BD4NRG
Big Data for Next Generation Energy

Deliverable D4.1: BD4NRG High Performance Data Processing

WP4 – Data Modelling & Open Modular AI-based edge-level Analytics Toolbox

Month YYYY

www.bd4nrg.eu

Big Data for Next Generation Energy

Disclaimer

Copyright Message

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613

Grant Agreement Number: 872613 Acronym: BD4NRG

Full Title	Big Data for Next Generation Energy		
Topic	DT-ICT-11-2019 Big data solutions for energy		
Funding scheme	H2020- IA: Innovation Action		
Start Date	January 2021	Duration	36
Project URL	http//www.bd4nrg.eu		
Project Coordinator	ENG		
Deliverable	BD4NRG High Performance Data Processing		
Work Package	WP4 – Data Modelling & Open Modular AI-based edge-level Analytics Toolbox		
Delivery Month (DoA)	M10	Version	1.0
Actual Delivery Date	10/01/2022		
Nature	Report	Dissemination Level	Public
Lead Beneficiary	ENG		
Authors	Marzia Mammina [ENG], Diego Arnone [ENG], Alessandro Rossi [ENG], Elissaios Sarmas [NTUA], Vaggelis Marinakis [NTUA], Vangelis Spiliotis [NTUA], Vasiliki Georgiadou [TNO], Elena Lazovik [TNO], Chris Develder [IMEC]		
Quality Reviewer(s):	Haris Doukas [NTUA], David Carro Santome [CS]		
Keywords	Big Data Analytics, Artificial Intelligence, Architecture		

Preface

BD4NRG focuses on addressing emerging challenges in big data management for energy sector with an innovative open holistic solution for smart grid-tailored, near real time, energy-specific and AI-based open Big Data Analytics modular framework. The vision is to deliver holistic services for techno-economic optimal management of Electric Power and Energy Systems value chain. Services range from optimal risk assessment for energy efficiency investments planning, to optimized management of grid and non-grid owned assets, improved efficiency, and reliability of electricity networks operation, while at the same time contributing to achieve fair energy prices to the consumers and laying the foundations for an EU-level energy-tailored data sharing economy. The BD4NRG Toolbox will be implemented and validated in real-life pilots in 12 large-scale demo sites across 10 countries for:

- Increasing the efficiency and reliability of the electricity network BD-4-NET
- Optimising the management of assets connected to the grid BD-4-DER
- De-risking investments in energy efficiency and increasing the efficiency and comfort of buildings BD-4-ENEF

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Who We Are

	Participant Name	Short Name	Country Code	Logo
1	ENGINEERING – INGEGNERIA INFORMATICA SPA	**ENG**	IT	
2	NATIONAL TECHNICAL UNIVERSITY OF ATHENS	**NTUA**	GR	
3	RHEINISCH-WESTFAELISCHE TECHNISCHE HOCHSCHULE AACHEN	**RWTH**	DE	
4	EUROPEAN DYNAMICS LUXEMBOURG SA	**ED**	LU	
5	INTERNATIONAL DATA SPACES EV	**IDSA**	DE	
6	EUROPEAN NETWORK OF TRANSMISSION SYSTEM OPERATORS FOR ELECTRICITY AISBL	**ENTSO-E**	BE	
7	PANEPISTIMIO DYTIKIS ATTIKIS	**UNIWA**	GR	
8	ATOS SPAIN SA	**ATOS**	ES	
9	FUNDACION CARTIF	**CARTIF**	ES	
10	UNIVERZA V LJUBLJANI	**UNILJ**	SL	
11	ENEL X SRL	**ENELX**	IT	
12	REN - REDE ELECTRICA NACIONAL SA	**REN**	PT	
13	CENTRO DE INVESTIGACAO EM ENERGIA REN - STATE GRID SA	**RDN**	PT	
14	UNINOVA-INSTITUTO DE DESENVOLVIMENTO DE NOVAS TECNOLOGIASASSOCIACAO	**UNINOVA**	PT	
15	ENERCOUTIM - ASSOCIACAO EMPRESARIALDE ENERGIA SOLAR DE ALCOUTIM	**ENERC**	PT	
16	FIWARE FOUNDATION EV	**FIWARE**	DE	
17	CENTRICA BUSINESS SOLUTIONS BELGIUM	**CENTRICA**	BE	
18	NEDERLANDSE ORGANISATIE VOOR TOEGEPAST NATUURWETENSCHAPPELIJK ONDERZOEK TNO	**TNO**	NL	
19	ASM TERNI SPA	**ASM**	IT	

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

	Participant Name	Short Name	Country Code	Logo
20	VIDES INVESTICIJU FONDS SIA	**LEIF**	LV	
21	COMSENSUS, KOMUNIKACIJE IN SENZORIKA, DOO	**COMSENSUS**	SL	
22	HOLISTIC IKE	**HOLISTIC**	GR	
23	INTERUNIVERSITAIR MICRO-ELECTRONICA CENTRUM	**IMEC**	BE	
24	TERRASIGNA SRL	**TS**	RO	
25	UBIMET GMBH	**UBIMET**	AT	
26	ELEKTRO LJUBLJANA PODJETJE ZADISTRIBUCIJO ELEKTRICNE ENERGIJE D.D.	**EKL**	SL	
27	BORZEN, OPERATER TRGA Z ELEKTRIKO, D.O.O.	**BORZEN**	SL	
28	AJUNTAMIENTO DE SANT CUGAT DEL VALLES	**AJSCV**	ES	
29	ELES, D.O.O., SISTEMSKI OPERATER PRENOSNEGA ELEKTROENERGETSKEGA OMREŽJA	**ELES**	SL	
30	E-LEX - STUDIO LEGALE	**ELEX**	IT	
31	OSMANGAZI ELEKTRIK DAGITIM ANONIM SIRKETI	**OEDAS**	TR	
32	VEOLIA SERVICIOS LECAM SOCIEDAD ANONIMA UNIPERSONAL	**VEOLIA**	ES	
33	STICHTING EG	**EGI**	NL	
34	CINTECH SOLUTIONS LTD	**CN**	CY	
35	EMOTION SRL	**EMOT**	IT	

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Contents

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Figures

Tables

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Executive Summary

Smart grid data analytics are playing an increasing and critical role in the business and physical operations of delivering electricity and managing consumption. Although utilities are starting from a difficult position with integrating data analytics into the enterprise, data science is a critical function if the mission of the modernised grid is to be achieved.

IoT applications bring further fundamental and technological challenges. The time is mature to rethink whether cloud computing is the only architecture able to support IoT applications. It is worth investigating an overall orchestration of the computational resources available today that can take advantage of the edge-fog-cloud continuum to guarantee a seamless execution of automated analytical tasks without compromising the accuracy of their outcomes. Furthermore, managing retention times between automated analytical tasks is critical for handling high/low latency of existing data life-cycles that are encountered when supporting IoT applications.

BD4NRG intends to tackle the barriers for full-scale realisation of the data economy in the energy value chain, and to create a real data economy for energy by proposing an incentivizing conceptual and technological framework for facilitating and motivating energy stakeholders to share data. Combined data from different domains (energy, climate, economics and financing, user behaviour and societal trends, regulations and many more) could pave the way for a wide spectrum of novel services, while exploiting the edge/fog/cloud paradigm and the opportunities it provides, concerning both novel data management architectures and data processing approaches. The overall vision and main objective of BD4NRG is to deliver an innovative framework which leverages on AI-based cross-sector analytics for integrated and optimised smart energy grid management (including operation and planning), based on seamless data-information-knowledge exchange under respective sovereignty and regulatory principles.

This document describes data analytics, the challenges of implementing those analytics, and how BD4NRG intends to provide a comprehensive data analytics framework allowing for a reliable and efficient grid.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

1 Introduction

1.1. Purpose of the Document

Deliverable 4.1 "BD4NRG High Performance Data Processing" is produced within the Work Package (WP) 4 "Data Modelling & Open Modular AI-based edge-level Analytics Toolbox" as a part of Task 4.1 "Distribution Enablers & Open Fog Flexible Management at the Edge". The purpose of this document is to present the BD4NRG innovative capillary distributed architecture including the technology enablers for a flexible fog management at the edge. The proposed architecture enables to manage the heterogeneous set of edge devices and make them transparent to the upper architectural layers, as well as to enable edge node selection and/or micro service placement, together with capabilities for infrastructure management. To this end, the document also leverages the results from other WP4 tasks and from WP2 and WP3.

