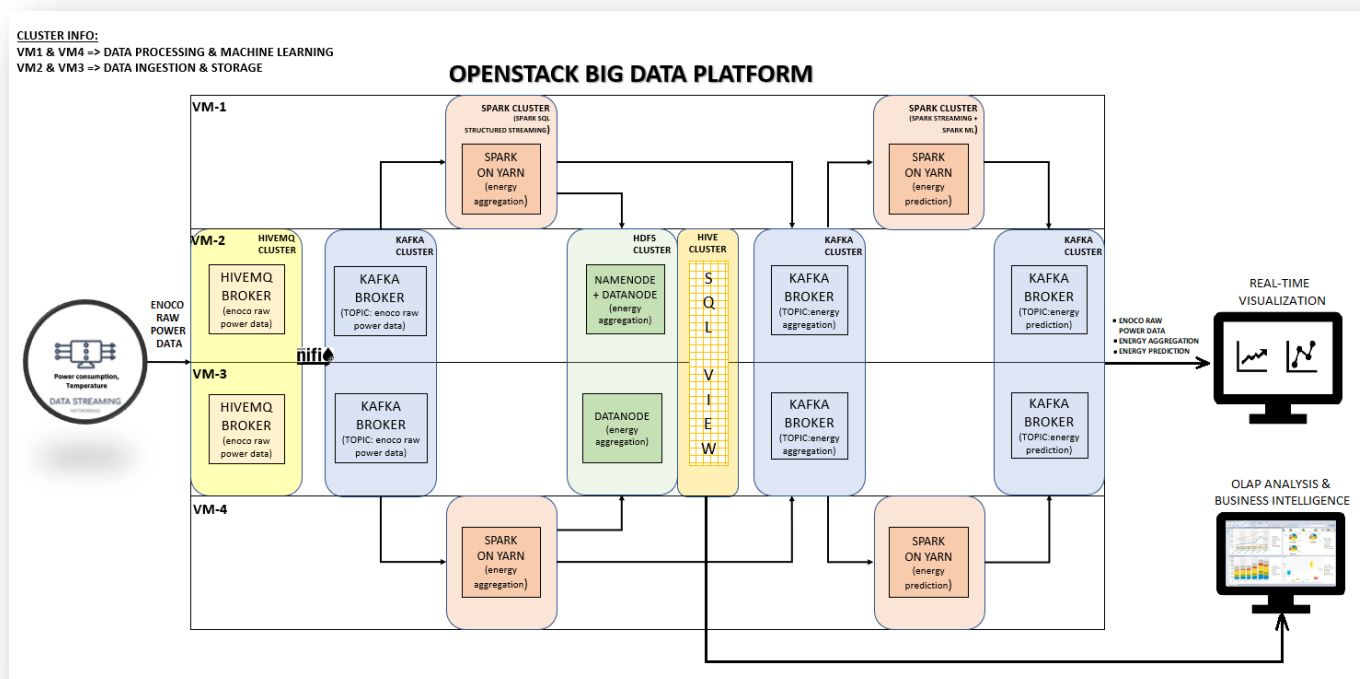


FIGURE 4.32 THE INITIAL BIG DATA CLUSTER

STREAMING USE CASE

Concerning data ingestion, the transmission is made from Norway (Customer Enoco premises) to France (AKKA premises) through the lightweight messaging protocol MQTT, increasingly adopted in the IOT domain. The OpenStack big data system built is composed of a cluster of four VMs, of which two are provided with high RAM (processing machines) and the other two are provided with high disk space (storage machines). Meter, sub-meter and temperature data are handled in a near-real time inflow as follows:

- 1) Data firstly arrive over an encrypted TSL channel to a MQTT broker (HiveMQ) distributed on the storage machines. Each metering device publishes its data into a separated *topic* (a logical space to distinguish the data coming different IOT devices).
- 2) Apache NiFi consumes the data from each topic and merges/publishes them into a global topic (*Enoco raw power data*) hosted into an Apache Kafka broker. The name of the MQTT topic where the data come from is used as device identifier. The Apache Kafka brokers are distributed again over two storage

machines and each topic, including those that will be created in the next steps, contains two copies of the incoming data for fault tolerance purpose.

- 3) A Spark SQL Structured Streaming, running on two YARN *NodeManagers* on the storage machines, consumes the data from the Kafka topic. Those raw data are represented as JSON strings containing power value in kW, device id and timestamp. Data preparation includes dynamic type parse and windowed transformation (i.e. produce chunks of 15 minutes data grouped by sub-meter id). Energy aggregation (kWh) is computed every 15 minutes as the definite integral of the power values in this interval.
- 4) The sequence of energy aggregation is forwarded to another Kafka topic as part of the streaming pipeline. Alongside, the energy aggregation is stored into the Hadoop File System (HDFS) in Parquet format to exploit data compression and to optimize the memory usage. Apache Hive is used to build an SQL view of the data stored in HDFS in order to provide a support for OLAP and batch analytics.
- 5) The prediction task concerns the energy aggregation and is scheduled under the conditions already mentioned in the paragraph “The ‘streaming data’ use case (Data analytics)”. The algorithm for the prediction is a streaming linear regressor where the model is constantly adjusted after each batch of incoming data. Training data are consumed by a Spark Streaming Job from the Kafka topic defined in the previous point. The predictions are computed by Spark ML and the results are made available to a third Kafka topic.
- 6) Each Kafka topic (*Enoco raw power data*, *energy aggregation* and *energy prediction*) is available to an external front-end which displays such information in a dynamic chart.

BATCH USE CASE

The collection of the historical and forecast weather data from Met.no is done through a REST API. The model is deployed by serializing the PyTorch model into a Pickle format (binary file). A python script is used to load the file in memory to perform predictions. The results are sent to the prediction requestor through the data exchange platform Estfeed.

4.7.5 CONCLUSIONS

Developed use cases concretely demonstrate a big data system that can be used to serve the changing energy market by supporting services that require near real-time processing, advanced prediction techniques, high scalability, fault tolerance and other big data capabilities. This big data system can be interfaced with various data sources and different data exchange platforms, such as Estfeed from Elering. As a result, big data benefits can be exploited by third parties or to external applications. More broadly, this demo illustrates the concept of big data as a service in the context of the electricity sector and it demonstrates the effective accuracy of the improved

Seq2Seq algorithm with residual LSTM and attention mechanisms to predict energy consumption of Norwegian regions and Estonian districts.

BIBLIOGRAPHY

- Abdullayeva, G. (2019). *Application and Evaluation of LSTM Architectures for Energy Time-Series Forecasting*. Master's thesis, University of Tartu.
- BEUC, The European Consumer Organisation. (2018). *Electricity aggregators: Starting off on the right foot with Consumers*. Retrieved from https://www.beuc.eu/publications/beuc-x-2018-010_electricity_aggregators_starting_off_on_the_right_foot_with_consumers.pdf. Accessed March 5, 2020.
- Bogdanov, D., Kamm, L., Kubo, B., Rebane, R., Sokk, V., & Talviste, R. (2016). Students and taxes: a privacy-preserving study using secure computation. *Proceedings on Privacy Enhancing Technologies*, 2016(3), 117-135. Retrieved from <https://content.sciendo.com/view/journals/popets/2016/3/article-p117.xml>. Accessed March 5, 2020.
- Deng, Y. (2013). *Applied parallel computing*.
- Directive (EU) 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity and amending Directive 2012/27/EU, OJ L 158/125 (the Electricity Directive). Retrieved from <http://data.europa.eu/eli/dir/2019/944/oj>. Accessed September 28, 2020.
- Directive (EU) 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity and amending Directive 2012/27/EU. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0944>. Accessed March 5, 2020.
- EirGrid, SEMO, and SONI. (2014). *DS3 System Services: Portfolio Capability Analysis*. Retrieved from <http://www.eirgridgroup.com/site-files/library/EirGrid/DS3-System-Services-Portfolio-Capability-Analysis.pdf>. Accessed September 28, 2020.
- EirGrid, SEMO, and SONI. (2017). *I-SEM Training TSO Scheduling*. Retrieved from <https://www.semo.com/documents/training/TSO-Scheduling.pdf>. Accessed September 28, 2020.
- EirGrid, SONI (2017). *DS3 System Services Contracts for Regulated Arrangements RECOMMENDATIONS PAPER*. Retrieved from http://www.eirgridgroup.com/site-files/library/EirGrid/DS3-System-Services-Contracts-Recommendations_final.pdf. Accessed September 28, 2020.
- EirGrid, SONI. (2019). *Balancing Market Principles Statement: A Guide to Scheduling and Dispatch in the Single Electricity Market*. Retrieved from <http://test.soni.ltd.uk/media/documents/EirGrid-and-SONI-Balancing-Market-Principles-Statement-V3.0.pdf>. Accessed September 28, 2020.
- EirGrid, SONI. (2019). *DS3 System Services Protocol – Regulated Arrangements*. Retrieved from <http://www.soni.ltd.uk/media/documents/DS3-System-Services-Protocol-Recommendations-Paper-with-responses.pdf>. Accessed September 28, 2020.
- EirGrid, (2010). *All Island TSO Facilitation of Renewables Studies*. Retrieved from <http://www.eirgridgroup.com/site-files/library/EirGrid/Facilitation-of-Renewables-Report.pdf>. Accessed September 28, 2020.

- EirGrid, (2017). *Ireland's Grid Development Strategy*. Retrieved from [https://issuu.com/designtactics/docs/eirgrid - ireland s grid developmen](https://issuu.com/designtactics/docs/eirgrid_-_ireland_s_grid_developmen). Accessed September 28, 2020.
- EirGrid. (2018). *Operational Constraints Update 21/09/2018*. Retrieved from http://www.eirgridgroup.com/site-files/library/EirGrid/OperationalConstraintsUpdateVersion1_74_Sep_2018.pdf. Accessed September 28, 2020.
- Electricity Market Information (2015). *Existing generating plant dataset*.
- Elering, AST, Litgrid. (2017). *Demand response through aggregation – a harmonized approach in Baltic region*. Retrieved from <https://elering.ee/sites/default/files/public/Elektritur/Demand%20Response%20through%20Aggregatio n%20%20a%20Harmonized%20Approach%20in%20the%20Baltic....pdf>. Accessed September 28, 2020.
- EnerNOC. (2009). *The Demand Response Baseline*. Retrieved from www.enernoc.com
- ENTSO-E. (2018). *All TSOs' proposal for the Key Organisational Requirements, Roles and Responsibilities*. Retrieved from https://eepublicdownloads.blob.core.windows.net/public-cdn-container/clean-documents/Network%20codes%20documents/Implementation/sys/1.a.180227_KORRR_final.pdf. Accessed September 28, 2020.
- European Commission. (2016). *Network Code on Demand Connection*. Retrieved from <http://data.europa.eu/eli/reg/2016/1388/oj>. Accessed September 28, 2020.
- European Commission. (2016). *Network Code on Requirements for Grid Connection of Generators*. Retrieved from <http://data.europa.eu/eli/reg/2016/631/oj>. Accessed September 28, 2020.
- European Commission. (2017). *Guideline on Electricity Balancing*. Retrieved from <http://data.europa.eu/eli/reg/2017/2195/oj>. Accessed September 28, 2020.
- European Commission. (2017). *Guideline on System Operation*. Retrieved from <http://data.europa.eu/eli/reg/2017/1485/oj>. Accessed September 28, 2020.
- G. A. Elliott, J. H. Anderson. (2011). *Real-World Constraints of GPUs in Real-Time Systems. IEEE 17th International Conference on Embedded and Real-Time Computing Systems and Applications, Toyama, pp. 48-54*.
- Gong, G., An, X., Mahato, N. K., Sun, S., Chen, S., & Wen, Y. (2019). Research on short-term load prediction based on Seq2seq model. *Energies*, 12(16), 3199.
- ITU-T, International Telecommunication Union Telecommunication Standardization Sector. (2018). *Cloud computing – Functional architecture of big data as a service*.
- Kema, D.N.V. (2013). *Development of Demand Response Mechanism - Baseline Consumption Methodology – Phase 2 Results, Final Report*
- Kema, D.N.V. (2013). *Development of Demand Response Mechanism Baseline Consumption Methodology - Phase 1 Results*.
- Kim, J., El-Khamy, M., & Lee, J. (2017). *Residual LSTM: Design of a deep recurrent architecture for distant speech recognition*. arXiv preprint arXiv:1701.03360

- Kubo, B., Sahk, A., Berendsen, V. Saluveer, E. (2019). Privacy by design in statistics: Should it become a default/standard? *Statistical Journal of the IAOS*, vol. 35, no. 4, pp. 623–631. DOI 10.3233/SJI-190532. 12.2019.
- Kurylenko, O. (2020). Development of CNN-Based Models for Short-Term Load Forecasting in Energy Systems. *Master's thesis, University of Tartu*.
- Maceina, T. J., & Manduchi, G. (2017). Assessment of general purpose GPU systems in real-time control. *IEEE Transactions on Nuclear Science*, 64(6), 1455-1460.
- National Grid ESO. (2019). *ESO Balancing Services: A guide to contracting, tendering and providing response and reserve service*. Retrieved from <https://www.nationalgrideso.com/sites/eso/files/documents/ESO%20Balancing%20Services%20Guidance%20Document%20V1.pdf>. Accessed September 28, 2020.
- National Grid UK. (2017). *Firm Frequency Response (FFR) Interactive Guidance*. Retrieved from https://www.nationalgrid.com/sites/default/files/documents/Firm%20Frequency%20Response%20%28FFR%29%20Interactive%20Guidance%20v1%200_0.pdf. Accessed September 28, 2020.
- National Grid, 2020. *Frequency Response Services*. Retrieved from <https://www.nationalgrideso.com/balancing-services/frequency-response-services>. Accessed January 14, 2020.
- National Grid. (2019). *Our Forward Plan 2019-2021*.
- Nolan S. et al. (2019). *Product Definition for Innovative System Services - D3.1*. Retrieved from https://eu-sysflex.com/wp-content/uploads/2019/08/D3.1_Final_Submitted.pdf. Accessed September 28, 2020.
- North American Electric Reliability Corporation. (2008). *Glossary of Terms Used in Reliability Standards. Regulation (EU) 2019/943 of the European Parliament and of the Council of 5 June 2019 on the internal market for electricity (OJ L 158, 14.6.2019, pp. 54-124)*. Retrieved from <http://data.europa.eu/eli/reg/2019/943/oj>. Accessed September 28, 2020.
- Sharemind MPC (2018). *Sharemind MPC (Multi-Party Computation) webpage*. Retrieved March 5, 2020, from <https://sharemind.cyber.ee/sharemind-mpc/>.
- Sharemind Privacy Ecosystem (2020). *Sharemind Privacy Ecosystem Technical Overview*. Retrieved March 5, 2020, from <https://repo.cyber.ee/sharemind/www/files/technology/sharemind-technical-overview.pdf>.
- Sonone, V. (2019). *A gentle start in this series- time series and it's analysis*. Retrieved from <https://medium.com/datadriveninvestor/a-gentle-start-in-this-series-time-series-and-its-analysis-bb0a67503d6e>. Accessed September 28, 2020.
- Transpower. (2019). *Instantaneous Reserve Ancillary Service Schedule*.
- Transpower. (2019). *System Security Forecast 2018*.
- Transpower. (2020). *Ancillary Services Tender*. Retrieved January 17, 2020 from <https://www.transpower.co.nz/ancillary-services-tender>.
- Transpower. *Planning for the future*. Retrieved January 17 2020 from <https://www.transpower.co.nz/about-us/our-purpose-values-and-people/planning-future-0>.

- Woolf, M., Ustinova, T., Ortega, E., O'Brien, H., Djapic, P., & Strbac, G. (2014). Distributed generation and demand response services for the smart distribution network. *Report A7 for the "Low Carbon London" LCNF project: Imperial College London*.
- Wollman, D. (2012). *NIST Presents Green Button: Technical*. United States Department of Energy. Retrieved from <https://www.osti.gov/sciencecinema/biblio/1044999>. Accessed September 28, 2020.
- Yang, M., Otterness, N., Amert, T., Bakita, J., Anderson, J. H., & Smith, F. D. (2018). Avoiding pitfalls when using NVIDIA GPUs for real-time tasks in autonomous systems. *In 30th Euromicro Conference on Real-Time Systems (ECRTS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. Retrieved from <https://drops.dagstuhl.de/opus/volltexte/2018/8984/pdf/LIPIcs-ECRTS-2018-20.pdf>. Accessed September 28, 2020.

COPYRIGHT

Copyright © EU-SysFlex, all rights reserved. This document may not be copied, reproduced, or modified in whole or in part for any purpose. In addition, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.

Changes in this document will be notified and approved by the PMB. This document will be approved by the PMB.

The EC / Innovation and Networks Executive Agency is not responsible for any use that may be made of the information it contains.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under EC-GA No 773505.

ANNEX I – BIG DATA FRAMEWORKS: SUPPLEMENTARY INFORMATION ABOUT COMPONENTS

The following table includes examples of how the selected big data components that could be used to implement the Identification of technical requirements. It is pointed out that it does not represent a definitive solution but only possible options, links or recommendations. The table is structured as follows: The first two columns include the A1 requirement IDs and high-level descriptions, the third column contains the list of the big data components involved in the exemplified implementation of the solution. The solution is described in more detail in the last two columns, after an analysis a priori of the scenario, hypothesis and possible use cases.

ANNEX I: TABLE A.1 BIG DATA COMPONENTS EXAMPLES

Requirement ID	Short description	big data components	Hypothesis / Scenarios / Typical Use cases	Architecture Pattern / Solution
AGG-ED-REQ-3	Data source (e.g. meter data hub) ability to aggregate data	<ul style="list-style-type: none"> • Cassandra • Hive • Drill • Spark 	<p>Data to be aggregated are in the big data system (in the data warehouse). Aggregation is the process which transforms these data into aggregated view. The queries to get the aggregated views are already defined or triggered by OLAP sessions (meaning interactive business analysis).</p> <p><u>Typical use cases:</u></p> <ol style="list-style-type: none"> 1) a data owner (customer) wants to get his consumption aggregated by period. 2) an application wants to get aggregated consumption for every town or geographical area. 3) a business analyst wants to explore the data warehouse for statistical or strategical purposes. 	<p>1 & 2) The aggregation process is managed by Spark which takes the data from the warehouse and stores the aggregated views in Cassandra. Periodical batch processes regularly update the aggregated views.</p> <p>The serving layer receives requests from data user and application. These requests consist of pre-defined queries submitted through the Customer Portal / Applications. The query engine Presto handles these requests and get the data from the batch views.</p> <p>3) The business analyst opens an OLAP session through Drill and interacts directly with the data warehouse performing online aggregations (without using pre-defined queries).</p>
AGG-ED-REQ-4	DEP ability to forward aggregated data from data source to data user	<ul style="list-style-type: none"> • Cassandra • Presto 		
ANO-ED-REQ-3	Data source (e.g. meter data hub) ability to anonymize data	<ul style="list-style-type: none"> • Cassandra • Hive • Spark • ARX 	<p>Data to be anonymized are in the big data system (in the data warehouse). The processes for anonymization are already defined in the BDS (big data System) according with the GDPR guideline.</p> <p>After receiving data from an external source, the BDS must be able to anonymize it and store the anonymized view into the serving layer.</p> <p><u>Typical use case:</u> an application wants to get anonymized data for academic studies, benchmarking, reporting.</p>	<p>The anonymization process is managed by ARX and Spark which take the data from the warehouse and store the anonymized views in Cassandra. Periodical batch processes regularly update the anonymized views.</p> <p>The serving layer receives requests from application. The query engine Presto handles these requests and get the dataset from the batch views.</p>
ANO-ED-REQ-4	DEP ability to forward anonymized data from data source to data user	<ul style="list-style-type: none"> • Cassandra • Presto 		
AUTH-REQ-3	Ability to share information related to representation rights between data users and concerned Customer Portals	<ul style="list-style-type: none"> • MongoDB • Ranger • Knox 	<p>big data system's objective is not to be a server of authentication and authorization but must include these policies. These policies are stored in the BDS.</p> <p><u>Typical use cases:</u></p> <ol style="list-style-type: none"> 1) a user/application wants to access to his data stored in the BDS. 2) a data source/application wants to connect to the BDS to send his data. 	<p>1 & 2) Knox is used as a authentication proxy to verify the identity of the external entity. Once the entity is authenticated, Ranger validates the authorization to access/send the requested data and deny the access to unauthorized entities.</p> <p>N.B: representation rights, permissions & authentication information are (semi)static data, so they are stored in MongoDB.</p>
AUTH-REQ-4	Ability to share authentication information between data users, Customer Portals and Authentication Service Provider	<ul style="list-style-type: none"> • MongoDB • Knox 		
AUTHZN -REQ3	Ability to share access permissions between data owners, concerned DEPs, applications and data sources	<ul style="list-style-type: none"> • MongoDB • Ranger 		