1.2. Relation to other documents

This document relies on the deliverable D2.5 "BD4NRG Reference Architecture First Version", which provides information on BD4NRG-specific systems that address the relevant aspects of the reference architecture (RA), such as data analysis, governance, and marketplace. Additional information on BD4NRG-specific systems is provided by the deliverable D3.1 "BD4NRG GOVERNANCE – First Technology Release", where an overall view of WP3 components, especially those present in the 1st Technology release of the BD4NRG Governance Layer, is given.

1.3. Acronyms and Abbreviations

Table 1: Acronyms and abbreviations

Acronym	Description
AI	Artificial Intelligence
a.k.a.	also known as
AMI	Advanced Metering Infrastructure
API	Application Programming Interface
ASA	Active Smart Grid Analytics
BDA	Big Data Analytics
BDAaaS	Big Data Analytics-as-a-service
DaaS	Data-as-a-Service
DLT	Distributed Ledger Technology
DSO	Distribution System Operator

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

DSP	Distributed Stream Processing
EPS	Electro Power Systems
ESCO	Energy Service Company
GE	General Electric
GIS	Geographic Information System
HPCaaS	high-performance Computing as a Service
HTAP	Heterogenous Data Capture and Adaptive Polyglot Persistence
IDS	International Data Spaces
IDSA	International Data Spaces Association
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
ISO	International Organization for Standardization
IT	Information Technology
LSP	Large Scale Pilot
MAPE-K	Monitor-Analyze-Plan-Execute over a shared Knowledge
ML	Machine Learning
MLaaS	Machine Learning as a Service
OT	Operational Technology
PaaS	Platform-as-a-Service
PMU	Phasor Measurement Unit
QoS	Quality of Service
RES	Renewable Energy Sources
SCADA	Supervisory Control And Data Acquisition
TSO	Transmission System Operator

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

USD	US Dollar
VM	Virtual Machine

1.4. Document Structure

This document is organised as follows. Chapter 2 introduces the big data analytics (BDA) in the context of smart grid: state of the art, challenges, opportunities and explains how BD4NRG intends to take those opportunities. Chapter 3 introduces potential approaches for high-performance big data processing and presents the approaches adopted in BD4NRG. Chapter 4 describes the edge/fog/cloud paradigm and the opportunities, both in terms of architectures and data processing approaches for exploiting it; the architecture of main BD4NRG systems is presented and how BD4NRG approaches the edge/fog/cloud continuum is described. In Chapter 5 conclusions are given.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

2 Big Data Analytics in Smart Grids

The global electricity sector is undergoing a rapid and significant transition in the way energy is produced, distributed and consumed that is motivated by the urgency to ensure secure energy supply, achieve sustainable development, and limit climate change. Renewable energy sources (RES), such as solar and wind sources, are not just added to the generation side by the utility companies, but also by the end consumers and microgrids. As a counter effect of the increasing diffusion of renewables, innovative flexibility measures are considered to balance the demand and supply all the time, given the variability in the electrical system due to the unpredictability of RES[1]. Novel approaches are required to improve the flexibility of the energy system ranging from supply to demand side. New technologies are revolutionizing the end-user side by changing the electricity consumption patterns in order to improve energy efficiency and optimizing allocation of power. Also, vehicle to grid technology has provided opportunities for power flow management and the management of its flow from vehicles to the grid. In its transition, the power system is being witnessed a massive installation of smart meters for monitoring the near real-time usage of electricity while also collecting and communicating the data to the utilities.

The principle of smart grids solves the problems of power demand by providing two-way power and information flow between consumers and utility. The development of a smart grid is fully associated with the big data flow. Currently, many utilities are working on installing a large number of smart meters and utilizing this data for effective demand response and resource management and the generated data are significantly large. The high volume of raw data generated is not directly comprehensible or useful without a dependable and consistent ability to process, analyze, and understand the information contained. Thus, data should be transformed into useful information before action can be taken based on the data. The high volume of data obtained from various smart grid sources satisfies the characteristics of big data in terms of Volume, Velocity, Variety, Veracity, and Value[2]. These characteristics are the challenges when dealing with BDA along with other major concerns such as security, privacy, etc. There are various prospective applications of BDA on smart grid data such as real-time and automatic processing of the electrical consumers' energy consumption, automatic billing, intelligent energy planning and pricing analysis, detection of outages due to faults and anomalies, load and generation forecast under high unpredictability, load management with demand response, and asset management. An overview of the application of BDA in the smart grids is given in the following.

2.1 State-of-the Art

Discovering new value from high volumes of Internet of Things (IoT) data is one of the biggest opportunities in energy today. Utility companies can utilize smart meters to measure how energy resources are used. The opportunity to process and analyse streaming smart meter data and billion row datasets in real time could lead to optimisation of energy generation and uptime, prediction and

[1] https://ieeexplore.ieee.org/document/7977693

[2] D. Syed, A. Zainab, A. Ghrayeb, S. S. Refaat, H. Abu-Rub and O. Bouhali, "Smart Grid Big Data Analytics: Survey of Technologies, Techniques, and Applications," in IEEE Access, vol. 9, pp. 59564-59585, 2021

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

prevention of power outages, charting energy trends, understanding customer energy usage, and improving load forecasts.

Despite the success information technology has achieved in the field of BDA, electrical industries are only beginning to deploy big data. Only few commercial solutions for BDA specifically for smart grids are available in the market. The smart grid data analytics market was valued at US dollar (USD) 1,615 million in 2020, and it is expected to reach USD 3,289.15 million by 2026, registering a CAGR of 12.76%, during the period of 2021-2026. The smart grid data analytics market is fragmented and highly competitive in nature. Because of the emergence of new startups offering a broad range of innovative solutions catering to diverse industry requirements, the market has been witnessing intensifying competitive rivalry. The major players in the market have been considered to be synonymous with good performance and as such they are expected to have a competitive edge. Among others, major players include Siemens AG, Itron Inc., General Electric Company (GE), IBM Corporation, and AutoGrid Systems Inc. (Figure 1).[3]

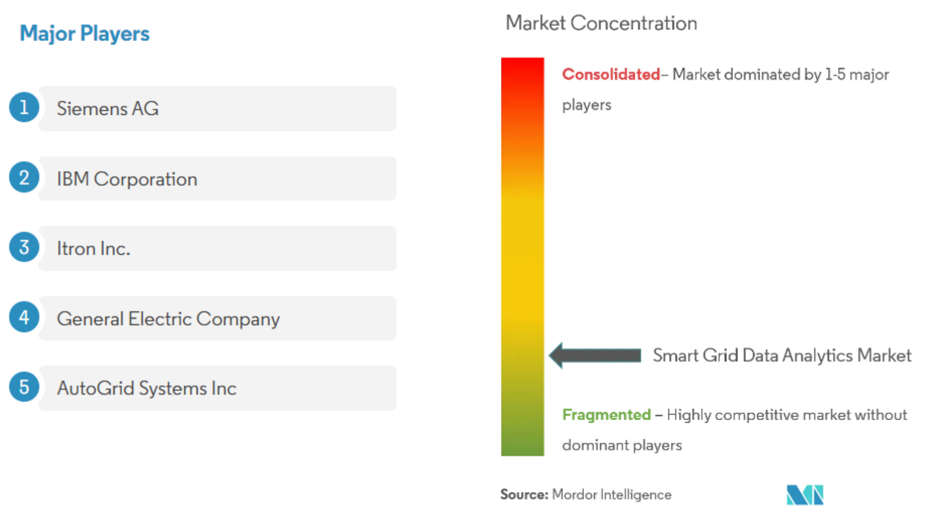

Figure 1: Smart Grid Analitycs Market Top Players[3]

Siemens has developed a big data platform, called EnergyIP Analytics[4], which adds big data to smart grid application. EnergyIP is an extremely powerful, flexible, and scalable platform able to handle millions of sensors and the huge volumes of data generated. EnergyIP Analytics – Power Quality leverages the existing AMI infrastructure and SIEMENS in-depth knowledge of grid architectures to monitor continuously parameters like voltage, reactive power, or momentary outages on the most detailed level possible. The increased level of detail puts utilities back in control and offers instant benefits. Grid analytics can allow utilities to use big data for multiple functionalities, including home energy management, grid energy management, and predictive/corrective controls.

Itron released a set of specialised analytic software applications built on the Active Smart Grid Analytics™ (ASA) platform. Each application is bundled with Itron Analytic Accelerators and Extension Toolkits, which help utilities take full advantage of their smart grid data. Built to leverage Itron

[3] https://www.mordorintelligence.com/industry-reports/smart-grid-analytics-market

[4] https://assets.new.siemens.com/siemens/assets/public.1541967053.5d98bdbca52e366847a73bd6621f7e00d417c42c.energyip-analytics-suite-power-quality-en.pdf

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

industry-leading metering, communication network and meter data management solutions, these analytic applications provide actionable business intelligence to enable specific smart grid solutions for distribution operations, revenue cycle services, and customer engagement. Each analytic application comes with an Accelerator and Extension Toolkit. Each Accelerator offers immediate value by providing a fully functional set of analytic dashboards and reporting tools, but also establishes an ideal launching point to develop customized analytic solutions using the ASA Extension Toolkits.[5]

GE has developed an industrial IoT platform, called PREDIX[6], which provides secure and scalable edge-to-cloud data connectivity, analytics processing, and many other underlying services, including dashboarding and visualisations, interactive analysis, policy workflow engine, alerting, case management and more.

In the Redguide™ publication[7], IBM® proposes a Smart Grid Analytics and Sensemaking framework is based upon various devices, data, and analytics. A smart grid is an ecosystem of large interconnected nonlinear systems. This proposed solution instance focuses on the use of context-awareness analytics to maintain correct values (current and historical) for nodes and edges. Key to the analytics is the use of the Mehta Value, which is composed of a base reference, drift, and context-referenced phase angle data. Real-time decisions, such as load shedding or pathway selection, can then be made based upon the combination of contextually correct data and analytics, such as the Mehta Value.

AutoGrid Systems Calif. developed an "energy data platform" that uses smart meter, sensor and other utility data to generate a real-time assessment of energy consumption and recommendations for efficient power distribution. The predictive control technology help Electro Power Systems (EPS) customers forecast energy demand and optimize energy storage and distribution systems for variable sources like solar and combined heat and power plants. Analytics algorithms implemented allow to aggregate energy storage and distribution resources; the combined resources can then be used for grid applications, such as managing demand response or determining how much renewable energy can be traded in the electricity market.

ABB has integrated cloud computing and BDA in the ABB Asset Health Center[8], which provides solutions for processing big data for smart grid applications. ABB Asset Health Center is a software-assisted service for systematic diagnosis and analysis of ABB control system settings and conditions. It automatically collects system performance, configuration, and life-cycle parameters and delivers actionable analytics and detailed recommendations for maintaining optimum control system performance to users.

GE Grid IQ™[9] insight is a BDA architecture, which works based on the foundation of PREDIX data analytics platform and allows utilities to monitor, manage, and control their grid more intelligently. It is a cloud-based grid management fee-for-service system aimed at meeting the operational technology needs of small and mid-market utilities that want to avoid the overhead expenses involved with developing their own smart grid management networks.

---

[5] https://investors.itron.com/static-files/08c18169-7fb9-454b-a4ac-664597e2ea7e

[6] https://www.ge.com/digital/iiot-platform

[7] https://www.redbooks.ibm.com/redpapers/pdfs/redp5082.pdf

[8] https://new.abb.com/power-generation/service/advanced-services/system-performance

[9] https://www.ge.com/news/press-releases/ges-grid-iq%E2%84%A2-solutions-service-key-internet-technology-helping-utilities-develop

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

The N-SIDE Energy Forecasting Platform[10] and related services enable system and network operators to build and customize predictive and prescriptive forecasting and decision support solutions for a wide range of demonstrated use case or for building a new one. Forecasting the evolution of parameters such as the system imbalance, the available flexibility on the market, or the load at a specific substation becomes crucial to anticipate and adapt to a rapidly changing situation, in a world where variable, decentralized energy resources and transnational coordination are adding complexity to system operation.

GridCure[11] is an alternative to on-premises data analytics, leveraging a secure cloud environment to offer sophisticated operational insight to any user no matter where they or the systems-of-record are located.

Those industries are offering utilities a way to gain understanding of what is the state of the grid and developing a launching pad for smart grid BDA applications over time. The next step is to integrate prognosis and diagnosis into the BDA framework effectively in order to facilitate utilities to provide situational awareness, informed predictive decisions, condition monitoring, health management of critical grid infrastructure, and supporting grid functionalities. Indeed, utilities are becoming more and more aware about the consequence the increasing data means to their traditional operations and have to adopt the necessary strategies to obtain value from this huge amount of data.

A clear understanding of big data architecture is necessary to identify how big data can be integrated with the existing power system control and operational architecture, how they differ from traditional computational environments, and what scientific and technological standardisation challenges need to be faced to deploy big data solutions.

BD4NRG intends to offer new market opportunities in the energy sector by unlocking and exploiting the potential of big data to enable improved operation for all stakeholders in the energy value chain. In contrast to existing commercial solutions, which require a high level of expertise, time, and cost to add artificial intelligence (AI) to embedded products, BD4NRG aims to provide access to advanced tools and services that can be obtained through a marketplace and an eco-system of collaborative stakeholders. BD4NRG aims to provide the necessary capabilities for pushing forward big data management and processing towards the edge, with a view to reduce some of the major barriers actually hampering big data and analytics full scale deployment in the smart energy grid domain, e.g., the excessive communication costs and the need for keeping generated data as much as possible closer to the generation point and the owner. In that respect, a number of technology enablers will be integrated and deployed to support AI-based learning, including privacy preserving federated learning on semantically structured data streams for mitigating the risk to share sensitive data, cross-context transfer learning combined with self-learning to promote the reuse of machine learning (ML) models in contexts with few or poor quality of data.

The flexible BD4NRG marketplace will leverage and combine tokenisation principles of Distributed Ledger Technology (DLT)/blockchain and smart contracts to enable the value- and incentive-based mono-commodity and/or multi-commodity sharing and trading of data, AI models/algorithms, edge storage and computing resources. Furthermore, an analytics toolbox will allow to combine and make

[10] https://energy.n-side.com/

[11] https://gridcure.com/

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

available to end users user-friendly graphical working environment capabilities to support custom selection of local and/or third party assets including: discovery of heterogeneous data sources formats, a huge variety of analytics ML algorithms/models to be selected among correlation, categorisation and feature extraction to implement a variety of analytics types (descriptive, predictive, prescriptive), a large variety of presentation modalities (reports, charts, etc.), edge-level available storage and computing resources.

Moreover, IDSA-compliant DLT/blockchain/smart contract -based implementation of an energy data decentralized governance technological enabler. It will include data quality assessment, certification management and regulatory/privacy/cybersecurity context-aware semantic data access brokerage service and provenance tracking management. The main aim will be to facilitate and enable seamless open multi-lateral B2B cross-stakeholder data sharing, trusted data exchange and handling, while allowing full data sovereignty and control of respective data ownership, access, security and protection.

2.2 Key Challenges

The term “big data“ is sometimes a source of confusion or controversy even if it is self-explanatory. For example, what can be considered big data by an electric utility, can be seen, instead, as moderate size data for data-centric enterprises, although the relativity of the term is recognized even within the information technology (IT) community. A definition often used for big data is a ‘‘high-volume, high-velocity and high-variety information asset that requires and demands cost-effective, innovative forms of information collection, storage, and processing for enhanced insight and decision making’’[12]. From this definition, it is evident that the size, that is the volume, is not the unique factor that characterises big data, as there are other factors too.

Volume refers to the amount of data. Many IT-related organisations define bid data in terabyte, sometimes petabytes. The amount of data being generated by electric utilities is being increasing exponentially at an exponential rate and due to increased deployment of intelligent devices (e.g., smart meter, electric vehicles, and inverters) on the consumer side and their active engagement on different grid services, the data management is being increased at the consumer level as well, in the order of hundreds of tera bytes. Therefore, effective management of huge volume of data is becoming increasingly challenging issues for utilities. New innovative technologies, such as distributed and scalable computing architecture are necessary to afford the management of those huge quantities of data. For example, a reduced representation of the data set that is much smaller in volume while maintaining the integrity of the original data, the so-called dimensionality reduction, can allow to reduce significantly the data complexities.