Requirement ID	Short description	big data components	Hypothesis / Scenarios / Typical Use cases	Architecture Pattern / Solution
DC-REQ1.1	Get near-real-time data (up to 1 hour) from meters	<ul style="list-style-type: none"> • Kafka • NiFi 	Data to be collected come from the meters in a continuous streaming with high throughput. <u>Typical use case</u> : collect fast IOT data.	Data are received through NiFi and temporarily buffered in Kafka. Kafka broker allow to: - reduce data loss: in case of failure of the producer or the consumer, data are kept in Kafka waiting for system recovery; - adapt the speed of the producer to the speed of the consumer (and vice-versa); - define a secure data exchange protocol and handle more streams in parallel.
DC-REQ1.2	Get historical data (monthly) from conventional meters	<ul style="list-style-type: none"> • NiFi 	Conventional meters support some storage capabilities since can maintain historical data. This kind of collection is considered likewise a data transfer from the external data hub to the BDS. <u>Typical use case</u> : collect historical data.	Historical data are received through NiFi.
DC-REQ1.3	Store data in meter data hub	<ul style="list-style-type: none"> • Kafka • NiFi • HDFS 	Data to be stored have been just collected inside a broker (DC-REQ1.1) or come directly from an external source (DC-REQ1.2). <u>Typical use case</u> : store data into the BDS data lake.	NiFi routes the data from a source (Kafka or external sources) to HDFS.
DC-REQ2.1	Get near-real-time (up to 1 hour) data from market	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as DC-REQ1.1 with the difference that here data come from market.	See pattern solution for DC-REQ1.1
DC-REQ2.2	Get historical data from market	<ul style="list-style-type: none"> • NiFi 	Same as DC-REQ1.2 with the difference that here data come from market.	See pattern solution for DC-REQ1.2
DC-REQ2.3	Store data in market data hub	<ul style="list-style-type: none"> • Kafka • NiFi • HDFS 	Same as DC-REQ1.3 with the difference that here data come from market.	See pattern solution for DC-REQ1.3
DC-REQ3.1	Get very-near-real-time (up to 1 minute) data from grid	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as DC-REQ1.1 with the difference that here data come from grid.	See pattern solution for DC-REQ1.1

Requirement ID	Short description	big data components	Hypothesis / Scenarios / Typical Use cases	Architecture Pattern / Solution
DC-REQ3.2	Get near-real-time (up to 1 hour) data from grid	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as DC-REQ1.1 with the difference that here data come from grid. There is no difference between near-real-time and very-near-real-time.	See pattern solution for DC-REQ1.1
DC-REQ3.3	Get historical data from grid	<ul style="list-style-type: none"> • NiFi 	Same as DC-REQ1.2 with the difference that here data come from grid.	See pattern solution for DC-REQ1.2
DC-REQ3.4	Store data in grid data hub	<ul style="list-style-type: none"> • Kafka • NiFi • HDFS 	Same as DC-REQ1.3 with the difference that here data come from grid.	See pattern solution for DC-REQ1.3
DT-REQ1	Transfer of data	<ul style="list-style-type: none"> • Kafka • NiFi 	<p>Data to be transferred are in external data hubs or outcoming from the data sources. This requirement concerns only data transfer: storage is out of the scope.</p> <p><u>Typical use cases:</u></p> <ol style="list-style-type: none"> 1) From data sources to BDS 2) From data hubs to BDS 	<ol style="list-style-type: none"> 1) Data transfer is managed using both Kafka and NiFi. Kafka acts as a broker in order to mediate the velocity of the data stream, decouple the senders (IoT meters or event sources) from the receiver (BDS) and reduce the risk of data loss. 2) Data transfer is managed using only NiFi. Kafka is not necessary since the data to be transferred already stored in the data hub. As a consequence, an eventual data loss will be solved simply by resending the data.
DT-REQ3	Data owner's access to data through DEP (and foreign DEP)	<ul style="list-style-type: none"> • Cassandra • Presto 	<p>Data to be transferred are in the big data system.</p> <p><u>Typical use case:</u> a data owner (customer) wants to get his data.</p>	This requirement is considered as a generalization of the AGG-ED-REQ, where the data to be transferred are not necessarily aggregated. See pattern solution 1 & 2 of AGG-ED-REQ and exclude the Spark aggregation batch job.
DT-REQ4	Application's access to data through DEP (and foreign DEP)	<ul style="list-style-type: none"> • Kafka • NiFi • Cassandra • Drill • Presto 	<p>Data to be transferred are in external data hubs, outcoming from the data sources or already stored in the BDS.</p> <p>This requirement concerns only data transfer: storage is out of the scope.</p> <p><u>Typical use cases:</u></p> <ol style="list-style-type: none"> 1) From BDS to application. 2) From external data source to application through a near-real time streaming (no BDS storage). 	<ol style="list-style-type: none"> 1) This requirement is considered as a generalization of the AGG-ED-REQ, where the data to be transferred are not necessarily aggregated. See pattern solution 1 & 2 & 3 of AGG-ED-REQ and exclude the Spark aggregation batch job. 2) Data flow through the speed layer of the BDS (Kafka + NiFi). Since the requirement concerns only data transfer, was not included any processing framework (e.g. Spark Streaming).
DER-SCADA-REQ4	Ability of DEP to forward real-time data from DER's to System Operators	<ul style="list-style-type: none"> • Kafka • NiFi 	<p>Data to be forwarded are in DER.</p> <p>The BDS is not supposed to store these data.</p> <p><u>Typical use case:</u> from DER to SO.</p>	See pattern solution 2 of DT-REQ4.

Requirement ID	Short description	big data components	Hypothesis / Scenarios / Typical Use cases	Architecture Pattern / Solution
DER-SCADA-REQ5	Ability of DEP to forward very-near-real-time (up to 1 minute) data from DER's to System Operators	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as DER-SCADA-REQ4. There is no difference between real-time and near-real-time in this work	See pattern solution 2 of DT-REQ4.
DER-SCADA-REQ6	Ability of DEP to forward near-real-time (up to 1 hour) data from DER's to System Operators	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as DER-SCADA-REQ4. There is no difference between real-time and near-real-time in this work	See pattern solution 2 of DT-REQ4.
DER-SCADA-REQ7	Ability of DEP to forward activation requests from System Operators to DER	<ul style="list-style-type: none"> • Kafka • NiFi 	Activation requests to be forwarded are sent by SO. The BDS is not supposed to store these activation requests. <u>Typical use case</u> : from SO to DER.	See pattern solution 2 of DT-REQ4.
FA-REQ2	Exchange of activation requests through DEP and flexibility platform	<ul style="list-style-type: none"> • Kafka • NiFi 	Activation requests to be forwarded are sent by SO. The BDS is not supposed to store these activation requests. <u>Typical use case</u> : activation requests exchange from SO to FP. Hypothesis 1) FP is an external application. Hypothesis 2) FP is part of the BDS.	H1) See pattern solution 2 of DT-REQ4 (please note, the BDS is not supposed to store these data). H2) See pattern solution 1 of DT-REQ1.
FB-REQ1	Ability of flexibility platform to collect input for baseline calculation, incl. through DEP	<ul style="list-style-type: none"> • Kafka • NiFi 	Input for baseline calculation can be already stored into the BDS or come from external data sources. <u>Typical use case</u> : baseline calculation. Hypothesis 1) FP is an external application. Hypothesis 2) FP is part of the BDS.	This specific requirement concerns only the input data collection (not calculation). According to the SUC, the inputs for baseline calculation could be both sub-meters data and 'certified' meters data. As a consequence, the pattern to use is the same as DC-REQ1.1 for both hypotheses.
FB-REQ2	Ability of flexibility platform to compute baseline	<ul style="list-style-type: none"> • Spark Streaming • Kafka (H1 only) • NiFi (H1 only) 		According to the SUC, "Real-time data are used for the calculation". As a consequence, computation should be performed using the real-time data processing framework (Spark Streaming). The calculation will be based on a standard mathematical formula and not machine learning. In order to reduce the FB computational workload and reduce the data exchange, the proposal would be to compute the baselines always into the BDS, in particular: H1) The BDS compute the baseline through Spark Streaming with the input collected in the FB-REQ1 and then provides the real-time result to the FP through NiFi & Kafka (as a real-time view). H2) Same as before but without the need of NiFi & Kafka to return the result since the FB is part of the BDS.

Requirement ID	Short description	big data components	Hypothesis / Scenarios / Typical Use cases	Architecture Pattern / Solution
FBIDS-REQ2	Ability to exchange information on System Operators' flexibility need and FSPs' flexibility potential through flexibility platform (and DEP)	<ul style="list-style-type: none"> • Kafka • NiFi 	<p>Information to be exchanged are sent by SO to FP (flexibility needs) and by FSP to FP (flexibility bids). <u>Typical use case:</u> flexibility information exchange between SO and FPS through the FP. Hypothesis 1) FP is an external application. Hypothesis 2) FP is part of the BDS.</p>	See pattern solution FA-REQ2 for both hypotheses.
FBIDS-REQ4	Algorithm for prequalification of flexibility providers	<ul style="list-style-type: none"> • Spark Streaming (H1b and H2 only) • Kafka (H1b only) • NiFi (H1b only) 		<p>H1a) This process is implemented and run into the FP: the BDS is not involved. H1b) Similar to the H1 pattern solution of FB-REQ2: delegate the algorithm computation to the BDS (Spark Streaming) and then provides the real-time result to the FP through NiFi & Kafka (as a real-time view). H2) Same as before but without the need of NiFi & Kafka to return the result since the FB is part of the BDS.</p>
FBIDS-REQ6	Flexibility platform's ability to collect bids from FSPs	<ul style="list-style-type: none"> • Kafka • NiFi 		<p>H1) See pattern solution 2 of DT-REQ4 (please note, the BDS is not supposed to store these data). H2) See pattern solution 1 of DT-REQ1.</p>
FBIDS-REQ7	Selection of successful bids	<ul style="list-style-type: none"> • Spark Streaming (H1b and H2 only) • Kafka (H1b only) • NiFi (H1b only) 		See pattern solution FBIDS-REQ4.
FBIDS-REQ9	Calculation of grid impacts (congestion, imbalance)	<ul style="list-style-type: none"> • Spark Streaming (H1b and H2 only) • Kafka (H1b only) • NiFi (H1b only) 		See pattern solution FBIDS-REQ4.
FPRED-REQ1	Collection of data for prediction (long term - years)	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as DC-REQ1.1 & DC-REQ1.2.	See pattern solution DC-REQ1.1 & DC-REQ1.2.

Requirement ID	Short description	big data components	Hypothesis / Scenarios / Typical Use cases	Architecture Pattern / Solution
FPRED-REQ2	Computation of predictions (long term - years)	<ul style="list-style-type: none"> • Spark • Deep learning frameworks (Keras, TensorFlow, DL4J) • Cassandra (H1 only) 	<p>Data to use for the prediction are in the big data system (in the data warehouse) due to the FPRED-REQ1.</p> <p>A long-term prediction can be handled by a batch processing framework, not necessarily a real-time one.</p> <p><u>Typical use case:</u> long term flexibility predictions.</p> <p>Hypothesis 1) FP is an external application.</p> <p>Hypothesis 2) FP is part of the BDS.</p>	<p>The long-term predictions are periodically computed in a batch process implemented through a deep learning framework and running in a Spark cluster.</p> <p>These predictions will be stored in:</p> <p>H1) a batch view into Cassandra and eventually get by the FP through a query.</p> <p>H2) the FP.</p>
FPRED-REQ1	Collection of data for prediction (medium-term - days to years ahead)	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as DC-REQ1.1 & DC-REQ1.2.	See pattern solution DC-REQ1.1 & DC-REQ1.2.
FPRED-REQ2	Computation of predictions (medium-term - days to years ahead)	<ul style="list-style-type: none"> • Spark • Deep learning frameworks (Keras, TensorFlow, DL4J) 	<p>Data to use for the prediction are in the big data system (in the data warehouse) due to the FPRED-REQ1.</p> <p>A medium-term prediction can be handled by a batch processing framework (as well as the long term), not necessarily a real-time one.</p> <p><u>Typical use case:</u> medium-term flexibility predictions.</p> <p>Hypothesis 1) FP is an external application.</p> <p>Hypothesis 2) FP is part of the BDS.</p>	<p>The medium-term predictions are periodically computed in a batch process implemented through a deep learning framework and running in a Spark cluster.</p> <p>These predictions will be stored in:</p> <p>H1) a batch view into Cassandra and eventually get by the FP through a query.</p> <p>H2) the FP.</p>
FPRED-REQ1	Collection of data for prediction (short term - intraday operation)	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as DC-REQ1.1 & DC-REQ1.2.	See pattern solution DC-REQ1.1 & DC-REQ1.2.
FPRED-REQ2	Computation of predictions (short term - intraday operation)	<ul style="list-style-type: none"> • Spark Streaming • Deep learning frameworks (Keras, TensorFlow, DL4J) 	<p>Data to use for the prediction are in the big data system (in the data warehouse) due to the FPRED-REQ1.</p> <p>A near-real-time processing framework should handle a short-term prediction.</p> <p><u>Typical use case:</u> short term flexibility predictions.</p> <p>Hypothesis 1) FP is an external application.</p> <p>Hypothesis 2) FP is part of the BDS.</p>	<p>The short-term predictions are continuously computed in a near-real-time process implemented through a deep learning framework and running in a Spark cluster (using the Spark Streaming functionalities).</p> <p>These predictions will be stored in:</p> <p>H1) a real-time view into Kafka and eventually consumed by the FP through NiFi.</p> <p>H2) the FP.</p>
FVERIF-REQ1	Calculation of actually delivered flexibility as a response to an activation request	<ul style="list-style-type: none"> • Spark (H2 only) 	<p>Data to use for the verification are in the Flexibility Platform.</p> <p>The flexibility verification process is computed once a month: it is assumed the batch case (no need for real-time processing).</p> <p>Typical use case: calculation and verification of the delivered flexibility.</p>	<p>H1) The BDS does not own information about the delivered flexibilities.</p> <p>The calculation and the verification processes happen on the FP side and the BDS is not involved.</p> <p>H2) The BDS includes the FP, then it is provided of the information about the delivered flexibilities. The calculation and the verification processes can be managed inside the BDS</p>

Requirement ID	Short description	big data components	Hypothesis / Scenarios / Typical Use cases	Architecture Pattern / Solution
FVERIF-REQ2	Verification that flexibility delivered matches with flexibility requested	<ul style="list-style-type: none"> • Spark (H2 only) 	<p>Hypothesis 1) FP is an external application.</p> <p>Hypothesis 2) FP is part of the BDS.</p>	through Spark.
LOGS-REQ1	Ability to share information related to security logs between data owners, concerned DEPs, applications and data sources	<ul style="list-style-type: none"> • Ranger • Knox 	<p>All the information related to accesses, authentications and authorizations need to be recorded into security logs.</p> <p><u>Typical use case:</u> detect possible suspicious activities and prevent data breaches.</p>	The components already proposed in the SUC authentication and SUC Authorization (Knox and Ranger) are also provided of auditing functionalities to produce automatically the security logs.
SUBMET-REQ1	Collection of data from sub-meters	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as DC-REQ1.1.	See pattern solution DC-REQ1.1.
SUBMET-REQ3	Storing sub-meter data in data hub	<ul style="list-style-type: none"> • Kafka • NiFi • HDFS 	Same as DC-REQ1.3, with the terminological distinction that the BDS stores these data into its "data lake" instead of the "data hub".	See pattern solution DC-REQ1.3.
SUBMET-REQ2	Ability of DEP to forward sub-meter data from data hub to customer (data owner) and an application (energy service provider)	<ul style="list-style-type: none"> • Kafka • NiFi • Cassandra • Drill • Presto 	Same as DT-REQ3 (for data owner) & DT-REQ4 (for energy service provider).	See pattern solution DT-REQ3 & DT-REQ4.
SUBMET-REQ7	Ability of DEP to forward activation orders from a customer (data owner) or application (energy service provider) to devices	<ul style="list-style-type: none"> • Kafka • NiFi 	Same as FA-REQ2.	See pattern solution FA-REQ2.

ANNEX II – IDENTIFICATION OF TECHNICAL REQUIREMENTS

ANNEX II: TABLE A.2 TECHNICAL REQUIREMENTS FOR SELECTED SUCs

SUCs	REQUIREMENTS	DESCRIPTION OF RELATED DATA						TYPE OF REQUIREMENT					
		Volume of data to be <u>collected</u> by time period (e.g.: 125 MB/sec, 1 GB/min)	Volume of data to be <u>processed</u> by time period (e.g.: 125 MB/sec, 1 GB/min)	Type of processing of data (e.g.: prediction, reformatting, anonymization)	Type of data (e.g.: Structured, semi-structured, unstructured data, Times series, Streaming, Sequence, Graph, Spatial)	Other information	Accuracy (is it necessary to complete, filter, transform, to correct the data?)	Performance	Big data	Personal data	Security	Functional	
SUC: Aggregate energy data		Aggregation is internal processing of existing data – no data is collected	Daily aggregated reports from data hub to SOs, suppliers, aggregators	Aggregation	Semi-structured sequence data – meter data		Depends on accuracy of the underlying data – no further actions needed to make data more accurate						
AGG-ED-REQ-1	○ Standard rules to aggregate data in order not to enable the identification of persons behind data										V	V	
AGG-ED-REQ-2	○ Standard rules to aggregate data in order to ensure the comparability of aggregated data sets												V
AGG-ED-REQ-3	○ Data source (e.g. meter data hub) ability to aggregate data							V	V			V	
AGG-ED-REQ-4	○ DEP ability to forward aggregated data from data source to data user							V	V			V	

SUC: Anonymize energy data		Anonymization is internal processing of existing data – no data is collected	Based on the request of data user. Try to test at least one case where we will anonymize historical data of e.g. 100 000 consumers – hourly data of 5 years.	Anonymization	Semi-structured sequence data – personal meter data, personal market data		Depends on accuracy of the underlying data – no further actions needed to make data more accurate						
ANO-ED-REQ-1	<ul style="list-style-type: none">Standard rules to anonymize data not to enable the identification of persons behind data										V	V	
ANO-ED-REQ-2	<ul style="list-style-type: none">Standard rules to anonymize data in order to ensure the comparability of anonymized data sets												V
ANO-ED-REQ-3	<ul style="list-style-type: none">Data source (e.g. meter data hub) ability to anonymize data							V	V			V	
ANO-ED-REQ-4	<ul style="list-style-type: none">DEP ability to forward anonymized data from data source to data user							V	V			V	
SUC: Authenticate data users		Each time when it is necessary to authenticate the user – millions of users. Thousands of authentication cases per day	Thousands of authentication cases per day, cases are not simultaneous necessarily		Structured sequence data – information related to authentication and representation rights		Accuracy needs to be checked with the role Authentication Service Provider						
AUTH-REQ-1	<ul style="list-style-type: none">Right to access own data									V			
AUTH-REQ-2	<ul style="list-style-type: none">Authentication tools											V	
AUTH-REQ-3	<ul style="list-style-type: none">Ability to share information related to representation rights between data users and concerned Customer Portals							V	V			V	
AUTH-REQ-4	<ul style="list-style-type: none">Ability to share authentication							V	V			V	

	information between data users, Customer Portal and Authentication Service Provider											
SUC: Manage access permissions		Each time when it is necessary to authorize the user – millions of users. Thousands of access permissions per day	Thousands of access permission per day		Structured sequence data – access permissions		Accuracy needs to be checked with the role Consent Service Provider (=DEP)					
AUTHZN-REQ1	○ Every person needs access permission									V	V	
AUTHZN-REQ2	○ Valid identity of the person receiving access permissions											V
AUTHZN-REQ3	○ Ability to share access permissions between data owners, concerned DEPs, applications and data sources							V	V			V
SUC: Collect energy data												
DC-REQ1	○ Collection of meter data		N/A	Collection and storing. Assume personal meter data is stored for at least 5 years. EC: https://ec.europa.eu/info/law/law-topic/data-protection/refo	Structured time-series data – consumption and generation data from certified meters	From near-real-time (15min/1h) to historical (monthly) readings. Data should be read and available in data hub immediately	Data from certified meters has to be max accurate because is used for imbalance settlement and billing. Responsibility of Meter data operator					
DC-REQ1.1	▪ Get near-real-time data (up to 1 hour) from meters	20 million hourly values per day. 1 message containing 24 hourly values for 1 metering point = 3kB						V	V	V	V	
DC-REQ1.2	▪ Get historical	Few values per						V	V	V	V	

	data (monthly) from conventional meters	month		rm/rules-business-and-organisations/principles-gdpr/how-long-can-data-be-kept-and-it-necessary-update-it_en								
DC-REQ1.3	<ul style="list-style-type: none"> Store data in meter data hub 	Hourly readings for 700 thousand metering points for at least 5 years							V			V
DC-REQ2	<ul style="list-style-type: none"> Collection of market data 		N/A	Collection and storing	Structured time-series, streaming or sequence data – public market data (e.g. weather, price information), individual market data (e.g. bids, schedules)	From near-real-time to historical. Data should be available in data hub immediately	Responsibility of Data provider					
DC-REQ2.1	<ul style="list-style-type: none"> Get near-real-time (up to 1 hour) data from market 	Thousands of values per day						V	V	V	V	
DC-REQ2.2	<ul style="list-style-type: none"> Get historical data from market 	Thousands of values per month						V	V	V	V	
DC-REQ2.3	<ul style="list-style-type: none"> Store data in market data hub 								V			V
DC-REQ3	<ul style="list-style-type: none"> Collection of grid data 	Grid outages data is sent by TSO and collected by ENTSO-E Transparency Platform right	N/A	Collection and storing of data on planned grid outages	Structured time-series, streaming or sequence data – different types of grid data	From very-near-real-time to historical. Data should be available in data hub immediately	Responsibility of Data provider					
DC-REQ3.1	<ul style="list-style-type: none"> Get very-near-real-time (up to 1 minute) data from grid 							V	V		V	
DC-REQ3.2	<ul style="list-style-type: none"> Get near-real-time 							V	V		V	

	(up to 1 hour) data from grid	after it occurs – size of a message is few kBs										
DC-REQ3.3	<ul style="list-style-type: none"> Get historical data from grid 							V	V		V	
DC-REQ3.4	<ul style="list-style-type: none"> Store data in grid data hub 								V			V
SUC: Transfer energy data		N/A	Thousands of values per second. Millions of values per minute. Assume 10 applications. A message with grid (outage) data depending on the number of values may be 5-50 kB. A message with meter data containing 24 hourly values for 1 metering point is 3 kB	Forwarding existing data from data hubs. EC: Data portability aims to empower data subjects regarding their own personal data, as it facilitates their ability to move, copy or transmit personal data easily from one IT environment to another (whether to their own systems, the	Structured time-series, streaming or sequence data – 15min/1h/monthly consumption and generation data from certified meters; public market data (e.g. weather information, price information); individual market data – bids, schedules; different grid data (tbd); “IoT” data collected from sub-meters	From real-time to historical data. Data needs not to be exchanged immediately after collecting it but with some time delay – from 1 second to several years later	Data transferred has to be of same accuracy as provided by data providers. Responsibility of Data exchange platform operator					
DT-REQ1	<ul style="list-style-type: none"> Transfer of data must be secured, by means of encryption or communication protocol 							V	V	V	V	
DT-REQ2	<ul style="list-style-type: none"> Data portability (applies to personal data - Article 20 of the GDPR) 											V
DT-REQ3	<ul style="list-style-type: none"> Data owner’s access to data through DEP (and foreign DEP) 								V	V	V	V
DT-REQ4	<ul style="list-style-type: none"> Application’s access to data through DEP (and foreign DEP) 								V	V	V	V

				systems of trusted third parties or those of new data controllers).								
SUC: Exchange data between DER and SCADA				Collection and forwarding data								
DER-SCADA-REQ4	○ Ability of DEP to forward real-time data from DER's to System Operators	Hundreds of real-time values	Hundreds of values exchanged in less than 1 second		Structured time-series data – sub-meter data for verification and settlement of flexibility activations	From real-time to near-real-time. Data should be available immediately after submitting	Flexibility bids and meter data from devices has to be accurate enough to satisfy the parties involved in flexibility and other energy services	V	V	V	V	
DER-SCADA-REQ5	○ Ability of DEP to forward very-near-real-time (up to 1 minute) data from DER's to System Operators	Hundreds of values per minute	Hundreds of values exchanged in less than 1 minute		Structured sequence data – flexibility bids. Structured time-series data – sub-meter data for verification and settlement of flexibility activations			V	V	V	V	
DER-SCADA-REQ6	○ Ability of DEP to forward near-real-time (up to 1 hour) data from DER's to System Operators	Hundreds of values per hour	Hundreds of values exchanged in less than 1 hour		Structured sequence data – flexibility bids. Structured time-series data – sub-			V	V	V	V	

					meter data for verification and settlement of flexibility activations								
DER-SCADA-REQ7	<ul style="list-style-type: none">Ability of DEP to forward activation requests from System Operators to DER	Hundreds of values per minute	Hundreds of values exchanged in less than 1 minute		Structured sequence data – flexibility activation signals	Very-near-real-time (real-time activations are excluded)	Flexibility activation signals have to be received by Flexibility providers either directly or through Aggregator	V	V	V	V		
DER-SCADA-REQ2	<ul style="list-style-type: none">Communication link between DEP and SO’s SCADA											V	
DER-SCADA-REQ1	<ul style="list-style-type: none">Encrypted data exchange									V		V	
DER-SCADA-REQ3	<ul style="list-style-type: none">Safety of DER’s IT infrastructure											V	
SUC: Manage flexibility activations		Hundreds of values per minute	Hundreds of values exchanged in less than 1 minute (Very fast products have to be activated as the response to the frequency deviations in the	Collection and forwarding data	Structured sequence data – flexibility activation signals	Very-near-real-time (real-time activations are excluded)	Flexibility activation signals have to be received by Flexibility providers either directly or through Aggregator						
FA-REQ2	<ul style="list-style-type: none">Exchange of activation requests through DEP and flexibility platform							V	V		V	V	
FA-REQ1	<ul style="list-style-type: none">Automated activation of devices is possible											V	

			grid. But otherwise, for slower products the activation request can be sent via DEP.)									
SUC: Calculate flexibility baseline		Thousands of values per day	Thousands of values per day	Collection and forwarding data	Meter and sub-meter data; usages of rooms and devices, climate conditions, etc.	Each time when baseline needs to be calculated (product specific)	Accuracy depends on the needs of computation methodology					
FB-REQ1	○ Ability of flexibility platform to collect input for baseline calculation, incl. through DEP							V	V	V	V	V
FB-REQ2	○ Ability of flexibility platform to compute baseline								V			V
SUC: Manage flexibility bids				Collection and forwarding data		Product specific	Estimations					
FBIDS-REQ2	○ Ability to exchange information on System Operators' flexibility need and FSPs' flexibility potential through flexibility platform (and DEP)							V	V	V	V	V
FBIDS-REQ4	○ Algorithm for prequalification of flexibility providers								V			V

FBIDS-REQ6	○ Flexibility platform's ability to collect bids from FSPs	Few values per minute/hour. Size of the bid – 120 kB	Few values exchanged in less than 1 hour. Size of the bid – 120 kB		Flexibility bids	Product specific	Responsibility of flexibility providers		V	V	V	V
FBIDS-REQ7	○ Selection of successful bids								V			V
FBIDS-REQ8	○ Flexibility platform's ability to collect grid validation results from SOs	Few per minute	Few per minute (Algorithm for grid impact assessment will not be developed as part of WP9.)		Grid impact assessment	Continuous	Accuracy of grid impact assessment is responsibility of SO	V		V		V
FBIDS-REQ9	○ Calculation of grid impacts (congestion, imbalance)								V		V	V
FBIDS-REQ3	○ Auction process supervised by Market Operator											V
FBIDS-REQ5	○ Automated exchange of bids is possible											V
SUC: Predict flexibility availability												
FPRED-REQ1	○ Collection of data for prediction (long term - years)	Low – MB/week	Low – MB/week	Collection	Unstructured assessments of future electrical requirements, generation and constraints		Low	V	V			V
FPRED-REQ2	○ Computation of predictions (long term - years)	Low – MB/week	Low – MB/week	Manual analysis	Spreadsheet etc. analysis		Low	V	V			V
FPRED-REQ3	○ Collection of data for prediction (medium term - days to years)	Low – MB/day	Low – MB/day	Collection	Structured – from auction results, external generation		Medium/High	V	V			V

	ahead)				and usage predictions etc.							
FPRED-REQ4	<ul style="list-style-type: none"> Computation of predictions (medium term - days to years ahead) 	Low – MB/day	Low – MB/day	Combining data to gain single view	Reports providing days to years ahead		Medium/High	V	V			V
FPRED-REQ5	<ul style="list-style-type: none"> Collection of data for prediction (short term - intraday operation) 	High – MB to GB/sec	High – MB to GB/sec	Real-time collection of data	Structured readings of real time electricity usage/generation		High	V	V			V
FPRED-REQ6	<ul style="list-style-type: none"> Computation of predictions (long term - intraday operation) 	High – MB to GB/sec	High – MB to GB/sec	Real-time processing of data	Real-time understanding of current system and where flexibility exists and is needed		High	V	V			v
SUC: Verify and settle activated flexibilities		N/A	Hundreds of	Requesting								
FVERIF-REQ1	<ul style="list-style-type: none"> Calculation of actually delivered flexibility as response to activation request 		values per day	data necessary for the SUC	1sec to 1 hour meter and sub-meter data and activation requests as inputs	Each time when verification and settlement processes are required	Max accuracy of meter and sub-meter data		V			V
FVERIF-REQ2	<ul style="list-style-type: none"> Verification that flexibility delivered matches with flexibility requested 				Baseline and actually delivered flexibility as inputs				V			V
FVERIF-REQ3	<ul style="list-style-type: none"> Calculation of the penalty if flexibility 											V

	delivered is less than flexibility requested											
SUC: Provide list of suppliers and ESCOs					Updated information about the suppliers and service providers	According to the requests	All suppliers and service providers should be in the list					
ESCO-REQ1	<ul style="list-style-type: none"> List of suppliers and ESCOs is available through DEP; List of aggregators is available through flexibility platform additionally 	Few per day (adding new units to the list)	N/A									V
SUC: Erase and rectify personal data					Any personal data	According to the need	N/A					
PERSO-DATA-REQ1	<ul style="list-style-type: none"> Ability to share information related to erasure of personal data between data owners, concerned DEPs, applications and data sources 	Few per day	Few per day							V		V
PERSO-DATA-REQ2	<ul style="list-style-type: none"> Ability to share information related to rectification of personal data between data owners, concerned DEPs, applications and data sources 	Few per day	Few per day							V		V
SUC: Manage data logs					Information about the access to data:	Logs should be available when	Responsibility of DEP operator					
LOGS-REQ1	<ul style="list-style-type: none"> Ability to share information 	Thousands of	Thousands of						V	V	V	V

	related to data logs between data owners, concerned DEPs, applications and data sources	messages per second	messages per second		which party, when, which data	requested						
SUC: Manage sub-meter data				Collection, storing and forwarding data								
SUBMET-REQ1	○ Collection of data from sub-meters	Hundreds of values per second	~1kb/sec per meter		1sec to 15min consumption and generation data of devices	From real-time to historical	Meter data from devices has to be accurate enough to satisfy the parties involved in flexibility and other energy services	V	V	V	V	
SUBMET-REQ3	○ Storing sub-meter data in data hub	Hundreds of values per second	~2 MB/day per meter						V	V	V	V
SUBMET-REQ2	○ Ability of DEP to forward sub-meter data from data hub to customer (data owner) and to application (energy service provider)	N/A	Hundreds per second					V	V	V	V	V
SUBMET-REQ7	○ Ability of DEP to forward activation orders from customer (data owner) or application (energy service provider) to devices	N/A	Hundreds per second (simultaneous activation of few hundred aggregated devices – e.g. heat pumps)		Flexibility activation signals	Near-real-time (real-time activations are excluded)	Flexibility activation signals have to be received by Flexibility providers either directly or through Aggregator	V	V	V	V	V
SUBMET-REQ4	○ Data format of sub-metering				Structured data/			V	V			

					times series data							
SUBMET-REQ5	<ul style="list-style-type: none"> Transmission protocols of sub-metering 	Hundreds of values per second	~1kb/sec per meter		Structured data / times series data			V	V	V	V	
SUBMET-REQ6	<ul style="list-style-type: none"> SLA between customer and energy service provider 					Needed when costumer is responsible for the data transmission infrastructure (e.g. internet connection)	low uptime might impact service quality	V				V

ANNEX III – COMPARATIVE STUDY OF EXISTING SOLUTIONS: DETAILED ESTIMATION OF SELECTED SOLUTION

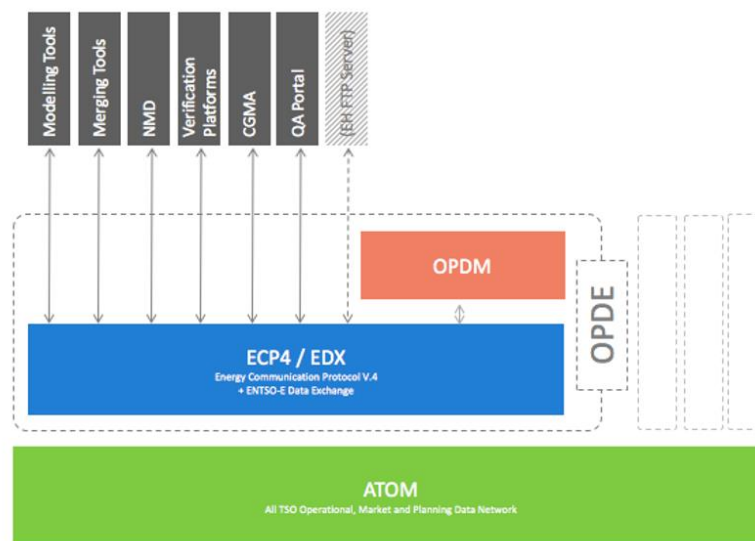
There are 4 existing solutions selected to be analysed regarding defined requirements:

- ENTSO-E OPDE
- Estonian Data Exchange Platform (Estfeed, Data Hub, e-Elering)
- Elhub
- Green button

More detailed information on the analysed solutions can be found in the tables below.

ANNEX III: TABLE A.3 ENTSO-E OPDE DESCRIPTION

Solution name	Operational Planning Data Environment (OPDE)
Owner	ENTSO-E
Operator	ENTSO-E
Purpose	OPDE is the data exchange system for the Common Grid Model (CGM).
Description	The CGM compiles the individual grid model (IGM) of each TSO, covering timeframes ranging from one year before real time to one hour before real time. TSOs' individual (in most cases, national) grid models are collected by Regional Security Centers (RSCs), who, following a quality assessment and pan-European alignment process, merge them into a pan-European Common Grid Model and feed the merged Common Grid Model back into the system. The OPDE, specified by Art. 114 of the SOGL, is the information platform that supports the data exchange associated with the CGM merging process. It is also the foundation of the data exchange platform for fulfilling the five core tasks of RSCs.
Implementation phase	<p>The implementation of the OPDE components by TSOs and RSCs is ongoing. At the end of 2018, data exchange via ENTSO-E's OPDE environment was automated for nine TSOs and two RSCs.</p> <p>Implemented process: Improved Individual Grid Models / Common Grid Model Delivery. Collecting and quality checking Individual Grid Models and merging them into pan-European or regional Common Grid Models to use the as a basis for the other services.</p> <p>Types of data exchanged via OPDE Platform:</p> <ul style="list-style-type: none"> • Common Grid Models (CGM), • Individual Grid Models (IGM), • Boundary Data Sets (BDS).
Modules	



ANNEX III: FIGURE A.1 ENTSO-E MODULES

ATOM (other names: COMMO, PCN – Physical Communication Network) – pan-European private (separate from the internet) MPLS network based on leased lines and TSO-owned communication lines for non-real-time data exchange dedicated for TSO's data exchange.

Implementation of the physical communication network by all TSOs was ongoing during 2018 and at the end of 2018 four TSOs and two RSCs were connected via the physical communication network.

ECP – The Energy Communication Platform v.4 designed and developed by Unicorn as a reference implementation of the MADES standard; based on AMQP protocol; no central communication point, one endpoint can communicate with different brokers.

- High performance – Sophisticated technology to ensure high throughput even over limited bandwidths,
- Guaranteed and traceable communication – With acknowledgment of delivery; Any message transported by ECP can be tracked down to gather trustworthy information about the state of delivery
- Secure and reliable – Encryption and digital signatures; Only recipient of the message is capable of reading the message content. The sender of any message can be unambiguously verified.
- Supported and maintained – Guarantee of high service availability,
- Straightforward integration – Large number of supported technological Interfaces,
- Platform and operating system independent – MS Windows, Linux, Unix, Solaris.

EDX (ENTSO-E Data Exchange) - service based distributed integration platform implementing publish/subscribe mechanism. Designed as universal and reusable for other projects. Developed by Unicorn.

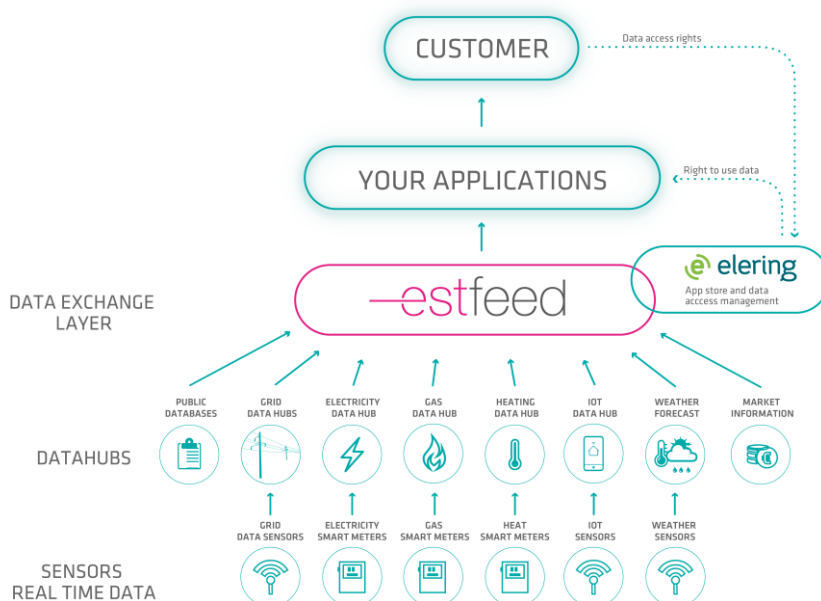
OPDM Client (Operational Planning Data Management Client) – designed for storing

	<p>grid models and related data. Could be extended for other types of data. Provides file storage, data validation, API and GUI for manual user interaction. Developed by Aprico.</p>
Technology	<p>OPDE Platform is designed for non-real-time data exchange communication. OPDE modules are described in “modules” section.</p> <p>[ECP/EDX tier]</p> <p>The ECP network consists of multiple ECP endpoints, component directories, and brokers. ECP uses two main protocols for messaging – AMQPS and HTTPS. Detailed information can be found in “ECP Installation Guide” published on ENTSO-E website: https://www.entsoe.eu/Documents/digital/ecp4/ECP_Installation_Guide_v4.4.0.pdf</p> <p>[OPDE Client]</p> <p>OPDE Client is built as client-server application using EDX Toolbox and ECP communication layer. Technologies used for OPDE Client are: Apache Karaf, Apache Hadoop, Apache Kafka, Elasticsearch and Kibana.</p>
Development plans	<p>Processes to implement:</p> <ul style="list-style-type: none"> • Coordinated Security Analysis Identifying operational security violation risks and planning remedial actions on a regional basis. • Coordinated Capacity Calculation Calculation of cross-border capacities including the optimization of available capacity within operational security limits. • Short and Medium Term Adequacy Performing short and medium term studies on the adequacy of the transmission network; • Outage Planning Coordination Identifying cross border outage incompatibilities between relevant assets.
References	<ol style="list-style-type: none"> 1. OPDE Requirements & Solution Proposal https://extra.entsoe.eu/SOC/IT/WP_3/160128_OPDE_-_Requirements_and_Proposed_Solution.docx 2. ENTSO-E Annual Report 2018 https://annualreport2018.entsoe.eu/wp-content/uploads/2019/06/1_entso-e_ar2018_08_190612.pdf 3. OPDE Client, EDX Documentation 4. ECP Documentation https://www.entsoe.eu/data/transparency-platform/data-providers/

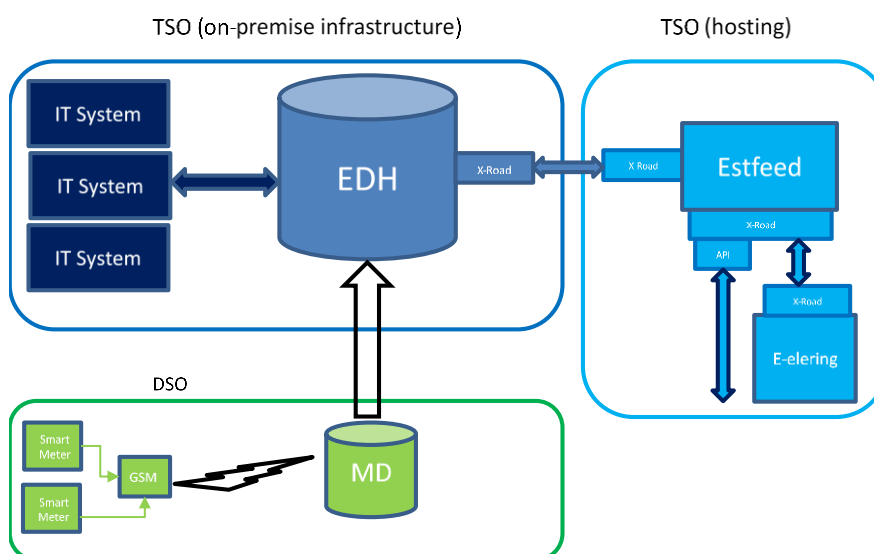
ANNEX III: TABLE A.4 ESTFEED DESCRIPTION

Solution name	Data Exchange Platform Estfeed + Electricity DataHub + e-Elering Portal
Owner	Elering AS
Operator	Elering AS
Implementation phase	Production phase since 2017.
Purpose	Estfeed platform is a digital environment for secure data exchange in the energy market between any stakeholders. Functionalities include consent management for exchanging personal and commercially sensitive data. Data Hub is for collecting and storing meter data but also for the management of supplier switching, joint invoicing, balance chain management. e-elering is customer for portal for accessing data and logs about data usage, granting access permissions and representation rights
Description	<p>According to the Electricity Market Act, all meter data and related master data exchange processes in the Estonian open electricity market take place via the data exchange platform, which must ensure the principles of efficient and equal treatment of market participants in data exchange processes. Data Hub was established for collecting, storing and making available meter data. It was launched in 2013. Data Hub provides equal access to electronic power meter data to all authorized market participants and enables a quick process of switching the supplier.</p> <p>Electricity and gas data hubs are primary data sources integrated with Estfeed platform. Estfeed is data transport layer to which any data source can be connected. On the other side, any application can request data via Estfeed from these sources. Estfeed's consent management solution enables exchange of personal and commercially sensitive data. Estfeed is GDPR compliant.</p> <p>e-elering was designed as customer interface of Estfeed.</p>

Modules



ANNEX III: FIGURE A.2 ESTFEED MODULES



ANNEX III: FIGURE A.3 ESTFEED MODULES

Electricity Data Hub (EDH) – The Estonian Data Hub system is a software/hardware solution that manages the exchange of electricity metering data between market participants, supports the process of changing electricity suppliers in the market, and archives the metering data of electricity consumption. The Estonian Data Hub assigns codes to market participants who operate on Estonian electricity market. It also codes all metering points which are needed in order to track the transfer of energy between market participants. Code assignment provides basis for defining the rights of the market participants and helps to track the supply chains.

X-Road (“X-Road” is currently used to refer to the technology), the data exchange layer for information systems, is a technological and organizational environment

enabling a secure Internet-based data exchange between information systems. X-Road has a versatile security solution: authentication, multi-level authorisation, high-level system for processing logs, and data traffic that is encrypted and signed.