The scope of big data also affects its quantification. For example, one smart meter, with resolution of seconds to minutes generates much fewer data than one phasor measurement unit (PMU), with resolution of milliseconds; an advanced metering infrastructure (AMI) may generate large volume of data coming from millions of customers.

Variety refers to the diversity of data types. Compared to traditional data systems, nowadays, data comes from a much greater variety of sources. Traditional structured data, (e.g., tables and other data

[12] De Mauro, A., Greco, M., Grimaldi, M., 2016. A formal definition of big data based on its essential features. Libr. Rev. 65 (3), 122–135.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

structures of relational databases, record formats, character-delimited rows of many flat files) are now joined by unstructured data (e.g., text, voice, and video), that is unorganized data that can come in different files or formats, and semi-structured data, that is data that have not been organized into specialized repository but have associated information, such as metadata, that makes it easier to process than unstructured data(e.g., JSON, RSS feeds, and hierarchical data). The challenge in variety concerns the standardisation and the distribution of the data.

Velocity refers to how quickly data are generated and how quickly data moves. This is a key aspect of big data: data are expected to be at the right times to make the best business decisions possible. An example is a continuous stream of data, as opposed to once-in-a-while event triggered data from a sensor. Although the majority of power system sensors are event-triggered, there are also sensors, e.g., PMUs both at transmission and distribution level, that produce data streams at high rates.[13]

The so called ''three Vs'' of big data, see Figure 2, are described as follows.

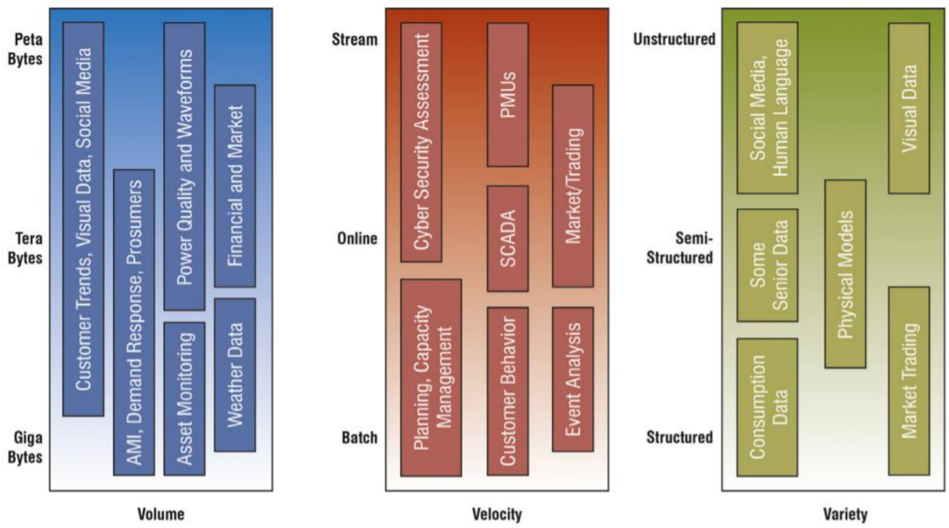

Figure 2: The 3 V attributes of BD and some examples in the context of power systems[14]

In addition to the ''three Vs'', other "Vs" have been defined.

Veracity refers to the quality and accuracy of data; represents the level of trust there is in the collected data. Indeed, inaccurate or messy data may not be able to provide real and valuable insight. Causes of uncertainties and loss of data quality can be due, for example, to sensor inaccuracy, communication latencies/delays, cyber-attack, physical damages of equipment, time unsynchronised data, missing/inconsistent data, noises etc. Those uncertainties may result from various reasons, for instance, readings of sensors are uncertain because of sensor aging or malicious attacks during data acquisition and control processes. This requires innovative techniques to deal with data mining and data analytics techniques.

[13] Bhattarai, B.P.; Paudyal, S.; Luo, Y.; Mohanpurkar, M.; Cheung, K.; Tonkoski, R.; Manic, M. Big data analytics in smart grids: State-of-the-art, challenges, opportunities, and future directions. IET Smart Grid. 2019, 1–15.

[14] Akhavan-Hejazi H, Mohsenian-Rad H. Power systems big data analytics: An assessment of paradigm shift barriers and prospects. Energy Rep 2018;4:91–100.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Value relates to what can be done with the data that have been collected. The value of big data increases depending on the insight that can be obtained from them.

Smart grid data mostly involve consumer privacy information, commercial secrets, and financial transactions; therefore, data security (e.g., privacy, integrity, authentication) is very crucial.

The power consumption of consumer normally provides insights on their behaviour. This is the reason why the data privacy is a critical concern. Data aggregation is one of the common approaches to address data privacy issues, in addition to other techniques such as, for example, distributed aggregation, differential aggregation.

Data integrity refers to those methodologies adopted primarily to prevent unauthorised modification of information. However, due to close interdependencies between power and communication infrastructure, the power industry is also susceptible to increased cyber/physical-attacks, that could not only deliberately modify financial transactions, but also severely mislead the utility operational decisions. Privacy-preserving data aggregation schemes are an example of a methodology that can help ensuring data integrity through a digital signature or a message authentication code.

Furthermore, in order to distinguish legitimate and illegitimate identities, authentication systems have to be deployed to protect smart grid data. Different techniques such as data encryption and signature generation are normally used for data authentication and security management in smart grids.

The smart grid data also might present issues on indexing and query processing, particularly when real-time applications are requested for big data management, where traditional methods based for example on SQL may not suffice. Therefore, advanced data indexing and query-processing algorithms will play critical roles in smart grid BDA.

The increasing need for real-time control and communication in the smart grid requires great attention on time synchronisation. Synchronised data allow analysts to determine connections between events and help both forensic analysis of past events, near real-time situational awareness and informed predictive decisions. The determination of a sequence of past events and real-time situational awareness of the grid health can aid in providing a preventive or remedial solution.[14]

2.3 Opportunities

As depicted in Figure 3, the types of data that can form big data in power systems can be classified into domain data and off-domain data. The domain data can be further categorized by their sources, for example:

- Telemetry and Supervisory Control And Data Acquisition (SCADA) data, which enable continuous flow of measurements on grid equipment status and parameters and other grid variables.
- Oscillographic and Synchrophasor data, voltage and current waveform samples in time or frequency domains that can create a graphical record.
- Consumption data, generally provided by the smart meters.
- Asynchronous event data, often produced by devices with embedded processors generating messages under a variety of normal and abnormal conditions.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

- Grid metadata, which may include internal sensor data, calibration data, and other device-specific information.
- Financial data, such as day-ahead and real-time market bids and price data, bilateral transactions, and retail rates.

Power grid operation also relies on different forms of off-domain data, which are not specific to or necessarily intended for the power sector, such as weather data and Geographic Information System (GIS) data used to enhance power system operations. Many forms of off-domain data are emerging, which can be exploited for the power grid operations and energy enterprise (e.g., traffic data, social media data, trade indices).

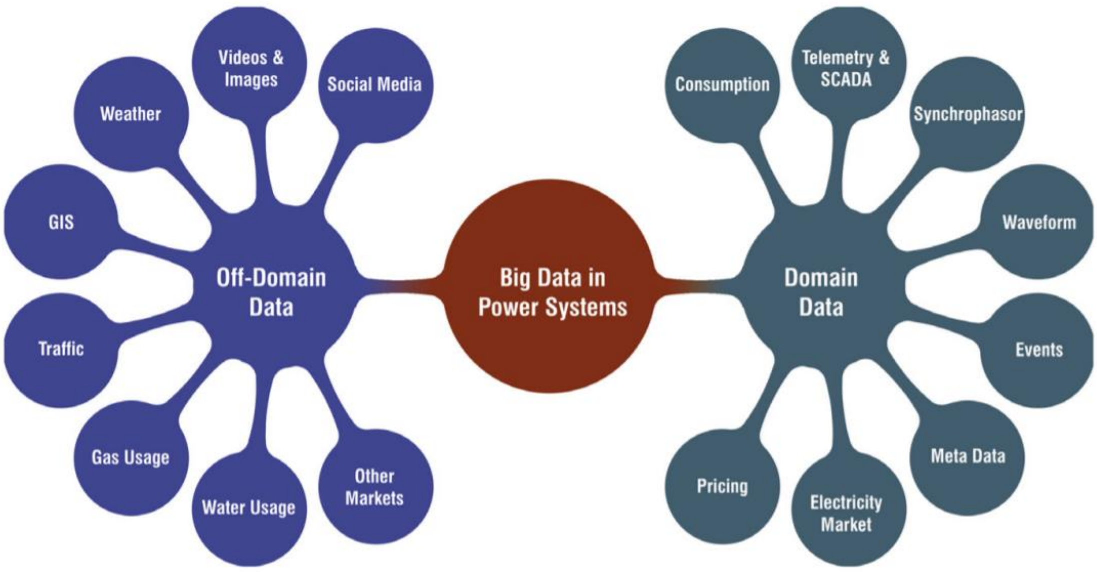

Figure 3: Classification and examples of big data types in power systems[14]

Key steps for big data analyses include data acquisition, data storage, data analytics, and operational integration. The huge amount of data and novel data analytics enables the technologies for better decision-making by providing methods and tools that can support the smart grid objectives of higher resilience, economic efficiency, and reduced emissions, while providing higher levels of service quality and customer satisfaction. Data analytics represent a transformational step toward the future grid.