- Independence of platform and architecture – X-Road enables the information systems of X-Road members on any software platform to communicate with the information systems of data service providers on any software platform.
- Multilateralism – X-Road members can request access to any data services provided through X-Road.
- Availability and standardisation – for managing and developing X-Road, international standards and protocols are used where possible.
- Security – exchanging data through X-Road does not affect the integrity, availability or confidentiality of the data.

X-Road is used for data exchange between components of Data Exchange Platform.

Estfeed is Elering's smart grid data sharing platform, which allows the energy sector to exchange messages securely. It is a platform managed by Elering as Estonia's national transmission network operator, via which a consumer's electricity, gas and heat energy metering data are made available to that consumer, as well as to a third party, if so authorised by the consumer, by the law. Different data sources and applications that want to use this data can interface with the platform. Energy service providers, application owners, and end consumers can use the smart grid platform to exchange messages and manage data via the e-elering customer portal. Estfeed uses X-Road as the transport layer.

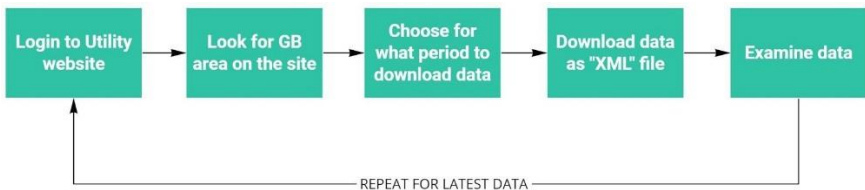
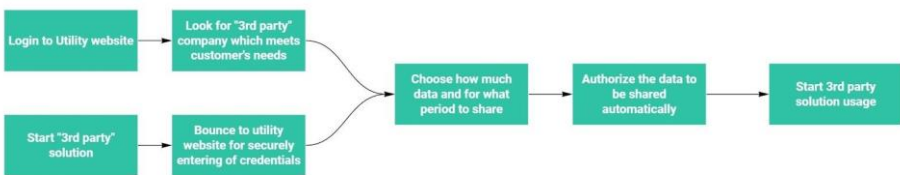
e-elering - customer portal that functions as a hub of all Elering's e-services and applications by other energy service providers and allows customers to access data related to their energy consumption and production. e-elering runs on Elering's smart grid platform Estfeed. e-elering can be used by customers to view data on the energy consumption and production of measuring points, manage contracts of energy service providers, grant third parties rights to view their data or represent them, monitor who has access to specific data, and see who has used this data.

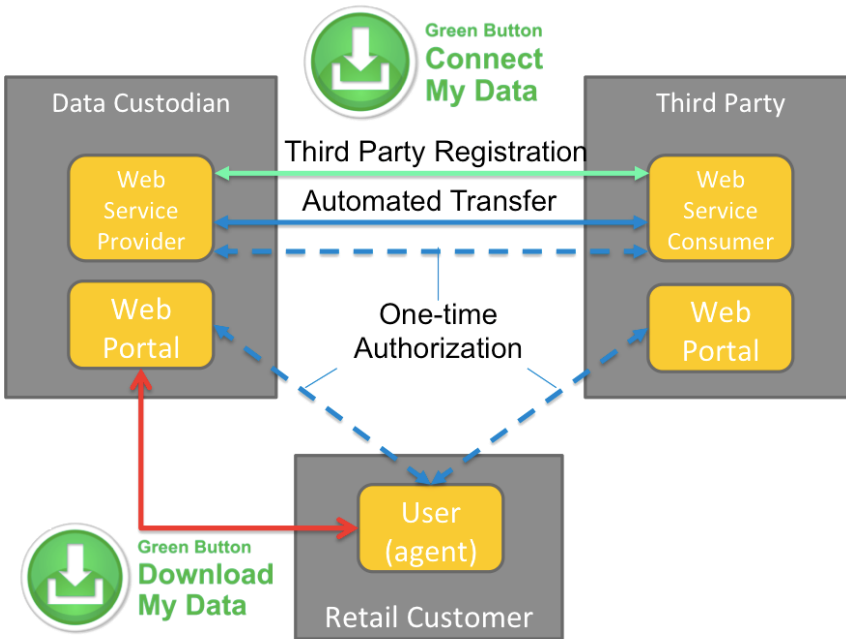
Besides, e-elering functions as a hub of applications, some of which are interfaced with Estfeed, Elering's smart grid platform. These interfaced applications use the Estfeed platform to request, receive, and forward data, and they have entered into respective agreements with Elering.

Elering customer portal gives market participants access to their meter data and enables them to download the data. The portal also provides the market participant with an overview of all information concerning them found on the Data Hub: agreement deadlines, open suppliers, hourly meter data, the market participant's EIC code, and the EIC codes of the metering points linked to the market participant. Each market participant can provide authorizations for accessing the meter data from previous periods via the customer portal; this is mainly to enable them to receive personalized offers from open suppliers. The market participant's data can be accessed by those market participants that have a statutory right to access the data

	or that have received an authorization from the market participant.
Technology	<p>Data Hub is based on MySQL database, meter data are transferred via AMQP protocol via Internet.</p> <p>Estfeed was built using Java and uses X-Road as a layer of secure data exchange and user authorization.</p>
Development plans	<p>Future development areas:</p> <ul style="list-style-type: none"> • connecting additional services – e.g. district heating meter data, IoT data, flexibility data, • exchange of near-real-time data, • cross border data exchange, • cross-sector data exchange.
References	<p>Elering website: https://elering.ee/en/data-exchange, https://elering.ee/en/smart-grid-development</p> <p>e-elering portal: https://e.elering.ee</p> <p>Republic of Estonia IS Authority portal: https://www.ria.ee/en/state-information-system/x-tee.html</p> <p>Data Hub Guide for Using and Joining Data Hub: https://www.elering.ee/sites/default/files/attachments/EL_Guide%20for%20Using%20and%20Joining%20Data%20Hub_2019_04.pdf</p>

ANNEX III: TABLE A.5 GREEN BUTTON DESCRIPTION

Solution name	Green Button
Owner	The Green Button Alliance
Operator	The Green Button Alliance
Implementation phase	Works since January of 2011
Purpose	Fostering global adoption of the Green Button standard to empower availability of metering water/energy consumption data to consumers
Description	<p>Green Button (GB) is a standard that helps utility companies provide consumption time-series data (e.g. electricity, gas, water data) to the customer directly from utility website (in CSV or XML format) or indirectly (via sharing data with 3rd party applications). GB standard works with the existing utility system. It does not require the installation of extra hardware inside the homes. GB framework is applied in U.S. and Canada. NAESB ESPI standard was carried out by OpenADE group and PAP 10. It allows transmit information securely and defines a standard energy usage data model. GB falls under international standards and works at different scales (industrial and residential).</p> <p>GB developed 2 use cases for customers to have access to their consumption data. The figures below represent the flow of this 2 use-cases:</p>  <p>ANNEX III: FIGURE A.4 DOWNLOAD MY DATA SCENARIO</p>  <p>ANNEX III: FIGURE A.5 CONNECT MY DATA SCENARIO</p> <p>Green Button Standards & Specifications</p> <p>OpenADE: Requirements specification for secure delivery of historical and ongoing usage information to 3rd Party</p> <p>PAP 10: Seed standard that defines a common energy data usage information model, for use across and interoperability between multiple standards</p> <p>NAESB ESPI: Standard that satisfies the requirements laid out in OpenADE and incorporates the data model from NAESB PAP 10 Energy Usage Information</p>

	<p>Green Button: File format subset of ESPI provides usage information to the consumer's via Web site [3]</p>
Modules	<p>The OpenESPI implementation consists of two instances: Data Custodian and a Third Party. Each implements the similar roles of the ESPI standard. These components can be utilized as the starting point for formal implementation based on the standard. Additionally, they can be used to test implementations against a working reference. It is designed that the DataCustodian and ThirdParty can be utilized by a test harness to orchestrate conformance tests that include proper and improper behaviour to verify the robustness of an implementation.</p> <p>All Green Button Actors presented in the figure below</p>  <p>The diagram illustrates the Green Button actors and their interactions. It features three main components: Data Custodian, Third Party, and Retail Customer. The Data Custodian contains a Web Service Provider and a Web Portal. The Third Party contains a Web Service Consumer and a Web Portal. The Retail Customer contains a User (agent). Interactions include: Third Party Registration (green arrow from Third Party to Data Custodian), Automated Transfer (blue arrow from Data Custodian to Third Party), One-time Authorization (dashed blue arrow from both Web Portals to the User), and data download (red arrow from Data Custodian Web Portal to User, and green arrow from Third Party Web Portal to User). Two green buttons are shown: 'Green Button Connect My Data' and 'Green Button Download My Data'.</p> <p>ANNEX III: FIGURE A.6 GREEN BUTTON ACTORS</p> <p>Each component exposes four interfaces:</p> <ol style="list-style-type: none"> 1. Authentication - used for implementing the OAuth authentication mechanism 2. Transfer - used for exchanging Energy Usage Information (EUI) according to the ESPI data model 3. Back End - used to simulate a back-end repository of usage information primarily in the Data Custodian 4. Test Orchestration - a test interface that can direct the code to implement scripted excellent and lousy behaviour designed to test the interface. <p>It is the stated goal of this development to address specifically the implementation and conformance testing requirements of the UCAIug OpenADE Task Force. These requirements are linked to this development and constrain the releases to perform</p>

to them. The OpenADE task force will provide to this project a document(s) describing the “Implementation Agreement and Certification and Test Suite for ESPI and Green Button” that will circumscribe the behaviour and capabilities of this software. [3]

Technology

OpenESPI

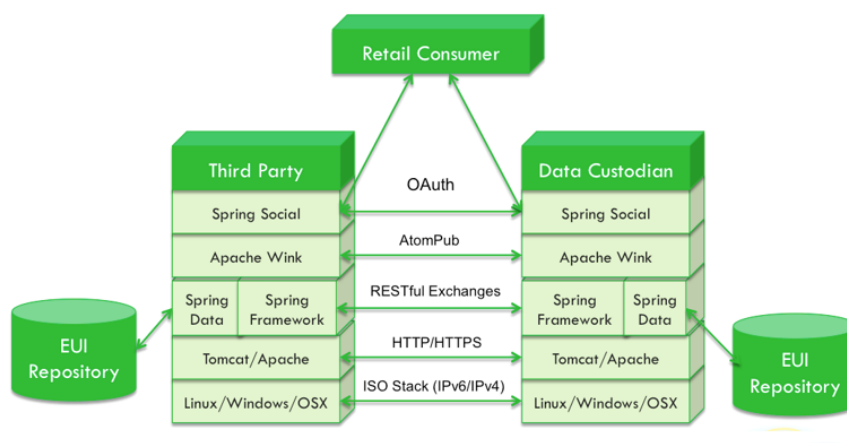
“The Energy Services Provider Interface (ESPI) provides a way for Energy Usage Information (EUI) to be shared, in a controlled manner, between participants in the energy services markets.

The OpenESPI project provides support for the development of deployable ESPI components that will help to rapidly and consistently engage the community with this exciting and enabling technology.” [3]

- **Actors:**
 - Consumer
 - Data Custodian
 - EnergyServiceProvider
- **InformationElements:**
 - Representations
 - EUI Data Repository
- **UseCases/Sequences:**
 - Scenario Automaton
 - OAuth/AtomPub Patterns
 - Orchestration and Deployment
- **Frameworks:**
 - Spring Model/View/Controller
 - Spring Social (OAuth)
 - Apache Wink (AtomPub)
- **Development Platform:**
 - Eclipse Projects
 - Ubuntu VM
 - Java Builds
 - C++ Builds (future)
- **Testing Plans:**
 - JUnit + Spring Testing Framework
 - Selenium Automations

ANNEX III: FIGURE A.7 OPENESPI SOFTWARE ARCHITECTURE

(Wollman, 2012, slide 41)

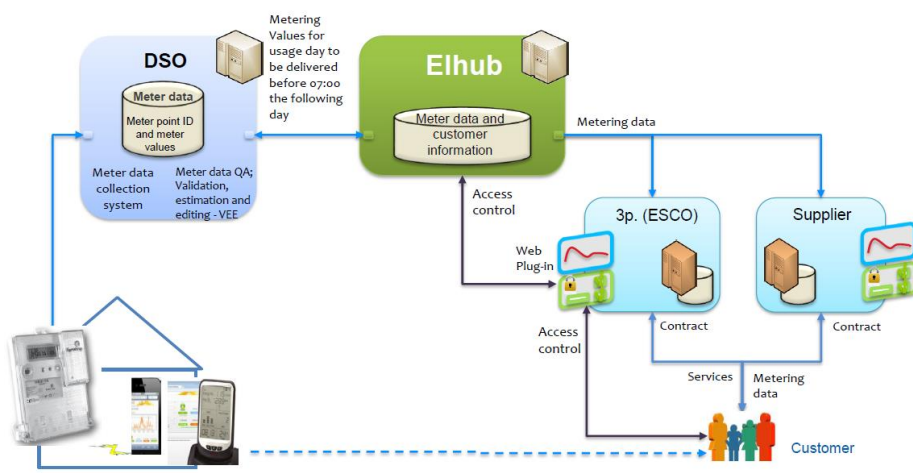


ANNEX III: FIGURE A.8 OPENESPI FRAMEWORKS

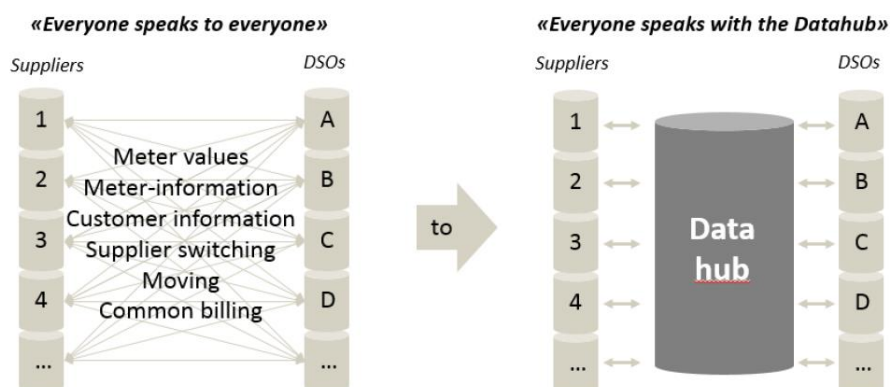
(Wollman, 2012, slide 42)

Development plans	<ol style="list-style-type: none"> 1. Offering GBA Certification for 3rd Parties' implementations /apps 2. Pursuing a green Button Directory service 3. Supporting International Green Button Ecosystem Growth 4. Driving Green Button Education & Member Co-Marketing 5. Featuring Green Button Case Studies 6. Welcoming new members to help lead the Utility Industry's Data-Access & Secure Data-Sharing Standard initiatives. 7. Development of GB Directory Service to connect utilities with 3rd parties, provide ratings for the apps [2]
References	<ol style="list-style-type: none"> 1. Wollman, D. (2012). <i>An Introduction to Green Button</i>. Presentation, slides 41, 42. 2. Green Button Alliance 2019 Annual General Meeting, 2019 3. Green Button Developer. Retrieved from: https://openei.org/wiki/Green_Button_Developer#Project_Description

ANNEX III: TABLE A.6 ELHUB DESCRIPTION

Solution name	Elhub
Owner	Elhub AS (subsidiary of Statnett, Norwegian TSO)
Operator	Elhub AS
Implementation phase	Production phase since Feb 2019
Purpose	The Elhub is a central IT system for collecting, storing, aggregating and sharing metering data for all energy consumption and production in Norway and for supporting power market processes.
Description	<p>Elhub has two main functions: handling metering data and supporting market processes.</p>  <p>The diagram illustrates the data flow from DSOs to the Elhub and then to various stakeholders. DSOs send metering data (including meter point ID and meter values) to the Elhub. The Elhub stores this data and provides access control to 3rd parties (ESCOs) and Suppliers. These parties then provide services and metering data to the end user (Customer). The diagram also shows a 'Web Plug-in' for access control and a 'Contract' between the 3rd parties and the Customer.</p> <p>ANNEX III: FIGURE A.9 METER DATA HANDLING AND END USER ACCESS PERSONAL DATA PROCESS</p> <p>Elhub supports the distribution and aggregation of metering data for all energy consumption and production in Norway. Grid companies are required to provide all metering data (hourly values) by 07:00 for the previous day, and Elhub will make available the metering data to energy suppliers, third parties and end users by 09:00. Elhub also calculates the basis for balance settlement and reports it to eSett.</p>

Grid companies are responsible for data quality, but Elhub provides the necessary aggregations and reporting.



ANNEX III: FIGURE A.10 SUPPORTING MARKET PROCESSES

Elhub provides one neutral interface between monopoly (DSO's) and the market (Suppliers). All messages regarding Norwegian energy market's processes such as supplier switches, relocations and updates of master data are sent directly to the data hub, which has the task of verifying and responding to the messages, as well as passing correct information on to the affected parties.

Modules

Elhub – the central part of the system responsible for processing metering data, processing messages regarding energy market processes and other functions such as data aggregation, settlement calculation, reporting.

Elhub Web Plugin (Plugin) is the part of Elhub where the end user can access their information via the market participants' websites. In the Plugin, the end user finds all their metering points with associated metering data and contracts. Besides, there is an overview of accesses, processing accesses and management of storage period for metering data. Only the end user can view and change information through the Plugin. Authorised user can also manage rights to access own data for other users and third-party companies.

The plugin is integrated in the form of a visible link to the plugin from the energy market players' websites. The plugin is a standalone web application that uses the ID port as the authentication solution.

Technology

Elhub is based on Siemens EnergyIP solution and implemented using Oracle EE Database and Oracle Fusion Middleware (SOA, OSB, WLS, OAG, OVD, OEM).

Development plans

Future development areas:

- Single invoice,
- Full "supplier centric market model",
- New tariff model allowing for capacity tariffs for residential and SMB.

References

Elhub AS: <https://elhub.no/>

	Statnett, Norwegian TSO: https://www.statnett.no/	
	Siemens	EnergyIP info:
	https://new.siemens.com/global/en/products/energy/energy-automation-and-smart-grid/grid-applications/energyip.html	

Analysis of existing solutions regarding requirements

The following subsections contain an analysis of existing solutions. Each of the requirements was evaluated and rated as described in the table below.

ANNEX III: TABLE A.7 EVALUATION RATES

Rating	Description
N/A	Not analysed. Requirement applies to components out of scope of analysis.
0	The analysed solution is not designed for this kind of requirement
1	The analysed solution does not meet the requirement, but some of its functionality can be used to meet this kind of requirements
2	The solution partially meets the requirement (no more than 75%)
3	The solution meets the requirement (75% and more)

ANNEX III: TABLE A.8 ENTSO-E OPDE EVALUATION TABLE

SUCs / Requirements		Solution: ENTSO-E OPDE	
ID	Name	Rating	Comment
SUC: Aggregate energy data			
AGG-ED-REQ-1	○ Standard rules to aggregate data in order not to enable the identification of persons behind data	0	OPDE data sources do not contain any personal data or data enabling the identification of individual persons.
AGG-ED-REQ-2	○ Standard rules to aggregate data in order to ensure the comparability of aggregated data sets	0	See AGG-ED-REQ-1
AGG-ED-REQ-3	○ Data source (e.g. meter data hub) ability to aggregate data	0	Data sources are out of OPDE.
AGG-ED-REQ-4	○ DEP ability to forward aggregated data from data source to data user	3	ECP is not restricted to specific data types, any kind of data could be sent via ECP.
SUC: Anonymize energy data			
ANO-ED-REQ-1	○ Standard rules to anonymize data not to enable the identification of persons behind data	0	OPDE data sources do not contain any personal data or data enabling the identification of individual persons.
ANO-ED-REQ-2	○ Standard rules to anonymize data in order to ensure the comparability of anonymized data sets	0	See ANO-ED-REQ-1.
ANO-ED-REQ-3	○ Data source (e.g. meter data hub) ability to anonymize data	0	Data sources are out of OPDE.
ANO-ED-REQ-4	○ DEP ability to forward anonymized data from data source to data user	3	ECP is not restricted to specific data types, any kind of data could be sent via ECP.

SUC: Authenticate data users			
AUTH-REQ-1	<ul style="list-style-type: none"> Right to access own data 	0	<p>OPDE data sources do not contain any personal data or data belonging to individual persons. There is no need to share data with citizens.</p> <p>Every data owner (eg. TSO) has access to own data through OPDE client application.</p>
AUTH-REQ-2	<ul style="list-style-type: none"> Authentication tools 	1	<p>OPDE client application uses authentication based on user/password pair. Local authentication is required to perform tasks in OPDE client applications. Password policies are defined by local administrator.</p> <p>Communication between nodes in OPDE network is secured by PKI. Nodes are verified by certificates.</p> <p>In the OPDE no other authentication methods are used.</p>
AUTH-REQ-3	<ul style="list-style-type: none"> Ability to share information related to representation rights between data users and concerned Customer Portals 	0	OPDE does not support representation of data owner.
AUTH-REQ-4	<ul style="list-style-type: none"> Ability to share authentication information between data users, Customer Portal and Authentication Service Provider 	0	OPDE does not use the Authentication Service Provider. See AUTH-REQ-2.
SUC: Manage access permissions			
AUTHZN-REQ1	<ul style="list-style-type: none"> Every person needs access permission 	0	<p>OPDE does not process any personal data or data belonging to individual persons. There is no need to share data with citizens.</p> <p>Recipient of message sent via OPDE is allowed to access to message content by default. No other parties (except sender and receiver) have access to the data.</p>
AUTHZN-REQ2	<ul style="list-style-type: none"> Valid identity of the person receiving access permissions 	1	Application access rights cover data access permission. Organizations for their employee create every application user account.
AUTHZN-REQ3	<ul style="list-style-type: none"> Ability to share access permissions between data owners, concerned DEPs, applications and data sources 	0	Sharing access permissions is not supported by OPDE.
SUC: Collect energy data			

DC-REQ1	○ Collection of meter data	-	OPDE does not process meter data. Some OPDE components (ECP/EDX) can be used for non-real-time data exchange.
DC-REQ1.1	▪ Get near-real-time data (up to 1 hour) from meters	0	OPDE is based (by design) on the ATOM network, which is designed for non-real-time operational exchanges.
DC-REQ1.2	▪ Get historical data (monthly) from conventional meters	0	See DC-REQ1
DC-REQ1.3	▪ Store data in meter data hub	0	OPDE does not store meter data.
DC-REQ2	○ Collection of market data	-	OPDE does not process market data. Some OPDE components (ECP/EDX) can be used for non-real-time data exchange.
DC-REQ2.1	▪ Get near-real-time (up to 1 hour) data from market	0	OPDE is based (by design) on the ATOM network, which is designed for non-real-time operational exchanges.
DC-REQ2.2	▪ Get historical data from market	0	See DC-REQ2
DC-REQ2.3	▪ Store data in market data hub	0	OPDE does not store market data.
DC-REQ3	○ Collection of grid data	-	OPDE does not process grid data (except of grid models). One of the Business Requirements for OPDE is to support Outage Planning Coordination (not implemented yet). Some OPDE components (ECP/EDX) can be used to non-real-time data exchange.
DC-REQ3.1	▪ Get very-near-real-time (up to 1 minute) data from grid	0	OPDE is based (by design) on the ATOM network, which is designed for non-real-time operational exchanges.
DC-REQ3.2	▪ Get near-real-time (up to 1 hour) data from grid	0	See DC-REQ3
DC-REQ3.3	▪ Get historical data from grid	0	See DC-REQ3.2
DC-REQ3.4	▪ Store data in grid data hub	0	OPDE does not store grid data.
SUC: Transfer energy data			
DT-REQ1	○ Transfer of data must be secured, by means of encryption or communication protocol	3	ECP/EDX is based on MADES standard and designed for transfer any types of data for non-real-time operations. OPDE can be used

			for large volume data exchange. Delivery time is not guaranteed. Data is transferred between endpoints, without central hub. EDX provides publish/subscribe mechanism.
DT-REQ2	<ul style="list-style-type: none"> Data portability (applies to personal data - Article 20 of the GDPR) 	0	OPDE does not process personal data
DT-REQ3	<ul style="list-style-type: none"> Data owner's access to data through DEP (and foreign DEP) 	0	OPDE data sources do not contain any personal data or data belonging to individual persons. There is no need to share data with citizens. Every data owner (eg. TSO) has access to own data through OPDE client application.
DT-REQ4	<ul style="list-style-type: none"> Application's access to data through DEP (and foreign DEP) 	1	ECP provides API for app access (webservice, amqp, file-based interface), excluding personal data.
SUC: Exchange data between DER and SCADA			
DER-SCADA-REQ4	<ul style="list-style-type: none"> Ability of DEP to forward real-time data from DER's to System Operators 	0	OPDE and all of its components are designed for non-real-time operations.
DER-SCADA-REQ5	<ul style="list-style-type: none"> Ability of DEP to forward very-near-real-time (up to 1 minute) data from DER's to System Operators 	0	See DER-SCADA-REQ4
DER-SCADA-REQ6	<ul style="list-style-type: none"> Ability of DEP to forward near-real-time (up to 1 hour) data from DER's to System Operators 	3	ECP is not restricted to specific data types, any kind of data could be transferred via ECP.
DER-SCADA-REQ7	<ul style="list-style-type: none"> Ability of DEP to forward activation requests from System Operators to DER 	1	OPDE platform can be used for communication for non-real-time flexibilities activation and cannot be use for flexibilities activated in real-time.
DER-SCADA-REQ2	<ul style="list-style-type: none"> Communication link between DEP and SO's SCADA 	1	OPDE is not connected to SCADA systems. Using the provided API, OPDE can be connected to SCADA systems on the client side. There is no real use-cases.
DER-SCADA-REQ1	<ul style="list-style-type: none"> Encrypted data exchange 	3	Messages transferred via ECP are electronically signed and encrypted at the ECP endpoint.
DER-SCADA-REQ3	<ul style="list-style-type: none"> Safety of DER's IT infrastructure 	N/A	DER's infrastructure is out of scope of analysis.

SUC: Manage flexibility activations			
FA-REQ2	<ul style="list-style-type: none"> Exchange of activation requests through DEP and flexibility platform 	1	OPDE platform can be used for communication for non-real-time flexibilities activation and cannot be use for flexibilities activated in real-time.
FA-REQ1	<ul style="list-style-type: none"> Automated activation of devices is possible 	0	OPDE does not support interoperability with devices
SUC: Calculate flexibility baseline			
FB-REQ1	<ul style="list-style-type: none"> Ability of flexibility platform to collect input for baseline calculation, incl. through DEP 	1	OPDE does not process data required for baseline calculation. ECP is not restricted to specific data types, any kind of data could be transferred via ECP and stored on any OPDE node.
FB-REQ2	<ul style="list-style-type: none"> Ability of flexibility platform to compute baseline 	0	OPDE does not support baseline computing.
SUC: Manage flexibility bids			
FBIDS-REQ2	<ul style="list-style-type: none"> Ability to exchange information on System Operators' flexibility need and FSPs' flexibility potential through flexibility platform (and DEP) 	1	OPDE does not support bidding process at all. ECP is not restricted to specific data types, any kind of data could be transferred via ECP and stored on any OPDE node.
FBIDS-REQ4	<ul style="list-style-type: none"> Algorithm for prequalification of flexibility providers 	0	OPDE does not support bidding process at all.
FBIDS-REQ6	<ul style="list-style-type: none"> Flexibility platform's ability to collect bids from FSPs 	1	ECP is not restricted to specific data types, any kind of data could be transferred via ECP and stored on any OPDE node.
FBIDS-REQ7	<ul style="list-style-type: none"> Selection of successful bids 	0	See FBIDS-REQ4
FBIDS-REQ8	<ul style="list-style-type: none"> Flexibility platform's ability to collect grid validation results from SOs 	1	See FBIDS-REQ6
FBIDS-REQ9	<ul style="list-style-type: none"> Calculation of grid impacts (congestion, imbalance) 	0	See FBIDS-REQ4
FBIDS-REQ3	<ul style="list-style-type: none"> Auction process supervised by Market Operator 	0	See FBIDS-REQ4
FBIDS-REQ5	<ul style="list-style-type: none"> Automated exchange of bids is possible 	1	See FBIDS-REQ6
FBIDS-REQ1	<ul style="list-style-type: none"> Secure communication 	3	Messages transferred via ECP are electronically signed and encrypted at the ECP endpoint.
SUC: Predict flexibility availability			
FPRED-	<ul style="list-style-type: none"> Collection of data for prediction (long term - 	1	OPDE does not support flexibility prediction

REQ1	years)		process at all. ECP is not restricted to specific data types, any kind of data could be transferred via ECP and stored on any OPDE node.
FPRED-REQ2	○ Computation of predictions (long term - years)	0	See FPRED-REQ1
FPRED-REQ1	○ Collection of data for prediction (medium term - days to years ahead)	1	See FPRED-REQ1
FPRED-REQ2	○ Computation of predictions (medium term - days to years ahead)	0	See FPRED-REQ1
FPRED-REQ1	○ Collection of data for prediction (short term - intraday operation)	1	See FPRED-REQ1
FPRED-REQ2	○ Computation of predictions (long term - intraday operation)	0	See FPRED-REQ1
SUC: Verify and settle activated flexibilities			
FVERIF-REQ1	○ Calculation of actually delivered flexibility as response to activation request	0	OPDE does not support flexibility verification process at all.
FVERIF-REQ2	○ Verification that flexibility delivered matches with flexibility requested	0	See FVERIF-REQ1
FVERIF-REQ3	○ Calculation of the penalty if flexibility delivered is less than flexibility requested	0	See FVERIF-REQ1
SUC: Provide list of suppliers and ESCOs			
ESCO-REQ1	○ List of suppliers and ESCOs is available through DEP; List of aggregators is available through flexibility platform additionally	1	Lists of suppliers, ESCOs and aggregators are not available on OPDE. ECP is not restricted to specific data types, any kind of data could be transferred via ECP and stored on any OPDE node and can be used as a communication layer between systems exchanging such information.
SUC: Erase and rectify personal data			
PERSO-DATA-REQ1	○ Ability to share information related to erasure of personal data between data owners, concerned DEPs, applications and data sources	1	OPDE do not contain or process any personal data or data belonging to individual persons. There is no need to comply with GDPR requirements. ECP is not restricted to specific data types, any kind of data could be transferred via ECP and stored on any OPDE node
PERSO-	○ Ability to share information related to rectification of	1	See PERSO-DATA-REQ1

DATA-REQ2	personal data between data owners, concerned DEPs, applications and data sources		
SUC: Manage data logs			
LOGS-REQ1	<ul style="list-style-type: none"> Ability to share information related to data logs between data owners, concerned DEPs, applications and data sources 	1	<p>There is no exchange of information on the use of data between parties on the OPDE platform.</p> <p>ECP is not restricted to specific data types, any kind of data could be transferred via ECP and stored on any OPDE node.</p>
SUC: Manage sub-meter data			
SUBMET-REQ1	<ul style="list-style-type: none"> Collection of data from sub-meters 	1	<p>OPDE does not store sub-meter data.</p> <p>ECP is not restricted to specific data types, any kind of data could be transferred via ECP and stored on any OPDE node</p>
SUBMET-REQ3	<ul style="list-style-type: none"> Storing sub-meter data in data hub 	0	<p>OPDE does not store sub-meter data.</p>
SUBMET-REQ2	<ul style="list-style-type: none"> Ability of DEP to forward sub-meter data from data hub to customer (data owner) and application (energy service provider) 	1	<p>ECP is not restricted to specific data types, any kind of data could be transferred via ECP and stored on any OPDE node, except of personal data.</p>
SUBMET-REQ7	<ul style="list-style-type: none"> Ability of DEP to forward activation orders from customer (data owner) or application (energy service provider) to devices 	1	<p>See SUBMET-REQ2</p>
SUBMET-REQ4	<ul style="list-style-type: none"> Data format of sub-metering 	1	<p>ECP is agnostic to the file type, any file type could be transferred via ECP</p>
SUBMET-REQ5	<ul style="list-style-type: none"> Transmission protocols of sub-metering 	0	<p>OPDE does not support transmission protocols for sub-metering: MQTT, ISO/IEC PRF 20922. TLS and RFC6176 are supported.</p>
SUBMET-REQ6	<ul style="list-style-type: none"> SLA between customer and energy service provider 	0	

ANNEX III: TABLE A.9 ESTONIAN DATA EXCHANGE PLATFORM ESTFEED, DATAHUB, E-ELERING PORTAL EVALUATION TABLE

SUCs / Requirements		Solution: 2.2. Estonian Data Exchange Platform	
ID	Name	Rating	Comment
SUC: Aggregate energy data			
AGG-ED-REQ-1	○ Standard rules to aggregate data in order not to enable the identification of persons behind data	3	<i>(Data Hub reports to suppliers, NOs, BRPs, public information)</i>
AGG-ED-REQ-2	○ Standard rules to aggregate data in order to ensure the comparability of aggregated data sets	3	See AGG-ED-REQ-1
AGG-ED-REQ-3	○ Data source (e.g. meter data hub) ability to aggregate data	3	See AGG-ED-REQ-1
AGG-ED-REQ-4	○ DEP ability to forward aggregated data from data source to data user	3	<i>Estfeed</i>
SUC: Anonymize energy data			
ANO-ED-REQ-1	○ Standard rules to anonymize data not to enable the identification of persons behind data	2	Not standard procedure. Its true for Data Hub. Trialled once by anonymising a set of data for a full year and all consumers, except large consumers
ANO-ED-REQ-2	○ Standard rules to anonymize data in order to ensure the comparability of anonymized data sets	2	See ANO-ED-REQ-1
ANO-ED-REQ-3	○ Data source (e.g. meter data hub) ability to anonymize data	2	See ANO-ED-REQ-1.
ANO-ED-REQ-4	○ DEP ability to forward anonymized data from data source to data user	3	<i>Estfeed</i>
SUC: Authenticate data users			
AUTH-REQ-1	○ Right to access own data	3	Customer portal e-Elering
AUTH-REQ-2	○ Authentication tools	3	Smart-ID, bank links, ID-card, Mobil-ID
AUTH-REQ-3	○ Ability to share information related to representation rights between data users and concerned Customer Portals	3	Standard functionality of e-Elering Portal
AUTH-REQ-4	○ Ability to share authentication information between data users, Customer Portal and Authentication Service Provider	3	See AUTH-REQ-2

SUC: Manage access permissions			
AUTHZN-REQ1	○ Every person needs access permission	3	Access permission can be granted in e-Elering
AUTHZN-REQ2	○ Valid identity of the person receiving access permissions	3	e-Elering needs to know the identity of the person receiving access permission
AUTHZN-REQ3	○ Ability to share access permissions between data owners, concerned DEPs, applications and data sources	3	<i>Estfeed</i>
SUC: Collect energy data			
DC-REQ1	○ Collection of meter data		
DC-REQ1.1	▪ Get near-real-time data (up to 1 hour) from meters	3	<i>Data Hub (hourly data)</i>
DC-REQ1.2	▪ Get historical data (monthly) from conventional meters	3	<i>Data Hub</i>
DC-REQ1.3	▪ Store data in meter data hub	3	<i>Data Hub</i>
DC-REQ2	○ Collection of market data		
DC-REQ2.1	▪ Get near-real-time (up to 1 hour) data from market	0	
DC-REQ2.2	▪ Get historical data from market	0	
DC-REQ2.3	▪ Store data in market data hub	0	
DC-REQ3	○ Collection of grid data		
DC-REQ3.1	▪ Get very-near-real-time (up to 1 minute) data from grid	0	
DC-REQ3.2	▪ Get near-real-time (up to 1 hour) data from grid	0	
DC-REQ3.3	▪ Get historical data from grid	0	
DC-REQ3.4	▪ Store data in grid data hub	0	
SUC: Transfer energy data			
DT-REQ1	○ Transfer of data must be secured, by means of encryption or communication protocol	3	<i>Estfeed</i>
DT-REQ2	○ Data portability (applies to personal data - Article 20 of the GDPR)	3	<i>Estfeed</i>

DT-REQ3	<ul style="list-style-type: none"> ○ Data owner's access to data through DEP (and foreign DEP) 	3	<i>Estfeed - Domestic data exchange currently only</i>
DT-REQ4	<ul style="list-style-type: none"> ○ Application's access to data through DEP (and foreign DEP) 	3	See DT-REQ3
SUC: Exchange data between DER and SCADA			
DER-SCADA-REQ4	<ul style="list-style-type: none"> ○ Ability of DEP to forward real-time data from DER's to System Operators 	1	Technically it is possible, no real use-cases
DER-SCADA-REQ5	<ul style="list-style-type: none"> ○ Ability of DEP to forward very-near-real-time (up to 1 minute) data from DER's to System Operators 	1	See DER-SCADA-REQ4
DER-SCADA-REQ6	<ul style="list-style-type: none"> ○ Ability of DEP to forward near-real-time (up to 1 hour) data from DER's to System Operators 	1	See DER-SCADA-REQ4
DER-SCADA-REQ7	<ul style="list-style-type: none"> ○ Ability of DEP to forward activation requests from System Operators to DER 	1	See DER-SCADA-REQ4
DER-SCADA-REQ2	<ul style="list-style-type: none"> ○ Communication link between DEP and SO's SCADA 	1	See DER-SCADA-REQ4
DER-SCADA-REQ1	<ul style="list-style-type: none"> ○ Encrypted data exchange 	1	See DER-SCADA-REQ4
DER-SCADA-REQ3	<ul style="list-style-type: none"> ○ Safety of DER's IT infrastructure 	N/A	
SUC: Manage flexibility activations			
FA-REQ2	<ul style="list-style-type: none"> ○ Exchange of activation requests through DEP and flexibility platform 	1	Technically it is possible, no real use-cases
FA-REQ1	<ul style="list-style-type: none"> ○ Automated activation of devices is possible 	1	See FA-REQ2
SUC: Calculate flexibility baseline			
FB-REQ1	<ul style="list-style-type: none"> ○ Ability of flexibility platform to collect input for baseline calculation, incl. through DEP 	1	Technically it is possible, no real use-cases
FB-REQ2	<ul style="list-style-type: none"> ○ Ability of flexibility platform to compute baseline 	0	
SUC: Manage flexibility bids			
FBIDS-REQ2	<ul style="list-style-type: none"> ○ Ability to exchange information on System Operators' flexibility need 	1	Technically it is possible, no real use-cases

	and FSPs' flexibility potential through flexibility platform (and DEP)		
FBIDS-REQ4	○ Algorithm for prequalification of flexibility providers	0	
FBIDS-REQ6	○ Flexibility platform's ability to collect bids from FSPs	1	See FBIDS-REQ2
FBIDS-REQ7	○ Selection of successful bids	0	
FBIDS-REQ8	○ Flexibility platform's ability to collect grid validation results from SOs	1	See FBIDS-REQ2
FBIDS-REQ9	○ Calculation of grid impacts (congestion, imbalance)	0	
FBIDS-REQ3	○ Auction process supervised by Market Operator	0	
FBIDS-REQ5	○ Automated exchange of bids is possible	1	See FBIDS-REQ2
FBIDS-REQ1	○ Secure communication	3	
SUC: Predict flexibility availability			
FPRED-REQ1	○ Collection of data for prediction (long term - years)	1	Technically it is possible, no real use-cases
FPRED-REQ2	○ Computation of predictions (long term - years)	0	
FPRED-REQ1	○ Collection of data for prediction (medium term - days to years ahead)	1	Technically it is possible, no real use-cases
FPRED-REQ2	○ Computation of predictions (medium term - days to years ahead)	0	
FPRED-REQ1	○ Collection of data for prediction (short term - intraday operation)	1	Technically it is possible, no real use-cases
FPRED-REQ2	○ Computation of predictions (long term - intraday operation)	0	
SUC: Verify and settle activated flexibilities			
FVERIF-REQ1	○ Calculation of actually delivered flexibility as response to activation request	1	Data transfer is technically possible, no real use-cases.
FVERIF-REQ2	○ Verification that flexibility delivered matches with flexibility requested	1	See FVERIF-REQ2
FVERIF-	○ Calculation of the penalty if flexibility delivered is less	0	

REQ3	than flexibility requested		
SUC: Provide list of suppliers and ESCOs			
ESCO-REQ1	<ul style="list-style-type: none"> List of suppliers and ESCOs is available through DEP; List of aggregators is available through flexibility platform additionally 	3	Estfeed (suppliers, apps)
SUC: Erase and rectify personal data			
PERSO-DATA-REQ1	<ul style="list-style-type: none"> Ability to share information related to erasure of personal data between data owners, concerned DEPs, applications and data sources 	3	All GDPR requirements are met. Estonian regulator has confirmed this. Data owners can modify and withdraw their data, access permissions, agreement with e-Elering Terms of Service, unless differently obliged by law.
PERSO-DATA-REQ2	<ul style="list-style-type: none"> Ability to share information related to rectification of personal data between data owners, concerned DEPs, applications and data sources 	3	See PERSO-DATA-REQ1
SUC: Manage data logs			
LOGS-REQ1	<ul style="list-style-type: none"> Ability to share information related to data logs between data owners, concerned DEPs, applications and data sources 	3	Estfeed
SUC: Manage sub-meter data			
SUBMET-REQ1	<ul style="list-style-type: none"> Collection of data from sub-meters 	0	
SUBMET-REQ3	<ul style="list-style-type: none"> Storing sub-meter data in data hub 	0	
SUBMET-REQ2	<ul style="list-style-type: none"> Ability of DEP to forward sub-meter data from data hub to customer (data owner) and application (energy service provider) 	1	Technically it is possible, no real use-cases
SUBMET-REQ7	<ul style="list-style-type: none"> Ability of DEP to forward activation orders from customer (data owner) or application (energy service provider) to devices 	1	See SUBMET-REQ2
SUBMET-REQ4	<ul style="list-style-type: none"> Data format of sub-metering 	1	Estfeed is agnostic to the file type, any file type could be transferred via Estfeed. Technically it is possible, no real use-cases
SUBMET-REQ5	<ul style="list-style-type: none"> Transmission protocols of sub-metering 	0	

SUBMET- REQ6	○ SLA between customer and energy service provider	0	
-----------------	---	---	--

ANNEX III: TABLE A.10 GREEN BUTTON EVALUATION TABLE

SUCs / Requirements		Solution: 2.3. Green Button	
ID	Name	Rating	Comment
SUC: Aggregate energy data			
AGG-ED-REQ-1	<ul style="list-style-type: none"> Standard rules to aggregate data in order not to enable the identification of persons behind data 	2	<p>GB provide standard rule how to organize consumption information but does not do it. This is the function of utility company. Any customer's consumption data cannot be shared without her/his permission. However, it is organized in a manner that it is impossible to identify the person behind the data.</p> <p>"There is no personally identifiable information (PII) contained within the standard Green Button data, it contains only measured interval usage information."</p>
AGG-ED-REQ-2	<ul style="list-style-type: none"> Standard rules to aggregate data in order to ensure the comparability of aggregated data sets 	2	The Green Button provides customers consumption data with in a standardized way.
AGG-ED-REQ-3	<ul style="list-style-type: none"> Data source (e.g. meter data hub) ability to aggregate data 	0	GB does not have data source. Utility companies manage data source.
AGG-ED-REQ-4	<ul style="list-style-type: none"> DEP ability to forward aggregated data from data source to data user 	3	Green Button can represent the user data in .xml or .csv format (can be extended to other formats by utility solution)
SUC: Anonymize energy data			
ANO-ED-REQ-1	<ul style="list-style-type: none"> Standard rules to anonymize data not to enable the identification of persons behind data 	2	Green Button usage data files do not include personal information (names, addresses, meter numbers). Green Button personal data files do not include usage information. They are kept in separate streams. Any customer's consumption data can not be shared without her/his permission .
ANO-ED-REQ-2	<ul style="list-style-type: none"> Standard rules to anonymize data in order to ensure the comparability of anonymized data sets 	2	See ANO-ED-REQ-1
ANO-ED-REQ-3	<ul style="list-style-type: none"> Data source (e.g. meter data hub) ability to anonymize data 	0	GB does not have data source. Utility companies manage data source.
ANO-ED-REQ-4	<ul style="list-style-type: none"> DEP ability to forward anonymized data from data 	3	Green Button can represent the user data in