Data analytics are designed to extract hidden information and patterns within a large dataset that can be transformed into useful outcomes/knowledge by using various algorithms and procedures (e.g., clustering, correlation, classification, regression, and feature extraction). The use of big data techniques and technologies in the entire process, from data storage to data analytics increases the utility (Figure 4). Data from the distributed sources must be transmitted as fast as possible to the analytics tools; therefore, data integration and storage in smart grids are of utmost relevance. Once collected grid data are combined, cleaned, and scaled, they are ready for batch, near-real time, or real time analysis for different purposes, such as forecasting, event detection, customer profiling, anomaly detection. Lastly, visualisation technologies are used to interpret the results of the analysis or for monitoring purposes.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Analytics is the application of advanced analysis methodologies to the data – including predictive and prescriptive analytics, forecasting and optimisation. The opportunities for smart grid analytics are expanding because of the exponential data available to develop analytical models.

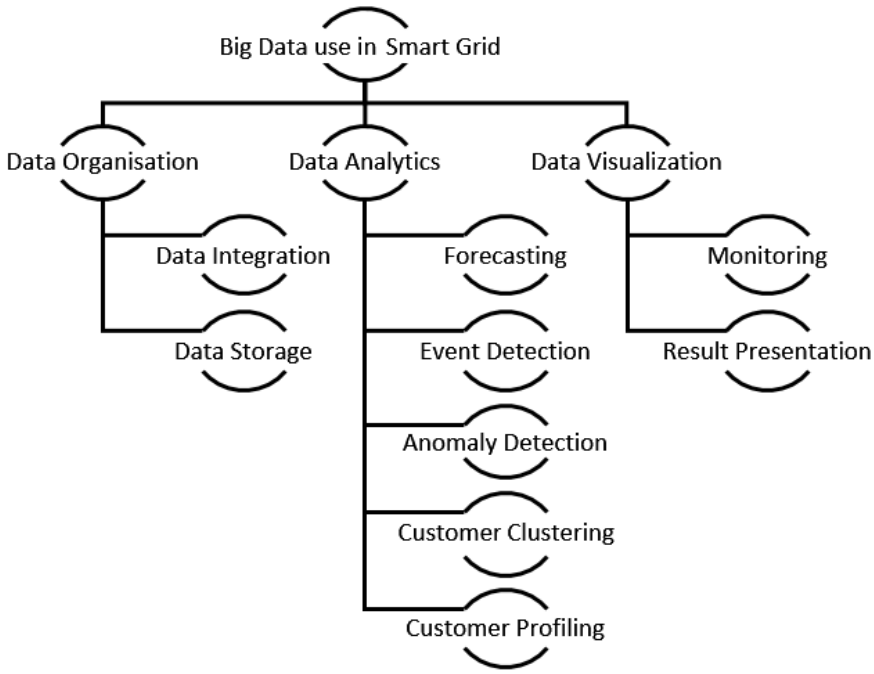

Figure 4: Big data for smart grid[15]

From the perspective of smart grid applications, data analytics can be grouped into four categories, as represented in Figure 5. Event analytics cover diagnosis/detection of the power systems events, such as faults and outage managements and also include a descriptive analysis of previous power system events using various techniques (e.g., classification, filtering, and correlation).

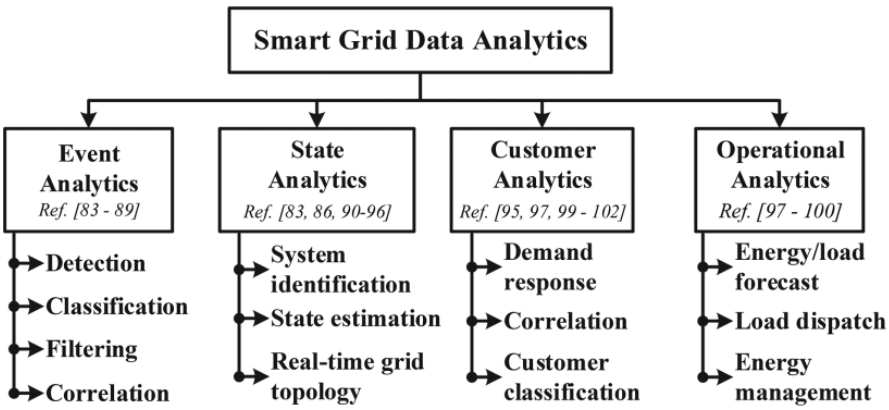

Figure 5: BDA for smart grid[13]

Some of the key application areas of event analytics are detection of abnormal operating conditions including fault detection, system outage detection, detection of malicious attacks. Application of the

[15] Sinanç, Duygu & Arslan, Bilgehan & Sagiroglu, Seref. (2018). Smart Grid Security Evaluation with a Big Data Use Case. 10.1109/CPE.2018.8372536.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

state analytics includes state estimation, system identification, and grid topology identifications. Sample applications for operational analytics include energy/load forecast, energy management and dispatch of resources. Similarly, key power system applications that fall under customer analytics include customer classification/categorisation, the correlation between consumer behaviour and energy consumption patterns, and demand response.

Analysts can mine data, such as real-time asset metrics and weather factors, and then apply smart grid analytics for optimizing the performance of connected devices in the field. The use of smart grid analytics is a key point for controlling operating costs, improving grid reliability, and delivering personalized energy services to consumers.

Data analytics is not new for the utilities, but smart grid analytics enables the ability to embed analytical insights into operational systems, often for automated responses, thus providing a more direct link between IT and OT (Operational Technology). Some areas where utilities are already seeing success are the ones reported below.

Smart generation: The modernisation of power plants and substations by putting sensors on the main components, such as turbines and transformers, looking for vibration or other anomalies and the implementation of asset analytics that could predict future failures have already paid off utilities by preventing major outages related to equipment.

Grid analytics: Since technicians cannot be everywhere in a service territory, analytics can be crucial in monitoring the field. Sensors are able to pull data in real-time for immediate identification of impending trouble and taking advantage of such technology, some utilities are beginning to look at the application of synchrophasor technology in the transmission and distribution systems, looking for any disturbance that might indicate a voltage collapse or asset failure.

Revenue protection: Utilities around the world lose billions a year in revenue from customers who do not pay their bills or who tamper with meters to avoid charges. Smart grid analytics can detect this theft and fraud.

Consumption analytics: Analytics help utilities by supporting smarter decisions about unit commitment, outage planning, fuel procurement, rate and regulatory factors. This can have a big influence on the bill of their customers; since, for example, about half of utility bill each month is for fuel consumption, if the utility can manage intelligently the fuel procurement, the bill will be reduced.

Enhanced Demand Response: Until a few years ago, only large utilities and large customers were involved demand response, due to the hurdles in management and value proposition for the operation of huge number of small loads. BDA enable utilities (through cloud-based platforms employed by utilities) to look simultaneously at the consumption patterns of millions of users and rapidly determine which customers would be willing to participate in a demand response event, how much these customers will charge for participation, and how much will actually be saved.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Customer research and analytics: Analytics are driving the work of designing new products and services that will improve value to customers.[16]

The great amount of domain and off-domain data, the BDA technologies, the possibility to cross the data with the once from other energy sectors, are the premise for many opportunities for exploiting the value of such data, by also implementing new business models.

2.4 Typical Utility Analytics Maturity Model

Tactical solutions tend to gain a long-term life, which inhibits the organisation ability to realize end-to-end benefits across the business. Utilities that are able to leverage end-to-end analytics capabilities have great possibilities to be better positioned as market leaders.