REQ-4	source to data user		.xml or .csv format (can be extended to other formats by utility specific implementation)
SUC: Authenticate data users			
AUTH-REQ-1	<ul style="list-style-type: none"> Right to access own data 	3	The end user can access information related to their meter points, contracts and meter values via Utility Company website or 3 rd parties' applications certified by Green Button. To access customer data it is required from 3rd applications to gain authorization from consumers using the Internet Engineering Task Force (IETF) OAuth 2.0 Authorization Framework standards [RFC6749] and [RFC6750].
AUTH-REQ-2	<ul style="list-style-type: none"> Authentication tools 	2	Green Button does NOT use logins and passwords, as these are used to Authenticate an entity. The Green Button standard requires utilities to implement "The OAuth 2.0 Authorization Framework" (RFC 6749) standard's "Client Credential", "Authorization Code", and "Refresh Token" Grant flows, which generate OAuth 2.0 access tokens. The OAuth 2.0 access tokens are then required to access the utilities customer's data by the Third Party.
AUTH-REQ-3	<ul style="list-style-type: none"> Ability to share information related to representation rights between data users and concerned Customer Portals 	3	Enabling of automated transfers of Green Button data from a utility to 3 rd party service only if a customer has granted explicit permission.
AUTH-REQ-4	<ul style="list-style-type: none"> Ability to share authentication information between data users, Customer Portal and Authentication Service Provider 	0	GB does not use the Authentication Service Provider
SUC: Manage access permissions			
AUTHZN-REQ1	<ul style="list-style-type: none"> Every person needs access permission 	3	Green Button "Connect My Data" is a mechanism for a customer to authorize a third-party service provider to automate access to their utility data. Green Button Connect My Data uses authorization standards defined by the Internet Engineering Task force (IETF) to ensure the participants in the Green Button initiative are aligned with

			the mainstream of internet evolution.
AUTHZN-REQ2	<ul style="list-style-type: none"> Valid identity of the person receiving access permissions 	3	A typical method requires the consumer to provide authorization using a webpage, similar to how Facebook and Google based applications request users to approve access to their accounts. Once this authorization is granted, the application can automatically retrieve the consumer's energy data without any further involvement of the consumer. Before giving information to the 3rd parties GB check the authorization permission.
AUTHZN-REQ3	<ul style="list-style-type: none"> Ability to share access permissions between data owners, concerned DEPs, applications and data sources 	3	The customer retains full control of authorization and revocation of permissions.
SUC: Collect energy data			
DC-REQ1	<ul style="list-style-type: none"> Collection of meter data 		
DC-REQ1.1	<ul style="list-style-type: none"> Get near-real-time data (up to 1 hour) from meters 	3	The Green Button allows users to download their monthly, daily, hourly, or 15-minute interval energy usage data. Metered data is not currently transmitted or collected using the Green Button Energy Usage Information schema, although it is possible.
DC-REQ1.2	<ul style="list-style-type: none"> Get historical data (monthly) from conventional meters 	3	See DC-REQ1.1
DC-REQ1.3	<ul style="list-style-type: none"> Store data in meter data hub 	0	<i>Data stored at utility providers DB</i>
DC-REQ2	<ul style="list-style-type: none"> Collection of market data 		
DC-REQ2.1	<ul style="list-style-type: none"> Get near-real-time (up to 1 hour) data from market 	0	
DC-REQ2.2	<ul style="list-style-type: none"> Get historical data from market 	0	
DC-REQ2.3	<ul style="list-style-type: none"> Store data in market data hub 	0	
DC-REQ3	<ul style="list-style-type: none"> Collection of grid data 		
DC-REQ3.1	<ul style="list-style-type: none"> Get very-near-real-time (up to 1 minute) data from grid 	0	
DC-REQ3.2	<ul style="list-style-type: none"> Get near-real-time (up to 1 hour) data from grid 	0	
DC-	<ul style="list-style-type: none"> Get historical data from grid 	0	

REQ3.3			
DC-REQ3.4	<ul style="list-style-type: none"> Store data in grid data hub 	0	
SUC: Transfer energy data			
DT-REQ1	<ul style="list-style-type: none"> Transfer of data must be secured, by means of encryption or communication protocol 	3	Consumer can provide direct access to third parties so that they can retrieve measurement values as long as the end-user is associated with the measurement point and consumer permission is granted.
DT-REQ2	<ul style="list-style-type: none"> Data portability (applies to personal data - Article 20 of the GDPR) 	3	Data owner can download data. The customer retains full control over his consumption data.
DT-REQ3	<ul style="list-style-type: none"> Data owner's access to data through DEP (and foreign DEP) 	3	The customer retains full control of authorization and revocation of permissions.
DT-REQ4	<ul style="list-style-type: none"> Application's access to data through DEP (and foreign DEP) 	3	Green Button provides API for utility companies to represent consumer data and to 3 rd parties to enable analysis of the consumption data and building recommendations.
SUC: Exchange data between DER and SCADA			
DER-SCADA-REQ4	<ul style="list-style-type: none"> Ability of DEP to forward real-time data from DER's to System Operators 	0	
DER-SCADA-REQ5	<ul style="list-style-type: none"> Ability of DEP to forward very-near-real-time (up to 1 minute) data from DER's to System Operators 	0	
DER-SCADA-REQ6	<ul style="list-style-type: none"> Ability of DEP to forward near-real-time (up to 1 hour) data from DER's to System Operators 	0	
DER-SCADA-REQ7	<ul style="list-style-type: none"> Ability of DEP to forward activation requests from System Operators to DER 	0	
DER-SCADA-REQ2	<ul style="list-style-type: none"> Communication link between DEP and SO's SCADA 	0	
DER-SCADA-REQ1	<ul style="list-style-type: none"> Encrypted data exchange 	0	
DER-SCADA-REQ3	<ul style="list-style-type: none"> Safety of DER's IT infrastructure 	N/A	

SUC: Manage flexibility activations			
FA-REQ2	○ Exchange of activation requests through DEP and flexibility platform	0	
FA-REQ1	○ Automated activation of devices is possible	0	
SUC: Calculate flexibility baseline			
FB-REQ1	○ Ability of flexibility platform to collect input for baseline calculation, incl. through DEP	0	
FB-REQ2	○ Ability of flexibility platform to compute baseline	0	
SUC: Manage flexibility bids			
FBIDS-REQ2	○ Ability to exchange information on System Operators' flexibility need and FSPs' flexibility potential through flexibility platform (and DEP)	0	Green Button does not participate in the wholesale market bidding and contract process.
FBIDS-REQ4	○ Algorithm for prequalification of flexibility providers	0	
FBIDS-REQ6	○ Flexibility platform's ability to collect bids from FSPs	0	
FBIDS-REQ7	○ Selection of successful bids	0	
FBIDS-REQ8	○ Flexibility platform's ability to collect grid validation results from SOs	0	
FBIDS-REQ9	○ Calculation of grid impacts (congestion, imbalance)	0	
FBIDS-REQ3	○ Auction process supervised by Market Operator	0	
FBIDS-REQ5	○ Automated exchange of bids is possible	0	
FBIDS-REQ1	○ Secure communication	0	
SUC: Predict flexibility availability			
FPRED-REQ1	○ Collection of data for prediction (long term - years)	0	
FPRED-REQ2	○ Computation of predictions (long term - years)	0	
FPRED-REQ1	○ Collection of data for prediction (medium term - days to years ahead)	0	
FPRED-	○ Computation of predictions	0	

REQ2	(medium term - days to years ahead)		
FPRED-REQ1	<ul style="list-style-type: none"> Collection of data for prediction (short term - intraday operation) 	0	
FPRED-REQ2	<ul style="list-style-type: none"> Computation of predictions (long term - intraday operation) 	0	
SUC: Verify and settle activated flexibilities			
FVERIF-REQ1	<ul style="list-style-type: none"> Calculation of actually delivered flexibility as response to activation request 	0	
FVERIF-REQ2	<ul style="list-style-type: none"> Verification that flexibility delivered matches with flexibility requested 	0	
FVERIF-REQ3	<ul style="list-style-type: none"> Calculation of the penalty if flexibility delivered is less than flexibility requested 	0	
SUC: Provide list of suppliers and ESCOs			
ESCO-REQ1	<ul style="list-style-type: none"> List of suppliers and ESCOs is available through DEP; List of aggregators is available through flexibility platform additionally 	3	<p>A Green Button Alliance testing mark on the data custodian's site signifies the solution has been tested to be standards-compliant; and a listing in the GBA database helps promote tested solutions to a broader audience. Speeds data-sharing implementation time-to-market and assures a seamless and secure data-sharing experience for customers. Green Button standards ensure secure transmission of data and customer privacy, thus well-positioning utilities to meet regulations for customer privacy.</p> <p>To see solutions customer should go to GB Alliance membership page</p>
SUC: Erase and rectify personal data			
PERSO-DATA-REQ1	<ul style="list-style-type: none"> Ability to share information related to erasure of personal data between data owners, concerned DEPs, applications and data sources 	0	Personal data is beyond the scope of the GB.
PERSO-DATA-REQ2	<ul style="list-style-type: none"> Ability to share information related to rectification of personal data between data owners, concerned DEPs, applications and data sources 	1	Customers are allowed to request authorized access to their data by Third Parties be revoked.
SUC: Manage data logs			

LOGS-REQ1	<ul style="list-style-type: none"> Ability to share information related to data logs between data owners, concerned DEPs, applications and data sources 	1	The NAESB REQ.21 ESPI standard defines the authorization process, which includes controlling what data can be provided to Third Parties by the utilities when implementing the standard. How a utility performs authentication, data access, authentication and authorization logging is beyond the scope of the standard.
SUC: Manage sub-meter data			
SUBMET-REQ1	<ul style="list-style-type: none"> Collection of data from sub-meters 	3	Based on NAESB REQ.21 ESPI standard GB can provide sub meter data in readable format (if collected and made available by the utility).
SUBMET-REQ3	<ul style="list-style-type: none"> Storing sub-meter data in data hub 	0	GB does not store sub-meter data. It is a task of utility service provider.
SUBMET-REQ2	<ul style="list-style-type: none"> Ability of DEP to forward sub-meter data from data hub to customer (data owner) and application (energy service provider) 	3	See SUBMET-REQ1
SUBMET-REQ7	<ul style="list-style-type: none"> Ability of DEP to forward activation orders from customer (data owner) or application (energy service provider) to devices 	0	Activation/deactivation of a new device is beyond the scope of the NAESB REQ.21 ESPI standard (generally done as a business practice by utilities)
SUBMET-REQ4	<ul style="list-style-type: none"> Data format of sub-metering 	3	GB organize customers data represent in XML or .csv format (based on built solution utility company can provide in .json and other formats)
SUBMET-REQ5	<ul style="list-style-type: none"> Transmission protocols of sub-metering 	1	OpenADE protocol is used for data transfer between the utility and the Third Party based on NAESB REQ.21 ESPI standard (TLS 3.higher)
SUBMET-REQ6	<ul style="list-style-type: none"> SLA between customer and energy service provider 	3	Customer is aware which data is collected (e.g. bills, demand charges, TOU, monthly kWh, interval data), granularity and availability.

ANNEX III: TABLE A.11 NORWEGIAN ELHUB

SUCs / Requirements		Solution: 2.4. Norwegian Elhub	
ID	Name	Rating	Comment
SUC: Aggregate energy data			
AGG-ED-REQ-1	○ Standard rules to aggregate data in order not to enable the identification of persons behind data	3	Reports, statistics. Calculation and aggregation services.
AGG-ED-REQ-2	○ Standard rules to aggregate data in order to ensure the comparability of aggregated data sets	3	See AGG-ED-REQ-1
AGG-ED-REQ-3	○ Data source (e.g. meter data hub) ability to aggregate data	3	See AGG-ED-REQ-1
AGG-ED-REQ-4	○ DEP ability to forward aggregated data from a data source to data user	3	See AGG-ED-REQ-1
SUC: Anonymize energy data			
ANO-ED-REQ-1	○ Standard rules to anonymize data not to enable the identification of persons behind data	0	Elhub does not have the function of sharing anonymized data.
ANO-ED-REQ-2	○ Standard rules to anonymize data in order to ensure the comparability of anonymized data sets	0	See ANO-ED-REQ-1
ANO-ED-REQ-3	○ Data source (e.g. meter data hub) ability to anonymize data	0	See ANO-ED-REQ-1
ANO-ED-REQ-4	○ DEP ability to forward anonymized data from data source to data user	0	See ANO-ED-REQ-1
SUC: Authenticate data users			
AUTH-REQ-1	○ Right to access own data	3	The end user can access information related to their meter points, contracts and meter values via Elhub Web Plugin.
AUTH-REQ-2	○ Authentication tools	3	National ID (MinID, BankID, Buypass, Commfides)
AUTH-REQ-3	○ Ability to share information related to representation rights between data users and concerned Customer Portals	3	via Elhub Web Plugin
AUTH-REQ-4	○ Ability to share authentication information between data users, Customer Portal and Authentication Service Provider	3	See AUTH-REQ-2

SUC: Manage access permissions			
AUTHZN-REQ1	○ Every person needs access permission	3	Access permission can be granted via Elhub Web Plugin.
AUTHZN-REQ2	○ Valid identity of the person receiving access permissions	3	Every user needs to be registered in Elhub before receiving access permissions.
AUTHZN-REQ3	○ Ability to share access permissions between data owners, concerned DEPs, applications and data sources	3	All data and access permissions reside in Elhub, each party needs permission granted by data owner to access data.
SUC: Collect energy data			
DC-REQ1	○ Collection of meter data	-	
DC-REQ1.1	▪ Get near-real-time data (up to 1 hour) from meters	0	Meter data is delivered by DSO once a day for the previous day. After processing, the received data is made available to customers with a delay of up to 1 ½ day.
DC-REQ1.2	▪ Get historical data (monthly) from conventional meters	3	See DC-REQ1.1
DC-REQ1.3	▪ Store data in meter data hub	3	All meter data is stored in Elhub
DC-REQ2	○ Collection of market data	-	Elhub does not store or process market data.
DC-REQ2.1	▪ Get near-real-time (up to 1 hour) data from market	0	See DC-REQ2
DC-REQ2.2	▪ Get historical data from market	0	See DC-REQ2
DC-REQ2.3	▪ Store data in market data hub	0	See DC-REQ2
DC-REQ3	○ Collection of grid data	-	Elhub does not store or process grid data.
DC-REQ3.1	▪ Get very-near-real-time (up to 1 minute) data from grid	0	See DC-REQ3
DC-REQ3.2	▪ Get near-real-time (up to 1 hour) data from grid	0	See DC-REQ3
DC-REQ3.3	▪ Get historical data from grid	0	See DC-REQ3
DC-REQ3.4	▪ Store data in grid data hub	0	See DC-REQ3
SUC: Transfer energy data			
DT-REQ1	○ Transfer of data must be secured, by means of encryption or communication protocol	3	The end-user can provide direct access to third parties so that they can retrieve measurement values as long as the end-user is associated with the measurement point.

			The end user can then choose to provide full or limited access, where full access also provides access to contract data and basic data at the meter point, in addition to meter data. In contrast, limited access provides a minimum set of information and access to meter values.
DT-REQ2	<ul style="list-style-type: none"> Data portability (applies to personal data - Article 20 of the GDPR) 	3	Data owner can download data.
DT-REQ3	<ul style="list-style-type: none"> Data owner's access to data through DEP (and foreign DEP) 	2	Currently only domestic data exchange (request-response)
DT-REQ4	<ul style="list-style-type: none"> Application's access to data through DEP (and foreign DEP) 	2	Currently only domestic data exchange (request-response)
SUC: Exchange data between DER and SCADA			
DER-SCADA-REQ4	<ul style="list-style-type: none"> Ability of DEP to forward real-time data from DER's to System Operators 	0	Elhub and all of its components are designed for non-real-time operations. Elhub is not designed for communication with SCADA systems.
DER-SCADA-REQ5	<ul style="list-style-type: none"> Ability of DEP to forward very-near-real-time (up to 1 minute) data from DER's to System Operators 	0	See DER-SCADA-REQ4
DER-SCADA-REQ6	<ul style="list-style-type: none"> Ability of DEP to forward near-real-time (up to 1 hour) data from DER's to System Operators 	0	See DER-SCADA-REQ4
DER-SCADA-REQ7	<ul style="list-style-type: none"> Ability of DEP to forward activation requests from System Operators to DER 	0	See DER-SCADA-REQ4
DER-SCADA-REQ2	<ul style="list-style-type: none"> Communication link between DEP and SO's SCADA 	0	See DER-SCADA-REQ4
DER-SCADA-REQ1	<ul style="list-style-type: none"> Encrypted data exchange 	0	See DER-SCADA-REQ4
DER-SCADA-REQ3	<ul style="list-style-type: none"> Safety of DER's IT infrastructure 	N/A	DER's infrastructure is out of scope of analysis.
SUC: Manage flexibility activations			
FA-REQ2	<ul style="list-style-type: none"> Exchange of activation requests through DEP and flexibility platform 	0	Elhub does not transfer data between parties.

FA-REQ1	<ul style="list-style-type: none"> Automated activation of devices is possible 	0	Elhub does not support interoperability with devices
SUC: Calculate flexibility baseline			
FB-REQ1	<ul style="list-style-type: none"> Ability of flexibility platform to collect input for baseline calculation, incl. through DEP 	0	Elhub does not process data required for baseline calculation.
FB-REQ2	<ul style="list-style-type: none"> Ability of flexibility platform to compute baseline 	0	Elhub does not support baseline computing.
SUC: Manage flexibility bids			
FBIDS-REQ2	<ul style="list-style-type: none"> Ability to exchange information on System Operators' flexibility need and FSPs' flexibility potential through flexibility platform (and DEP) 	0	Elhub does not support bidding process at all.
FBIDS-REQ4	<ul style="list-style-type: none"> Algorithm for prequalification of flexibility providers 	0	See FBIDS-REQ2
FBIDS-REQ6	<ul style="list-style-type: none"> Flexibility platform's ability to collect bids from FSPs 	0	See FBIDS-REQ2
FBIDS-REQ7	<ul style="list-style-type: none"> Selection of successful bids 	0	See FBIDS-REQ2
FBIDS-REQ8	<ul style="list-style-type: none"> Flexibility platform's ability to collect grid validation results from SOs 	0	See FBIDS-REQ2
FBIDS-REQ9	<ul style="list-style-type: none"> Calculation of grid impacts (congestion, imbalance) 	0	See FBIDS-REQ2
FBIDS-REQ3	<ul style="list-style-type: none"> Auction process supervised by Market Operator 	0	See FBIDS-REQ2
FBIDS-REQ5	<ul style="list-style-type: none"> Automated exchange of bids is possible 	0	See FBIDS-REQ2
FBIDS-REQ1	<ul style="list-style-type: none"> Secure communication 	0	See FBIDS-REQ2
SUC: Predict flexibility availability			
FPRED-REQ1	<ul style="list-style-type: none"> Collection of data for prediction (long term - years) 	0	Elhub does not support flexibility prediction process at all.
FPRED-REQ2	<ul style="list-style-type: none"> Computation of predictions (long term - years) 	0	See FPRED-REQ1
FPRED-REQ1	<ul style="list-style-type: none"> Collection of data for prediction (medium-term - days to years ahead) 	0	See FPRED-REQ1
FPRED-REQ2	<ul style="list-style-type: none"> Computation of predictions (medium-term - days to years ahead) 	0	See FPRED-REQ1
FPRED-	<ul style="list-style-type: none"> Collection of data for 	0	See FPRED-REQ1

REQ1	prediction (short term - intraday operation)		
FPRED-REQ2	<ul style="list-style-type: none"> Computation of predictions (long term - intraday operation) 	0	See FPRED-REQ1
SUC: Verify and settle activated flexibilities			
FVERIF-REQ1	<ul style="list-style-type: none"> Calculation of actually delivered flexibility as a response to an activation request 	0	Elhub does not support flexibility verification process at all.
FVERIF-REQ2	<ul style="list-style-type: none"> Verification that flexibility delivered matches with flexibility requested 	0	See FVERIF-REQ1
FVERIF-REQ3	<ul style="list-style-type: none"> Calculation of the penalty if flexibility delivered is less than flexibility requested 	0	See FVERIF-REQ1
SUC: Provide list of suppliers and ESCOs			
ESCO-REQ1	<ul style="list-style-type: none"> List of suppliers and ESCOs is available through DEP; List of aggregators is available through flexibility platform additionally 	3	
SUC: Erase and rectify personal data			
PERSO-DATA-REQ1	<ul style="list-style-type: none"> Ability to share information related to erasure of personal data between data owners, concerned DEPs, applications and data sources 	3	
PERSO-DATA-REQ2	<ul style="list-style-type: none"> Ability to share information related to rectification of personal data between data owners, concerned DEPs, applications and data sources 	3	
SUC: Manage data logs			
LOGS-REQ1	<ul style="list-style-type: none"> Ability to share information related to data logs between data owners, concerned DEPs, applications and data sources 	3	
SUC: Manage sub-meter data			
SUBMET-REQ1	<ul style="list-style-type: none"> Collection of data from sub-meters 	0	Elhub does not collect or store sub-meter data.
SUBMET-REQ3	<ul style="list-style-type: none"> Storing sub-meter data in data hub 	0	See SUBMET-REQ1
SUBMET-	<ul style="list-style-type: none"> Ability of DEP to forward sub-meter data from data hub to customer (data 	0	Elhub can share stored information, it does not transfer data between parties.

REQ2	owner) and application (energy service provider)		
SUBMET-REQ7	<ul style="list-style-type: none"> Ability of DEP to forward activation orders from a customer (data owner) or application (energy service provider) to devices 	0	Elhub does not communicate with devices.
SUBMET-REQ4	<ul style="list-style-type: none"> Data format of sub-metering 	0	See SUBMET-REQ7
SUBMET-REQ5	<ul style="list-style-type: none"> Transmission protocols of sub-metering 	0	Elhub does not communicate with devices.
SUBMET-REQ6	<ul style="list-style-type: none"> SLA between customer and energy service provider 	0	

ANNEX IV – DATA EXCHANGE BETWEEN DSO AND TSO: SUPPLEMENTARY INFORMATION ABOUT EU REGULATIONS

ANNEX IV: TABLE A.12 GUIDELINE ON SYSTEM OPERATION

GUIDELINE ON SYSTEM OPERATION	TSO-DSO data exchange	Application for flexibility usage
<p><i>Article 48 – Structural data exchange (between TSOs, DSOs and distribution-connected power generating modules)</i></p> <ul style="list-style-type: none"> Paragraph 1: Power generating facility owner of a power generating module which is an SGU and by aggregation of the SGUs connected to the distribution system shall provide at least the following data to the TSO and to the DSO to which it has a connection point: general data of the power generating module, including installed capacity and primary energy source or fuel type; FCR, FRR and RR data; protection data; reactive power control capability; capability of remote access to the circuit breaker; data necessary for performing dynamic simulation; voltage level and location of each power generating module. 	Explicit	Structural data may be required for flexibility optimization, prediction and prequalification
<p><i>Article 53 – Data exchange between TSOs and distribution-connected demand facilities or third parties participating in demand response</i></p> <ul style="list-style-type: none"> Paragraph 1: SGU which is a distribution-connected demand facility and which participates in demand response other than through a third party shall provide the following scheduled and real-time data to the TSO and to the DSO: structural minimum and maximum active power available for demand response and the maximum and minimum duration of any potential usage of this power for demand response; a forecast of unrestricted active power available for demand response and any planned demand response; real-time active and reactive power at the connection point; a confirmation that the estimations of the actual values of demand response are applied. Paragraph 2: SGU which is a third party participating in demand response shall provide the TSO and the DSO at the day-ahead and close to real-time and on behalf of all of its distribution-connected demand facilities, with the following data: structural minimum and maximum active power available for demand response and the maximum and minimum duration of any potential activation of demand response in a specific geographical area defined by the TSO and DSO; a forecast of unrestricted active power available for the demand response and any planned level of demand response in a specific geographical area defined by the TSO and DSO; real-time active and reactive power; a confirmation that the estimations of the actual values of demand response are applied. 	Explicit	Structural data may be required for flexibility optimization, prediction and prequalification. Real-time data may be required for flexibility activation and optimization. Scheduled data may be required for flexibility optimization, prediction, prequalification, baseline calculation
<p><i>Article 154 – FCR technical minimum requirements</i></p>	Explicit	May concern DSO-connected

<ul style="list-style-type: none"> - Paragraph 8: Each FCR provider shall make available to the reserve connecting TSO³⁶, for each of its FCR providing units and FCR providing groups, at least the following information: (a) time-stamped status indicating if FCR is on or off; (b) time-stamped active power data needed to verify FCR activation, including time-stamped instantaneous active power; (c) droop of the governor for type C and type D power generating modules acting as FCR providing units, or its equivalent parameter for FCR providing groups consisting of type A or type B power generating modules, or demand units with demand response active power control³⁷. - Paragraphs 9-11: Each FCR provider shall have the right to aggregate the respective data for more than one FCR providing unit if the maximum power of the aggregated units is below 1,5 MW and transparent verification of activation of FCR is possible. At the request of the reserve connecting TSO, the FCR provider shall make the information listed in paragraph 9 available in real-time, with a time resolution of at least 10 seconds. At the request of the reserve connecting TSO and where necessary for the verification of the activation of FCR, an FCR provider shall make available the data concerning technical installations that are part of the same FCR providing unit. 		flexibilities. Technical requirements are needed for flexibility activation and flexibility baseline calculation
<i>Article 155 – FCR prequalification process</i> <ul style="list-style-type: none"> - Paragraph 3: A potential FCR provider shall submit a formal application to the reserve connecting TSO together with the required information of potential FCR providing units or FCR providing groups. 	Explicit	May involve prequalification of DSO-connected flexibilities
<i>Article – 156 FCR provision</i> <ul style="list-style-type: none"> - Paragraph 5: Each FCR provider shall inform its reserve connecting TSO, as soon as possible, about any changes in the actual availability of its FCR providing unit or its FCR providing group, in whole or in part, relevant for the results of prequalification. 	Explicit	May involve the provision of DSO-connected flexibilities
<i>Article 158 – FRR minimum technical requirements</i> <ul style="list-style-type: none"> - Paragraph 1(b): A FRR providing unit or FRR providing group shall activate FRR in accordance with the setpoint received from the reserve instructing TSO. - Paragraph 1(e): A FRR provider shall ensure that the FRR activation of the FRR providing units within a reserve providing group can be monitored. For that purpose, the FRR provider shall be capable of 	Explicit	May concern DSO-connected flexibilities. Technical requirements are needed for flexibility activation and

³⁶ According to SOGL ‘reserve connecting TSO’ means the TSO responsible for the monitoring area to which a reserve providing unit or reserve providing group is connected. Thus flexibility itself can be physically connected to DSO grid as well.

³⁷ According to article 5 of RfG NC the type A power-generating modules’ connection point is below 110 kV and maximum capacity is 0,8 kW or more, the type B and C power-generating modules’ connection point is below 110 kV and maximum capacity at or above a threshold proposed by each relevant TSO and approved by the relevant regulatory authority or Member State, and type D power-generating modules’ connection point is at 110 kV or above. A power-generating module is also of type D if its connection point is below 110 kV and its maximum capacity is at or above a threshold proposed by each relevant TSO and approved by the relevant regulatory authority or Member State.

<p>supplying to the reserve connecting TSO and the reserve instructing TSO real-time measurements of the connection point or another point of interaction agreed with the reserve connecting TSO concerning: (i) time-stamped scheduled active power output; (ii) time-stamped instantaneous active power for each FRR providing unit, for each FRR providing group, and for each power generating module or demand unit of an FRR providing group with a maximum active power output larger than or equal to 1,5 MW.</p> <ul style="list-style-type: none"> - Paragraph 4(b): Each FRR provider shall inform its reserve instructing TSO about a reduction of the actual availability of its FRR providing unit or its FRR providing group or a part of its FRR providing group as soon as possible. 		flexibility baseline calculation
<p><i>Article 159 – FRR prequalification process</i></p> <ul style="list-style-type: none"> - Paragraph 3: A potential FRR provider shall submit a formal application to the relevant reserve connecting TSO or the designated TSO together with the required information of potential FRR providing units or FRR providing groups. 	Explicit	May involve prequalification of DSO-connected flexibilities
<p><i>Article 161 – RR minimum technical requirements</i></p> <ul style="list-style-type: none"> - Paragraph 1(f): A RR provider shall ensure that the RR activation of the RR providing units within a reserve providing group can be monitored. For that purpose, the RR provider shall be capable of supplying to the reserve connecting TSO and the reserve instructing TSO real-time measurements of the connection point or another point of interaction agreed with the reserve connecting TSO concerning: (i) the time-stamped scheduled active power output, for each RR providing unit and group and for each power generating module or demand unit of a RR providing group with a maximum active power output larger than or equal to 1,5 MW; (ii) the time-stamped instantaneous active power, for each RR providing unit and group, and for each power generating module or demand unit of a RR providing group with a maximum active power output larger than or equal to 1,5 MW. 	Explicit	May concern DSO-connected flexibilities. Technical requirements are needed for flexibility activation and flexibility baseline calculation
<p><i>Article 162 – RR prequalification process</i></p> <ul style="list-style-type: none"> - Paragraph 3: A potential RR provider shall submit a formal application to the relevant reserve connecting TSO or the designated TSO together with the required information of potential RR providing units or RR providing groups. 	Explicit	May involve prequalification of DSO-connected flexibilities
<p><i>Article 182 – Reserve providing groups or units connected to the DSO grid</i></p> <ul style="list-style-type: none"> - Paragraph 1: TSOs and DSOs shall cooperate in order to facilitate and enable the delivery of active power reserves by reserve providing groups or reserve providing units located in the distribution systems. - Paragraph 2: For the purposes of the prequalification processes for FCR, FRR and RR, each TSO shall develop and specify, in an agreement 	Explicit	May involve DSO-connected flexibilities. Technical requirements are needed for flexibility prequalification

<p>with its reserve connecting DSOs and intermediate DSOs, the terms of the exchange of information required for these prequalification processes for reserve providing units or groups located in the distribution systems and for the delivery of active power reserves.</p> <ul style="list-style-type: none"> - Paragraph 2(a)-(d): The prequalification processes for FCR in Article 155, FRR in Article 159 and RR in Article 162 shall specify the information to be provided by the potential reserve providing units or groups, which shall include: voltage levels and connection points of the reserve providing units or groups; the type of active power reserves; the maximum reserve capacity provided by the reserve providing units or groups at each connection point; the maximum rate of change of active power for the reserve providing units or groups. - Paragraph 4: During the prequalification of a reserve providing unit or group connected to its distribution system, each reserve connecting DSO and each intermediate DSO, in cooperation with the TSO, shall have the right to set limits to or exclude the delivery of active power reserves located in its distribution system, based on technical reasons such as the geographical location of the reserve providing units and reserve providing groups. - Paragraph 5: Each reserve connecting DSO and each intermediate DSO shall have the right, in cooperation with the TSO, to set, before the activation of reserves, temporary limits to the delivery of active power reserves located in its distribution system. The respective TSOs shall agree with their reserve connecting DSOs and intermediate DSOs on the applicable procedures. 		and activation
--	--	----------------

ANNEX IV: TABLE A.13 TSO PROPOSALS RELATED TO DATA EXCHANGE

All TSOs' proposal for the Key Organisational Requirements, Roles and Responsibilities (KORRR) relating to Data Exchange	TSO-DSO data exchange	Application for flexibility usage
<p><i>Article 3 – General responsibilities</i></p> <ul style="list-style-type: none"> - Paragraph 2: On the basis of Articles 48 to 50 and 53 of the SO GL, the KORRR renders the provision of data both to TSOs and DSOs as the default option. This approach can be revised at a national level in order to allow SGUs the provision of data only to the TSO or to the DSO to which they are connected unless otherwise required to provide services to the system. In those cases where an SGU only provides data to a TSO or to a DSO to which they are connected, the TSO and the DSO shall exchange between them the data related to that SGU. - Paragraph 3: Subject to approval by the competent regulatory authority or by the entity designated by the Member State and according to Article 40 of the SO GL, it shall be determined at a national level whether distribution connected SGUs in their TSOs control area shall provide the structural, scheduled and real-time data to the TSO directly or through their connecting DSOs or to both. The decision on the data exchange model may be independent for each type of information and SGU if required. When the data is provided to the DSO, the DSO shall provide the required data to the TSO with a data granularity necessary to comply with the requirements of the SO GL provisions. 	Explicit. TSO-DSO agreement required	Involves data relevant for flexibility processes

ANNEX IV: TABLE A.14 GUIDELINE ON ELECTRICITY BALANCING

GUIDELINE ON ELECTRICITY BALANCING	TSO-DSO data exchange	Application for flexibility usage
<p><i>Article 15 – Cooperation with DSOs</i></p> <ul style="list-style-type: none"> - Paragraph 1: DSOs, TSOs, balancing service providers and responsible balance parties shall cooperate in order to ensure efficient and effective balancing - Paragraph 2: Each DSO provides all necessary information to perform the imbalance settlement to the connecting TSO 	Explicit	Information relevant for flexibility baseline calculation and flexibility verification
<p><i>Article 16 – Role of balancing service providers</i></p> <ul style="list-style-type: none"> - Paragraph 1: Successful completion of the prequalification, ensured by the connecting TSO as a prerequisite for the successful completion of the qualification process to become a balancing service provider - Paragraphs 2-5: Each balancing service provider shall submit to the connecting TSO information related to its balancing bids 	Explicit. Includes flexibilities connected to DSO	Information relevant for flexibility prequalification and bidding
<p><i>Article 18 – Terms and conditions related to balancing</i></p> <ul style="list-style-type: none"> - Paragraph 4(b),(c): Allow the participation of the demand/ aggregated demand, aggregated distributed energy sources, storage to balancing (be balancing service provider) - Paragraph 5(d),(f),(g): Requirement on data for DSO-connected reserves should be defined in terms and condition of BSP (Terms and conditions of balancing service provider should contain the requirements on data and information to be delivered to the connecting TSO and where relevant the reserve connecting DSO during prequalification and operation; the requirements on data and information to be delivered to the connecting TSO and where relevant the reserve connecting DSO to evaluate the provision of balancing services and to calculate imbalance; the definition of a location for each product) - Paragraph 6(d): Requirement on data and information to be delivered to the connecting TSO to calculate the imbalance is defined in terms and condition of BRP 	Explicit	Information relevant for flexibility prequalification, bidding, activation, baseline calculation and verification

ANNEX IV: TABLE A.15 NETWORK CODE ON DEMAND CONNECTION

NETWORK CODE ON DEMAND CONNECTION	TSO-DSO data exchange	Application for flexibility usage
<p><i>Article 28 – Specific provisions for demand units with demand response active power control, reactive power control and transmission constraint management</i></p> <ul style="list-style-type: none"> - Paragraph 2(e): Demand units with demand response active power control, demand response reactive power control, or demand response transmission constraint management, either individually or, where it is not part of a transmission-connected demand facility, collectively as part of demand aggregation through a third party shall 	Explicit. Involves DSO-connected demand units	Information relevant for flexibility activation and verification

be equipped to receive instructions, directly or indirectly through a third party, from the relevant system operator or the relevant TSO to modify their demand and to transfer the necessary information. The relevant system operator shall make publicly available the technical specifications approved to enable this transfer of information.		
---	--	--

ANNEX IV: TABLE A.16 NETWORKCODE ON REQUIREMENTS FOR GRID CONNECTION OF GENERATORS

NETWORK CODE ON REQUIREMENTS FOR GRID CONNECTION OF GENERATORS	TSO-DSO data exchange	Application for flexibility usage
Article 14 – General requirements for type B power-generating modules <ul style="list-style-type: none"> - Paragraph 5(d)(i): Power-generating facilities shall be capable of exchanging information with the relevant system operator or the relevant TSO in real-time or periodically with time stamping, as specified by the relevant system operator or the relevant TSO. - Paragraph 5(d)(ii): The relevant system operator, in coordination with the relevant TSO, shall specify the content of information exchanges including a precise list of data to be provided by the power-generating facility. 	Where applicable, TSO and DSO agreement for data exchange required	Data from generators may be useful for different flexibility processes
Article 15 – General requirements for type C power-generating modules <ul style="list-style-type: none"> - Paragraph 6(b)(iv): The facilities for quality of supply and dynamic system behaviour monitoring shall include arrangements for the power-generating facility owner, and the relevant system operator and the relevant TSO to access the information. The communications protocols for recorded data shall be agreed between the power-generating facility owner, the relevant system operator and the relevant TSO. 	Where applicable, TSO and DSO agreement for data exchange required	Data from generators may be useful for different flexibility processes

ANNEX IV: TABLE A.17 DIRECTIVE ON COMMON RULES FOR THE INTERNAL MARKET IN ELECTRICITY

DIRECTIVE ON COMMON RULES FOR THE INTERNAL MARKET IN ELECTRICITY	TSO-DSO data exchange	Application for flexibility usage
Article 17 – Demand response through aggregation <ul style="list-style-type: none"> - Paragraph 3(c): Non-discriminatory and transparent rules and procedures for the exchange of data between market participants engaged in aggregation and other electricity undertakings that ensure easy access to data on equal and non-discriminatory terms while fully protecting commercially sensitive information and customers' personal data. 	TSO and DSO can ensure easy access to data, where they have the role of distributing meter data to third parties (see Metering Data Operator or Smart Meter Gateway Operator or Data Hub Operator or Data Exchange Platform Operator instead ³⁸)	Demand response data relevant for different flexibility processes

³⁸ Data Hub Operator and Data Exchange Platform Operator are roles proposed by EU-SysFlex WP5.

<p><i>Article 20 – Functionalities of smart metering systems</i></p> <ul style="list-style-type: none"> - Paragraph (a): Validated historical consumption data shall be made quickly and securely available and visualised to final customers on request and at no additional cost. Non-validated near real-time consumption data shall also be made easily and securely available to final customers at no additional cost, through a standardised interface or through remote access, in order to support automated energy efficiency programmes, demand response and other services. - Paragraph (e): If final customers request it, data on the electricity they fed into the grid and their electricity consumption data shall be made available to them, in accordance with the implementing acts adopted pursuant to Article 24, through a standardised communication interface or through remote access, or to a third party acting on their behalf, in an easily understandable format allowing them to compare offers on a like-for-like basis. - It shall be possible for final customers to retrieve their metering data or transmit them to another party at no additional cost and in accordance with their right to data portability under Union data protection rules. 	<p>Management of smart meter data may involve both TSO and DSO, where they have the role of distributing meter data to third parties (see Metering Data Operator or Smart Meter Gateway Operator or Data Hub Operator or Data Exchange Platform Operator instead)</p>	<p>Smart meter data relevant for different flexibility processes</p>
<p><i>Article 23 – Data management</i></p> <ul style="list-style-type: none"> - Paragraph 1: Authorities shall specify the rules on the access to data of the final customer by eligible parties in accordance with applicable Union legal framework. Data shall be understood to include metering and consumption data as well as data required for customer switching, demand response and other services. - Paragraph 2: Member States shall organise the management of data in order to ensure efficient and secure data access and exchange, as well as data protection and data security. Independently of the data management model applied in each Member State, the parties responsible for data management shall provide access to the data of the final customer to any eligible party. Eligible parties shall have the requested data at their disposal in a non-discriminatory manner and simultaneously. Access to data shall be easy, and the relevant procedures for obtaining access to data shall be made publicly available. 	<p>Management of final customer data may involve both TSO and DSO, where they have the role of distributing meter data to third parties (see Metering Data Operator or Smart Meter Gateway Operator or Data Hub Operator or Data Exchange Platform Operator instead)</p>	<p>Final customer data relevant for different flexibility processes</p>
<p><i>Article 24 – Interoperability requirements and procedures for access to data</i></p> <ul style="list-style-type: none"> - Paragraph 1: In order to promote competition in the retail market and to avoid high administrative costs for the eligible parties, Member States shall facilitate the full interoperability of energy services within the Union. - Paragraph 2: The Commission shall adopt, using implementing acts, interoperability requirements and non-discriminatory and transparent procedures for access to metering data. - Paragraph 3: Member States shall ensure that electricity undertakings apply the interoperability requirements and procedures for access to metering data. Those requirements and procedures shall be based on existing national practices. 	<p>Interoperability is relevant for TSO-DSO data exchanges where they have the role of distributing meter data to third parties (see Metering Data Operator or Smart Meter Gateway Operator or Data Hub Operator or Data Exchange Platform Operator instead)</p>	<p>Interoperability is relevant in all flexibility processes, incl. in the retail market</p>

	Platform Operator instead)	
--	----------------------------	--

ANNEX IV: TABLE A.18 REGULATION ON THE INTERNAL MARKET FOR ELECTRICITY

REGULATION ON THE INTERNAL MARKET FOR ELECTRICITY	TSO-DSO data exchange	Application for flexibility usage
<p><i>Article 57 – Cooperation between distribution system operators and transmission system operators</i></p> <ul style="list-style-type: none"> - Paragraph 1: DSOs and TSOs shall cooperate in planning and operating their networks. In particular, distribution system operators and transmission system operators shall exchange all necessary information and data regarding the performance of generation assets and demand-side response, the daily operation of their networks and the long-term planning of network investments, with the view to ensure the cost-efficient, secure and reliable development and operation of their networks. 	Explicit. TSO and DSO agreement for data exchange is useful	Information relevant for different flexibility processes

ANNEX V - PRIVACY-PRESERVING DATA ANALYSIS: PROOF OF CONCEPT

BACKGROUND

This deliverable has been produced as a part of the EU-SysFlex project work package 5 (WP5): "Data management for the facilitation of new flexibility solutions". Task 5.3 focuses on data storage and big data solutions and should deliver a report describing data collection, storage and processing requirements.

In general, the approach for this privacy-preserving data analysis, is two-fold. A PoC has been developed that could:

- provide useful maintenance, development such as insight on privacy-preserving development in the electricity sector;
- be turned into a useful privacy-preserving demonstrator as a part of the larger project.

hBaseline calculations are used to estimate energy consumption from historical meter readings. Data providers participating in the Flexibility Platform may not be willing to calculate baselines. To allow baseline calculation to be outsourced to a third party, we must ensure that the computation party does not learn the consumer's consumption history.