A typical analytics maturity model[17] describes the stages a utility must take to maximize end-to-end business benefits from big data and analytics solutions. As shown in Figure 6, four stages of analytics maturity are envisioned, from "nascent" stage, where the utility has low awareness of analytics and the value that can be gained, trough "pre-adoption", where the utility conducts an initial assessment of data analytic to prepare a strategic plan, and "mature" stage, where the utility has strategic and scalable technical capabilities to deliver enterprise wide analytics solutions across business areas that leverage smart grid, smart meter, and customer data, up to the "visionary" stage, where the utility has generated significant business value from analytics by continuously focusing on improving its technological capabilities.

The maturity model is based on the evaluation of the key parameters in Figure 7.

[16] https://www.sas.com/en_us/insights/articles/data-management/the-opportunity-of-smart-grid-analytics.html

[17] Digital Systems & Technology - A Practical Approach for Power Utilities Seeking to Create Sustaining Business Value with Analytics. Cognizant 20-20 Insights

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

MATURITY LEVEL	DESCRIPTION
NASCENT (STAGE 1)	A utility at this stage has low awareness of analytics or its value across much of the value chain. It may have stand-alone dashboards and reporting capabilities and disparate analytics processes within each business area.
PRE-ADOPTION (STAGE 2)	A utility at this stage has conducted an initial assessment of data analytics and has some form of strategic plan in place. There are tactical capabilities and processes implemented within the departments to leverage smart meter and grid data. However, difficult obstacles remain, such as data integration, as well as tools and human resource skill-set shortcomings.
MATURE (STAGE 3)	A utility at this stage has strategic and scalable technical capabilities to deliver enterprise-wide analytics solutions across business areas that leverage smart grid, smart meter and customer data. There is a significant shift toward data-driven business processes and a redefinition of the operating model.
VISIONARY (STAGE 4)	A utility at this stage has generated significant business value from analytics by continuously focusing on sharpening its technological capabilities. The utility applies modern approaches such as intelligent automation and advanced machine learning technologies for grid automation and customer service to stay ahead of the curve.

Figure 6: Typical stages of analytics maturity[17]

TOWERS	SUB-TOWERS		
1. Organizational Strategic Alignment/Use Case Approach	1A. Analytics Vision & Strategy	1B. Regulatory Readiness	1C. Customer Readiness
2. Data Quality/Governance	2A. Data Strategy	2B. Data Quality	2C. Data Governance
3. Technology/Infra/Internal Resource Capability	3A. Technology/Tools	3B. IT Architecture	3C. IT Infrastructure
4. Organizational Culture and Decision-Making Process	4A. Organizational Structure/Ops Model	4B. Decision-Making	4C. Human Resource Skills/Culture
5. Analytics Adoption	5A. In-Practice Capability	5B. In-Progress Capability	5C. Near-Future Capability

Figure 7: Key parameters of an analytics maturity model[17]

2.5 Research Directions

Few standard information models for smart grid interoperability have been defined so far, such as the International Electrotechnical Commission (IEC) 61850, IEC 61850-90-7, IEC 61970/61968, Institute of Electrical and Electronics Engineers (IEEE) 1815, and IEEE 2030.5, but none of them define information models to describe interoperability among various BDA platforms, architecture, and their operational integrations with utility decision frameworks. Storage, usage, dissemination, and sharing of data with utility operational frameworks are not unified. This lack leaves space for extensive research activities for implementing interoperability among different devices, network operations, data analytics platforms, big data architecture, data repository, and information models.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Few standards and regulatory frameworks for sharing data among utilities, weather forecasting providers, and other energy systems (e.g., transportation, oil, and gas sectors) are established.

Technical standards should be established for maximising the value of big data and for ensuring data exchanges among different entities. A regulatory framework could be also beneficial to bind the entities with legal rules and regulations in terms of data sharing. Moreover, the involvement of an impartial third party could help in making fair estimation and justification of the costs associated with big data deployments for regulated markets and different entities. Therefore, efforts from professional communities are expected to be invested in establishing standards for data sharing among platforms/architectures and identifying the elements of regulatory frameworks to bind utilities in deploying big data.

The International Data Spaces Association (IDSA) provides a standard that enables the sovereign sharing of data with clearly defined usage rights[18]. The DIN SPEC 27070 standard guarantees data security and data protection for all parties involved, establishes mutual trust among them, ensures a level playing field by means of a jointly developed and commonly accepted design, and enforces data sovereignty for all data providers. Based on the IDS Reference Architecture Model, DIN SPEC 27070 is the first global and interoperable standard – soon to be also an International Organization for Standardization (ISO) standard. Digital platforms, applications in the fields of AI, IoT, or big data projects should not be conceived without IDS concepts. *The International Data Spaces (IDS) standard is an indispensable element for a trustworthy data sharing infrastructure in Europe – enabling the creation of all kinds of data spaces, data sharing ecosystems and data marketplaces. The IDS standard guarantees data security and data protection for all parties involved, establishes mutual trust among them, ensures a level playing field by means of a jointly developed and commonly accepted design, and enforces data sovereignty for all data providers*[19].

Future smart grid applications shall use multiple sources of big data (such as data weather, traffic, oil and gas industry, social media etc.), which can help in assessing the dependence of critical infrastructure on power grids, by uncovering crucial hidden information otherwise not possible from electrical measurements only. Therefore, data hubs should be created and be readily accessible to advance resiliency of critical infrastructures.

Most of the existing big data application for electric utilities have been conceived for system monitoring and operational planning. BDA should also be integrated into real-time control and operational module so as to provide real-time situational awareness and informed predictive decisions. Processing of massive data in real time has inherent computational and scalability issues, that should be in the focus of research activities.

Given the huge volume of smart grid data, distributed and parallel intelligence is requested to address effectively data computation. Distributed intelligence and coordination algorithms should be developed. Meta-modelling, dimensionality reduction, and edge computing should be introduced to reduce the computational and communication burdens.

[18] https://internationaldataspaces.org/ids-officially-a-standard-din-spec-27070-is-published/

[19] https://investforesight.com/standard-for-sovereign-sharing-of-data/

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Advanced AI technique, such as deep ML is essential not only for exploiting patterns within the data, but also for making the decision process less reliance on human interference. Due to the growing in volume, variety, and veracity of smart grid data, the scalability of the ML models is very critical and, since the timely and accurate capture of hidden information is a key factor in the operation of electrical infrastructure, accuracy and computational efficiency should be taken into account as well.

Furthermore, since the key benefit of BDA is to help utilities in taking actions based on real-time situational awareness derived from the data analytics, integration of advanced visualisation mechanisms with automated operation provides directive information to the operators and avoids the need of intuitive decisions. The co-design of smart grid BDA and advanced visualisation can produce a seamless integrated framework, which can reduce security risk and help to take effective decisions.[13]

BD4NRG intends to exploit those research opportunities by addressing the emerging challenges in big data management in the energy sector by providing an open holistic solution that proposes a European approach to the business-to-business platforms able to provide competitive edge services to the stakeholders involved in the energy value chain and to create new market opportunities.

BD4NRG intends to tackle the barriers for full-scale realisation of the data economy in the energy value chain, and to create a real data economy for energy by proposing an incentivizing conceptual and technological framework for facilitating and motivating energy stakeholders to share data. Combined data from different domains (energy, climate, economics and financing, user behaviour and societal trends, regulations and many more) could pave the way for a wide spectrum of novel services. The overall vision and main objective of BD4NRG is to deliver an innovative framework which leverages on AI-based cross-sector analytics for integrated and optimised smart energy grid management (including operation and planning), based on seamless data-information-knowledge exchange under respective sovereignty and regulatory principles.

BD4NRG intends to align and integrate the existing specifications for data-driven initiatives and underlying B2B Reference Architectures at the interplay among smart energy grids, AI, IoT, big data, smart home/building, industry/manufacturing, to provide a living reference architecture specifically tailored to the energy value chain, which will lay the foundation for facilitating seamless B2B multi-lateral interoperability.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

3 Approaches for High-Performance Big Data Processing

According to a study conducted by Accenture[20], the data volume a utility must handle in a smart grid environment is multiple orders of magnitude greater than operating a traditional grid, because in addition to detailed meter information, a wide range of new intelligent devices and data types are emerging. The implementation of new tools, architectures, and processes is a prerequisite for managing smart grid data and effective measurement, control, and optimisation of a smart grid requires new approaches to the related analytics and visualisation capabilities. The effective use of the right analytics turns the mass of data into usable information. If designed properly, the data architecture will provide utilities with the capabilities that are needed to deal with future change and evolution in their smart grids and business environment. To this end, the architecture will need to include not only data storages, but also elements such as data management systems, services, and integration busses to share data and information effectively.