Sharemind MPC is a technology which enables processing data without leaking individual values. Using Sharemind MPC, a proof of concept will be developed that calculates a baseline for a customer or a set of customers without revealing their actual meter readings or the baseline(s) to the computing party. One can learn more about Sharemind MPC in the *Sharemind Privacy Ecosystem* (2020) document.

SHAREMIND MPC

A Sharemind MPC deployment consists of three Sharemind MPC servers which must be hosted by different entities. Distributed control ensures privacy since values in Sharemind MPC are encrypted so that no server host can see the original values. Distributed control also ensures that only agreed upon computations can be run on the encrypted values. A single host cannot run arbitrary computations in the distributed Sharemind MPC deployment on their own.

In Sharemind MPC, personal values are secret-shared before being imported to the system. That is, given a 32-bit private integer x , two random values x_1, x_2 are generated and x_3 is calculated such that $x \equiv x_1 + x_2 + x_3 \mod 2^{32}$. Each of the three servers receives one share of x . Since a share is random, it provides no information about the original value x . Using network protocols, the set of three servers can perform arithmetic on secret-shared values. The results of computations are also secret-shared.

The three servers can collectively publish results of computations. Each server sends their share of a published result to an output party who can combine the shares with learning the result. Note that distributed control means that all three independent server hosts must agree which results are published. A single server cannot learn intermediate results, nor can they publish a result without the cooperation of the other servers.

HIGH-FIVE-OF-TEN ALGORITHM

Since baseline calculations are often used in contracts, their formulas must be understandable and straightforward in order to earn the trust of both parties of the contract. In this proof of concept, the symmetric high-five-of-ten formula shall be used (Woolf, Ustinova, Ortega, O'Brien, Djapic & Strbac, 2014). It has to be assumed that data is metered hourly. The process for calculating the baseline energy consumption for a timestamp is as follows:

1. Find the preceding ten days that do not include events like unusual energy consumption due to extreme weather.
2. Order the ten days according to their total energy consumption and retain the top 5 days.
3. Calculate the average energy consumption of the corresponding hour in the top 5 days.
4. If the energy consumption in the 2 hours preceding the timestamp to be estimated above or below its baseline, shift the calculated baseline upwards or downwards.

Some formulas do not include step 4, but it is often used to shift the calculated baseline to account for unusual days. If the timestamp to be estimated is in a period of unusual consumption (for example, due to extreme weather or equipment failure) the baseline can under- or overestimate the reliable baseline. The full formula is as follows:

$$b_t = \frac{c_1 + c_2 + c_3 + c_4 + c_5}{5} + \frac{c_{t-1} - b_{t-1} + c_{t-2} - b_{t-2}}{2}$$

b_t designates baseline at timestamp t . c_i is the consumption of the corresponding timestamp on a day i of the top 5 days.

There is a distinction between weekdays and weekend days. For example, to calculate the baseline of a Saturday, starts with ten weekend days. Likewise, with weekdays it starts with ten weekdays excluding weekend days.

ASSUMPTIONS OF THE PROOF-OF-CONCEPT DESIGN

POC software will be designed in a way so that it could be integrated with the Flexibility Platform, but it will not be integrated. The POC will be stand-alone.

The POC only aims to provide privacy-preserving analytics. Therefore much of the business process around baseline calculation is outside of scope.

The component of the POC implemented using the Sharemind MPC platform does not act as a data warehouse. Only data required for the baseline calculation will be imported into Sharemind MPC, and it is deleted after the calculation.

The POC will only consider consumption data. Energy production data shall not be given as input.

For the POC meter data aggregated was received by postcode. Consumers, in this case, will be identified by their postcode.

Historical metering data will not be used for estimation. No weather data or other sources are used.

Using real consumption data for baseline calculation

The high-five-of-ten model requires consumption data of the two timestamps preceding the timestamp that is being estimated. For instance, If metering data is acquired each hour, this model only allows us to calculate the baseline for the next hour ahead.

Let us assume for an example that the baseline for 11:00 on January 11th should be calculated. Consumption data up until January 10th will be used. It is desired to adjust the baseline to account for potentially unusual consumption on January 11th. Due to this, the high-five-of-ten formula has an adjustment term which uses consumption data of 10:00 and 09:00 on January 11th.

Consumption data preceding the baseline timestamp for adjustment is needed, it is assumed that there is full consumption data for the period being estimated.

This is useful for performance estimation. It is observed how many baseline computations can be computed in a limited amount of time. Another use case is to get a comparison of actual consumption and computed baseline over a period.

Differences between the proof-of-concept implementation and a production-ready component integrated with the Flexibility Platform

The PoC application should not be integrated with the Flexibility Platform, but it has been designed to enable integration. A possible scheme for integration is described in the chapter *“Integration with the Flexibility Platform”* of this PoC document.

For simplicity, we assume that the baseline calculation query parameters, such as consumer identifiers and the period being estimated, are public in the described integration protocol but private in the actual implemented POC application.

In order for the Estfeed platform to be able to enforce mandates, it must be able to check mandate object codes in message payloads. POC application, however, keeps all of the data private. It means that the POC application does not learn consumption data, consumers included in the baseline calculation or the estimation period. However, the described integration protocol keeps query parameters public for the sake of simplicity but protects consumption data.

The data used for evaluating the POC has been aggregated by postcode. In the POC, we will identify consumers by postcode. In a potential Flexibility Platform integration via Estfeed, metering points can be identified by the combination of a person’s identifier and metering point’s EIC code.

INPUT DATA FORMAT

This chapter describes the format of the tabular data imported to Sharemind MPC POC implementation and the parameters of the baseline calculation. Note that as described in section “*Assumptions of the proof-of-concept design*” potential integration will keep the query parameters public and only protect data. However, the POC implementation does not reveal query parameters to the Sharemind MPC software. The message payload formats of a potential integration with Estfeed and the Flexibility Platform are described in chapter “*Estfeed service message payload formats*”.

The Sharemind MPC POC program shall receive as inputs:

- The beginning and end of the period for which the baseline will be calculated.
- The beginning and end of the historical metering data period used for calculating the baseline.
- The list of consumer identifiers whose data will be used to calculate the baseline.

The metering data is expected in a tabular format with the following columns:

- date - date of metering.
- hour - an hour of metering as an unsigned 8-bit integer.
- consumer - consumer identifier as an unsigned 64-bit integer.
- consumed W - metered electricity consumption at that time point as an unsigned 64-bit integer.

The date should be packed in 32 bits. The first 16 bits indicate the year, the next 8 bits indicate the month and the last 8 bits indicate the day.

It is expected that the input data is clean. There should be no missing timestamps, and every value must be valid. We expect that the consumption data is metered at hourly intervals. Likewise, we will compute the baseline time-series in one-hour steps.

The endpoints of the historical period and estimated period should be given as ISO 8601 date and time. E.g. 09:00 on 11th of January 2017 is “2017-01-11T09:00:00Z”.

The consumer identifiers are given as a list of consumer identifiers (postcodes in this case) separated by commas (“,”) without spaces.

OUTPUT DATA FORMAT

The output of the Sharemind MPC POC program is a table with three columns. The Estfeed message payload format of a potential integration with the Flexibility Platform is described in chapter “*Estfeed service message payload formats*”. The columns are:

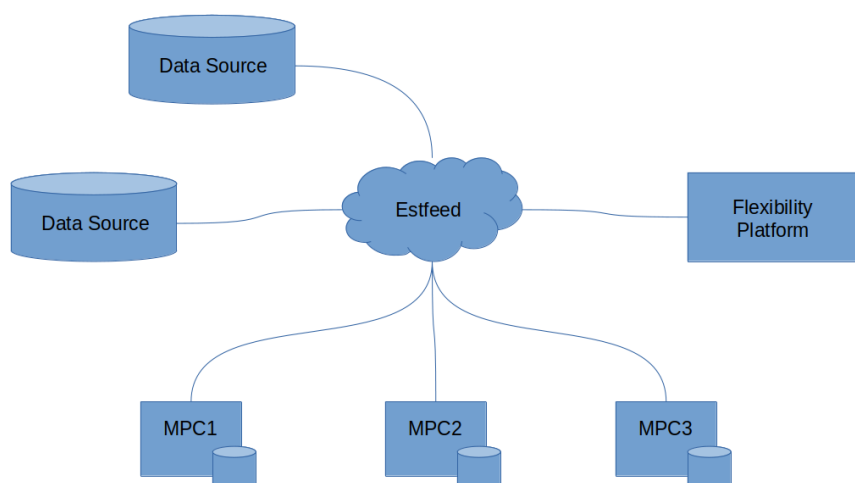
- date - date of a data point in the same format as used for the input.
- hour - hour of metering as an unsigned 8-bit integer.
- consumer - consumer identifier as an unsigned 64-bit integer.
- consumed W - computed baseline value for the timestamp in watt-hours as a 64-bit IEEE 754 floating-point number.

INTEGRATION WITH THE FLEXIBILITY PLATFORM

The application will not be integrated with the Flexibility Platform, however, designed POC enables integration in the future. In this chapter, the design will be described of the baseline calculation process in the Flexibility Platform.

One of the main features of the Estfeed exchange used for the Flexibility Platform is the ability for consumers to decide which applications are given access to their data. The scenario described in this section has four Estfeed applications which require a mandate: the Flexibility Platform and three Sharemind MPC servers. Consumers must give mandates to all four applications in order to participate in baseline calculations. A possible extension of the MyEstfeed portal used for giving and revoking mandates would be to support groups of multiple applications. This would allow the consumer to give mandates to all required Estfeed applications at once.

Components



ANNEX V: FIGURE A.11 DIAGRAM OF COMPONENTS PARTICIPATING IN THE BASELINE CALCULATION PROCESS

The Flexibility Platform, data providers and three Sharemind MPC servers should all be integrated into the Estfeed platform as services. We expect that there will be more than one data provider participating as input parties to the baseline calculation.

Privacy is ensured by the Sharemind MPC technology if the parties hosting the MPC servers are independent and do not collude to break the privacy of individual consumers. No Sharemind MPC server can individually break the privacy.

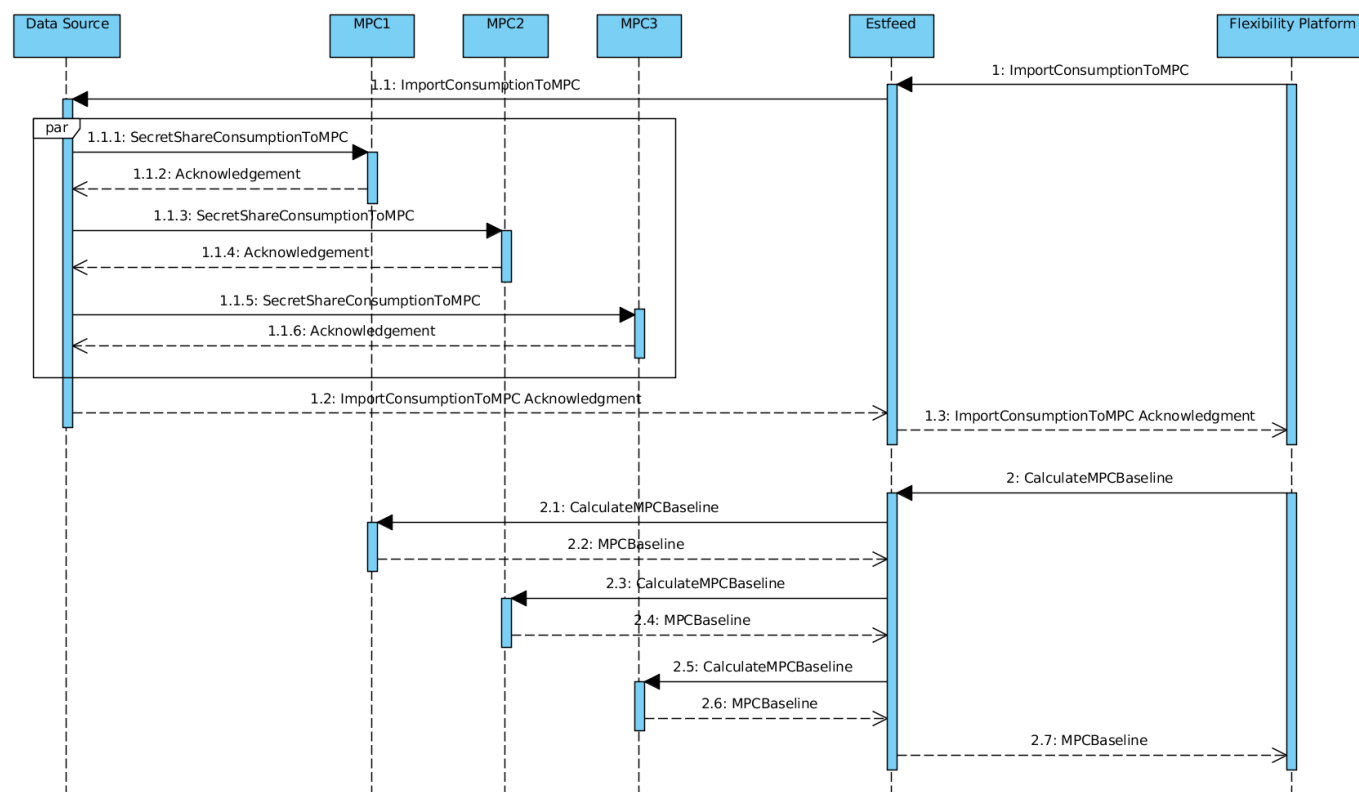
Before deployment, it must be decided which three organisations host the MPC servers. It would make sense for the MPC servers to be hosted by parties who are interested in the privacy-preserving baseline calculation such as aggregator businesses participating in the Flexibility Platform. One of the hosts could even be the party responsible for the Flexibility Platform or an organisation regulating the flexibility market.

Baseline calculation process

The process of calculating the baseline using Sharemind MPC can be split into two phases:

1. Secret-sharing and importing relevant data to Sharemind MPC.
2. Calculating the baseline.

The following diagram illustrates the baseline calculation process.



ANNEX V: FIGURE A.12 SEQUENCE DIAGRAM OF BASELINE CALCULATION PROCESS

We assume that there will be multiple data providers. This means that the Flexibility Platform will request data from multiple data providers and only start the baseline calculation if it has received acknowledgements of the import requests from all data providers.

Three new service types should be added to the Estfeed platform:

1. *ImportConsumptionToMPC*. Data providers implement this service. The Flexibility Platform issues an *ImportConsumptionToMPC* request to data providers to request them to secret-share and upload consumption data of the required period to Sharemind MPC servers.
2. *SecretShareConsumptionToMPC*. This service is implemented by Sharemind MPC server Estfeed integration component. A data provider issues a *SecretShareConsumptionToMPC* request when importing secret-shared data to a Sharemind MPC server.
3. *CalculateMPCBaseline*. This service is implemented by Sharemind MPC server Estfeed integration component. The Flexibility Platform issues a *CalculateMPCBaseline* request after data has been imported to MPC, and the baseline can be calculated.

Each value of the input data is secret-shared into three shares. An MPC server receives their share of each value.

While none of the shares can individually reveal the actual value, Estfeed sees all of the messages and could reconstruct the original data. To prevent this, the shares are encrypted using the public keys of the Sharemind MPC servers. Since Estfeed uses UXP for transmitting messages, we can use the public-key infrastructure of UXP. Encryption of the shares ensures that the Estfeed platform does not see the actual data while its ability to filter messages according to mandates given by consumers remains intact. The encrypted shares are encoded using base64 in order to transmit them in Estfeed message payloads as XML fields.

The result of the baseline calculation is also secret-shared. Each MPC server responds to the CalculateMPCBaseline with their shares of the calculated baseline. The shares are encrypted using the public key of the Flexibility Platform which ensures that the Estfeed platform can not see the calculated baseline.

Note that the message payloads include mandate object codes. It means that consumers need to opt-in to baseline calculations with Sharemind MPC. The goal is to earn the consumers trust by using privacy-preserving technology for baseline calculation but their choice whether to participate in the Flexibility Platform, is still respected.

ESTFEED SERVICE MESSAGE PAYLOAD FORMATS

Estfeed protocol, services, mandate objects codes and payloads are described in the Estfeed Protocol documentation. In this chapter, we will describe the payload formats of the messages used by the baseline calculation process. The payload formats are described in XSD (XML Schema Definition) markup.

ImportConsumptionToMPC

The Flexibility Platform issues one ImportConsumptionToMPC request with one or more payloads. Each payload identifies one electricity usage point. Each payload includes the mandate object code associated with the usage point.

Payload headers

Mandate object code: EIC code

Mandate object kind: *UsagePoint.Electricity*

Payload fields

Name	Type	Use	Description/Value
Person	String	Required	[ETSI] person ID
UsagePoint	String	Required	EIC code of usage point must match header
TimePeriod	DateTimeInterval	Required	ISO 8601 time interval of meter reading start timestamps, expressed by period start and end timestamp

Payload XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
```

```

    elementFormDefault="qualified "
    attributeFormDefault="unqualified">
<xs:element name="ImportConsumptionToMPC">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Person" type="xs:string"></xs:element>
      <xs:element name="UsagePoint" type="xs:string"></xs:element>
      <xs:element name="TimePeriod" type="xs:string"></xs:element>
    </xs:sequence>
    <xs:attribute name="xmlns:xsi" type="xs:string"></xs:attribute>
    <xs:attribute name="xsi:noNamespaceSchemaLocation"
      type="xs:string"></xs:attribute>
  </xs:complexType>
</xs:element>
</xs:schema>

```

SecretShareConsumptionToMPC

As described in Section 7, the consumption data is processed in the following way:

1. The data provider secret-shares each consumption value in the period. That is, for each value, three random values are generated such that their sum is the original value.
2. Shares destined for an MPC server are packed into a vector.
3. The vector is encrypted using the public key of the MPC server which hides the data from Estfeed.
4. The cryptogram is base64-encoded into a string to transmit it in the Estfeed message payload.

Payload headers

Mandate object code: EIC code

Mandate object kind: *UsagePoint.Electricity*

Payload fields

Name	Type	Use	Description/Value
Person	String	Required	[ETSI] person ID
UsagePoint	String	Required	EIC code of usage point must match header
TimePeriod	DateTimeInterval	Required	ISO 8601 time interval of meter reading start timestamps, expressed by period start and end timestamp.
Consumption	String	Required	a base64-encoded encrypted packed array of shares of consumption values

Payload XSD

```

<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified "
  attributeFormDefault="unqualified">
  <xs:element name="SecretShareConsumptionToMPC">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Person" type="xs:string"></xs:element>
        <xs:element name="UsagePoint" type="xs:string"></xs:element>

```



```

    <xs : element name="TimePeriod " type="xs : string"></xs : element>
    <xs : element name="Consumption" type="xs : string"></xs : element>
</xs : sequence>
<xs : attribute name="xmlns : xsi " type="xs : string"></xs : attribute >
<xs : attribute name=" xsi : noNamespaceSchemaLocation"
    type="xs : string"></xs : attribute >
</xs : complexType>
</xs : element>
</xs : schema>

```

CalculateMPCBaseline

Multiple payloads can be specified to calculate baselines for multiple consumers in parallel. The periods in all payloads must match.

Payload headers

Mandate object code: EIC code

Mandate object kind: *UsagePoint.Electricity*

Payload fields

Name	Type	Use	Description/Value
Person	String	Required	[ETSI] person ID
UsagePoint	String	Required	EIC code of usage point must match header
DataTimePeriod	DateTimeInterval	Required	ISO 8601 time interval of meter reading start timestamps, expressed by period start and end timestamp.
EstimationTimePeriod	DateTimeInterval	Required	ISO 8601 time interval of the period being estimated, expressed by period start and end timestamp

Payload XSD

```

<?xml version ="1.0" encoding="UTF-8"?>
<xs :schema xmlns : xs=" http : / /www.w3. org /2001/XMLSchema"
    elementFormDefault=" qualified "
    attributeFormDefault =" unqualified">
  <xs : element name="CalculateMPCBaseline">
    <xs : complexType>
      <xs : sequence>
        <xs : element name="Person" type="xs : string"></xs : element>
        <xs : element name="UsagePoint " type="xs : string"></xs : element>
        <xs : element name="DataTimePeriod " type="xs : string"></xs : element>
        <xs : element name=" EstimationTimePeriod " type="xs : string"></xs : element>
      </xs : sequence>
      <xs : attribute name="xmlns : xsi " type="xs : string"></xs : attribute >
      <xs : attribute name=" xsi : noNamespaceSchemaLocation"
        type="xs : string"></xs : attribute >
    </xs : complexType>
  </xs : element>
</xs : schema>

```

MPCBaseline

As described in Section 7, the baseline values are processed in the following way:

1. After calculating the baselines, each MPC server holds one share of each actual baseline value. If the Flexibility platform receives the shares, it can construct the public baseline values.
2. Each MPC server packs its shares of each consumer's baseline into a vector.
3. The vectors are encrypted using the public key of the Flexibility Platform which hides the data from Estfeed while transmitting it.
4. The cryptograms are base64-encoded into strings to transmit them in Estfeed message payloads. There will be one payload per consumer.

Payload headers

Mandate object code: EIC code

Mandate object kind: *UsagePoint.Electricity*

Payload fields

Name	Type	Use	Description/Value
Person	String	Required	[ETSI] person ID
UsagePoint	String	Required	EIC code of usage point must match header
Baseline	String	Required	a base64-encoded encrypted packed array of shares of calculated baseline

Payload XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified"
  attributeFormDefault="unqualified">
  <xs:element name="MPCBaseline">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Person" type="xs:string"/>
        <xs:element name="UsagePoint" type="xs:string"/>
        <xs:element name="Baseline" type="xs:string"/>
      </xs:sequence>
      <xs:attribute name="xmlns:xsi" type="xs:string"/>
      <xs:attribute name="xsi:noNamespaceSchemaLocation"
        type="xs:string"/>
    </xs:complexType>
  </xs:element>
</xs:schema>
```