Two utilities will never have the same smart grid; therefore, employing a flexible methodology to develop the right architecture and components for each utility environment is a critical point. In addition, the ability to design the proper analytics for each unique need of each utility could have a profound impact on the data management architecture. The possibility of using tested reference models, tools, and processes can help utilities optimize the predictability of the outcomes.

When designing analytics, two major aspects have to be taken into consideration: time scales (latency) and volume scalability. Due to varying application requirements, some analytics must be available at high speed and with low latency (milliseconds), primarily at the level of grid sensors and devices. Others require seconds-to-minutes, including those for operational processes, such as operational efficiency verification, real-time optimisation (load balancing), and outage management, while still others may require hours, days, weeks, and even months. In order to incorporate varying levels of latency into the data management architecture properly, utilities should construct a data latency hierarchy (an example is provided in Figure 8); this approach allows to treat and analyse data differently on the basis of their latency and applicability, ranging from the lowest-latency data, where real-time technical analytics feed into protection and control system, to the highest latency where operational analytics can be integrated with business intelligence management dashboards and reporting.[20]

Large volumes of data associated with distributed assets can make centralised computation of analytics problematic. The implementation of distributed data management and analytics is a proven solution to this issue. Distributed Stream Processing (DSP) systems are valuable solutions for extracting meaningful insights from large streams of real-time data. Application areas include IoT data processing, network monitoring, fraud detection, with high throughput rates and low end-to-end processing latencies to support time-sensitive decisions. Input streams, however, are dynamic in nature and processing loads, therefore, have the potential to change significantly over time. Consequently, DSP systems should be able to scale horizontally across a cluster of commodity nodes, in order to accommodate different processing loads. Research interest into approaches for the adaptive

[20] Achieving High Performance in Smart Grid Data Management - Making sense of the data deluge, Accenture https://www.smartgrid.gov/files/documents/Achieving_High_Performance_in_Smart_Grid_Data_Management_201012.pdf

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

management of resources by dynamically adjusting configurations at runtime is growing in in recent times.[21]

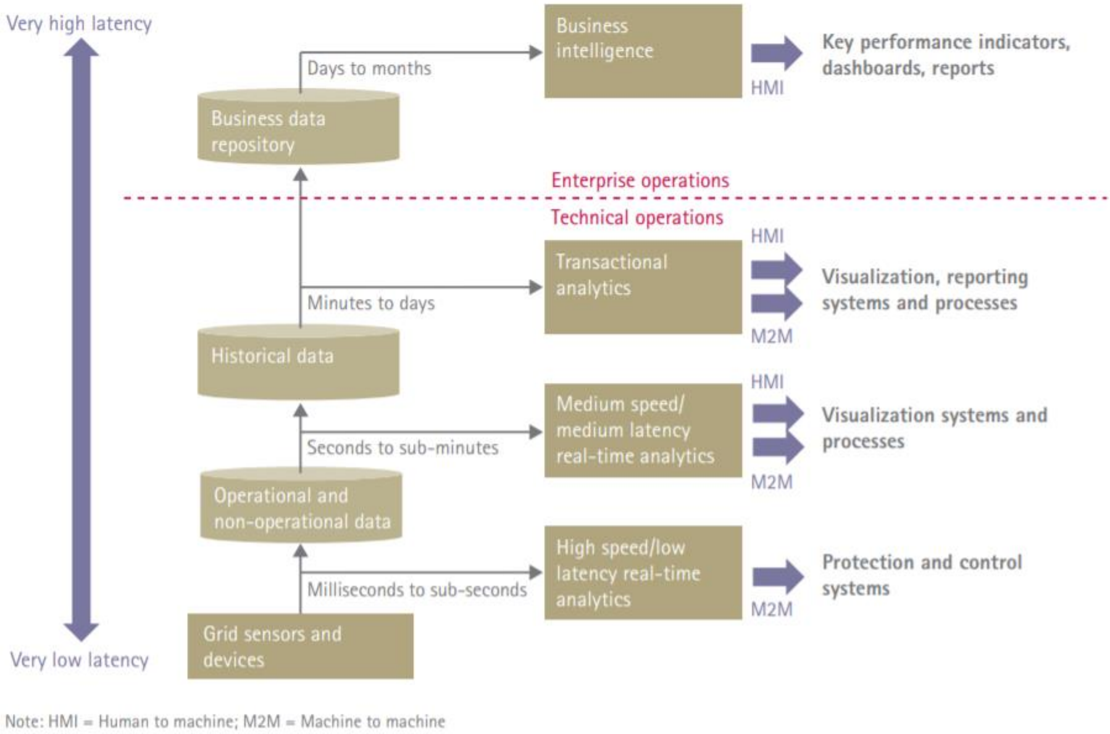

Figure 8: A data latency hierarchy for smart grid applications[20]

Cloud computing is a widespread approach adopted for delivering IT services that attracts the interest of both academic and industrial fields. Using advanced technologies, cloud computing provides end users with a variety of services, starting from hardware-level services to the application level. This means that computing services have moved from local data centers to hosted services that are offered over the Internet and paid for based on a pay-per-use model. In big data management, cloud computing plays an important role, since it provides organisations with the ability to analyse and store data economically and efficiently. A scalable HPC environment can handle the complexity and the challenges related to big data. For example, performing HPC in the cloud was introduced when data have started to be migrated and managed in the cloud. Performing HPC in the cloud is known as high-performance computing as a service (HPCaaS). Many solutions have been proposed and developed with the aim to improve computation performance of big data. Some of them tend to improve algorithm efficiency, provide new distributed paradigms, or develop powerful clustering environments.[22]

The increasing number of IoT applications often demand a low end-to-end latency between a sensor and a control center. These delay-sensitive applications often have stringent task service delay requirements, that is the total delay from the moment at which the task enters the system to when

[21] Gontarska, Kain & Geldenhuys, Morgan & Scheinert, Dominik & Wiesner, Philipp & Polze, Andreas & Thamsen, Lauritz. (2021). Evaluation of Load Prediction Techniques for Distributed Stream Processing.

[22] Achahbar, Ouidad & Abid, Mohamed & Bakhouya, M. & El Amrani, Chaker & Gaber, Jaafar & Essaidi, Mohamed & Ghazawi, T.A. (2015). Approaches for high-performance big data processing applications and challenges.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

the process is completed. The development of delay-sensitive IoT applications is presenting increasing challenges for the current cloud computing infrastructure. Not only the computation delay must be taken into consideration; the queuing delay and network delay must be taken into consideration as well. Even though cloud computing can provide a low-cost, easily expandable, and on-demand high-performance computation service, it relies on huge volumes of data transmitted from the IoT end devices to the remote cloud center, which can consume an extremely large amount of network bandwidth resources and can cause considerable network delay. For these reasons, cloud computing is becoming the bottleneck for the development of delay-sensitive IoT applications.

Fog computing has recently emerged as an extension of cloud computing that processes tasks at the edge devices of the network with the aim of preventing the large network delay, thus providing high-performance computing services for delay-sensitive IoT applications. By offloading tasks to a geographically proximal fog computing server instead of a remote cloud, the delay performance can be greatly improved. Nevertheless, the computation capability of a fog node is limited due to its geographical location and limited power. When the workload of the fog node is heavy, tasks have to be queued at the fog node, and may experience a long queuing delay, which in some cases may even exceed the network delay. Given the limited computation capability of a single fog node, it will be useful to explore and exploit cooperation models among multiple fog nodes and the cloud center to provide lower service delay. The cooperation of multiple closely connected fog nodes may significantly reduce the data transmission and network delay between the local fogs and the remote cloud center as well as the task queueing at a single fog node.[23]

While fog computing uses a centralised system that interacts with industrial gateways and embedded computer systems on a local area network, edge computing performs much of the processing on embedded computing platforms directly interfacing to sensors and controllers. It is powered by small form factor hardware with flash-storage arrays that provide highly optimised performance. Performing computations at the edge of the network reduces network traffic, thus reducing the risk of a data bottleneck. At its basic level, edge computing brings computation and data storage closer to the devices where data are gathered, rather than relying on a central location that can be thousands of kilometers away. This is done so that data, especially real-time data, do not suffer latency issues that can affect an application performance. In addition, companies can save money by having the processing done locally, reducing the amount of data that needs to be processed in a centralised or cloud-based location. Edge computing also improves security by encrypting data closer to the network core, while optimising data.

In BD4NRG, most of those approaches are being adopted. On the basis of pilot use cases and requirements, proper analytics have been defined and the correspondent expected latency and the necessary resources (both in terms of hardware and datasets) have been identified in order to individuate the most adequate layer of the edge/fog/cloud architecture for running each analytics. The distribution and parallelisation of data pipelines on the infrastructure nodes are being evaluated, trying to exploit as much as possible data locality. Data locality techniques allow bring computation to the

[23] Li, L.; Guo, M.; Ma, L.; Mao, H.; Guan, Q. Online Workload Allocation via Fog-Fog-Cloud Cooperation to Reduce IoT Task Service Delay. Sensors 2019, 19, 3830.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

data to optimize I/O speed, to comply with security rules (for example, in secure multi-party computation), and to avoid costly data migration (federated learning).

A large variety of technologies, algorithms, and optimisation techniques are being exploited, assessed, and implemented with the aim of efficiently enabling the training of complex models and the solving of computationally intensive problems. In the training phase, the type of the problem being solved (supervised or unsupervised learning/classification, regression, clustering), the size of the sample data, and the objectives of the algorithm (accuracy vs. efficiency) are taken into consideration for performing an incremental analysis and fine tune the model.

Moreover, different hyper parameters are examined for each one of the considered algorithms and the most successive ones will be adopted per case to maximise their potential and ensure that they are properly optimised for the particular training dataset. Distributed deep learning methods, allowing both data and model parallelism, will be implemented, such as data partitioning and ML models to distribute optimally the workload to multiple computational workers/nodes. In order to ensure high quality of services and automation, the solutions developed will involve self-learning elements, as well as real time evaluation frameworks and documentation, transforming black box models into actionable knowledge.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

4 Edge-Fog-Cloud Paradigms for Streaming and Batch Analytics

With the advent of IoT, connecting and controlling devices have given rise to smart grid technology, designed to provide robust and efficient energy management solutions lacking in the existing framework. The smart grid must be considered as a mission-critical asset that is part of an IoT framework. It can be used to remotely monitor and manage transmission lines, substations, cogeneration, outage sensors, early detection (e.g., power disturbances due to earthquakes and extreme weather). The smart grid does this through private, dedicated networks connecting devices that are distributed citywide, including smart meters, data concentrators, transformers, sensors. Smart grid IoT technologies enable two-way communication between connected devices and hardware that sense and respond to user demands. A smart grid is more resilient and less costly than the current power infrastructure. The overwhelming amount of data these devices produce poses a challenge for capturing, managing, processing, and analysing these data within a responsive acceptable time. This is a non-trivial process since a completely new IoT architecture is necessary that is also capable of performing streaming analytical tasks running in parallel to provide timely approximate and accurate results.

When selecting an IoT architecture, the most important issue is to balance the tradeoff between throughput and latency. Most approaches to handle this tradeoff have been based on a cloud computing environment where IoT data streams are pushed to and accumulated over a long period of time and are later processed and analysed in batches, despite the consequent issues, such as high latency, high data rates, low fault-tolerance. Batch-oriented processing frameworks have been used for processing large amounts of historical IoT data with high throughput but also with high latency. For example, the Hadoop MapReduce framework[24] is one of the most widely used cloud architectures for batch-oriented processing that supports distributed storage across many clusters of commodity servers. Apache Spark[25] is a multi-language engine for executing large-scale batch processing in memory on single-node machines or using distributed data sets in clusters.

Micro-batch frameworks (such as Spark Streaming) are capable to buffer and process IoT data streams in batch, thus increasing efficiency, but also increasing the time the data streams spend in the data pipeline. Stream-oriented frameworks (such as Storm and Flink), instead, typically provide time-sensitive computations and minimize the time a single tuple should spend in a processing pipeline, while they have relatively high data processing costs on a continuous stream of IoT data and minimise the time a single tuple spend in a processing pipeline.

From an analytics perspective, IoT data streams that are accumulated for a long period of time can be analysed in batches using traditional algorithms in ML and data mining such as clustering, classification, regression and dimensionality reduction, to name a few. As an example, a MapReduce based mining

[24] Dittrich, J.; Quiané-Ruiz, J.A. Efficient big data processing in Hadoop MapReduce. Proc. VLDB Endow. 2012, 5, 2014–2015.

[25] Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.; Franklin, M.J.; et al. Apache spark: A unified engine for big data processing. Commun. ACM 2016, 59, 56–65.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

algorithm has been proposed by Ismail et al.[26] to facilitate parallel productive periodic frequent pattern mining of health sensors data.[27]

In more recent times, when it comes to analytics, a new paradigm shift has emerged in the evolution of IoT architectures. Streaming analytics algorithms are being implemented to extract value from IoT data streams as soon as they arrive at a computational resource. In order to address to operate on continuously generated sensor data streams while addressing the specific Quality of Service (QoS) requirements, including low communication latency, fast computations, and high bandwidth, the centralised cloud computing paradigm has evolved towards new promising distributed edge and fog computing trends. The new edge/fog computing model aims to optimise smart applications by exploiting their functionalities closer to the geographic location of the IoT devices, rather than moving data to far away datacenters for computations. By decomposing the processing and analytical capabilities into a network of streaming tasks and distributing them into an edge-fog-cloud computing environment, it is possible to support streaming descriptive, diagnostic and predictive analytics efficiently.

The common approaches to the definition of a suitable edge or fog computing in the literature have been categorised into three groups: (i) studies that use a static amount of edge and fog resources as a complete replacement to cloud infrastructures; (ii) studies that intend to discover available edge and fog computing resources during runtime, when the workload increases and elasticity must be achieved; and (iii) replicating services in all edge devices, fog nodes and cloud resources to address reliability and resilience problems[28].

In an advanced edge/fog computing framework, new software engineering approaches can be adopted by exploiting lightweight Microservices[29] packaged into containers (managed and orchestrated through technologies like Docker, Kubernetes, OpenShift Origin, and Swarm), which allow to achieve a high degree of automation, deployment, elasticity, and reconfiguration of smart IoT applications at runtime. In that sense, an interesting example is the Analytics Everywhere Framework[23], which integrates a variety of computational resources for a flexible orchestration platform that exploit container functionalities. The primary goal of this framework is a seamless and automated execution of automated analytical tasks founded on a data life-cycle of an IoT application.

When designing analytics framework, three major factors have to be taken into account: Resource Capability, Analytical Capability and a Data Life-cycle.

Resource Capability In general, an IoT application requires a combination of different compute nodes running at the edge, fog and/or cloud and the main criteria to take into account when selecting a compute node can be:

- the geographical proximity of compute nodes to an IoT device, in which case edge nodes should be located near the IoT devices and should use short-range networks for the data streams (vicinity);

[26] Ismail, W.N.; Hassan, M.M.; Alsalamah, H.A. Mining of productive periodic-frequent patterns for IoT data analytics. Future Gener. Comput. Syst. 2018, 88, 512–523.

[27] Cao, H.; Wachowicz, M. An Edge-Fog-Cloud Architecture of Streaming Analytics for Internet of Things Applications. Sensors 2019, 19, 3594.

[28] Taherizadeh, Salman & Stankovski, Vlado & Grobelnik, Marko. (2018). A Capillary Computing Architecture for Dynamic Internet of Things: Orchestration of Microservices from Edge Devices to Fog and Cloud Providers. Sensors. 18. 10.3390/s18092938.

[29] Jarwar, M.A.; Kibria, M.G.; Ali, S.; Chong, I. Microservices in Web Objects Enabled IoT Environment for Enhancing Reusability. Sensors 2018, 18, 352.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

- the time to reach a compute node via a network (reachability);
- the retention time of IoT data streams, available memory size (memory resources);
- the amount of processing power available at a compute node for performing a set of automated analytical tasks (computation resources);
- bandwidth and the throughput required by the IoT application. The actual amount of data that have been transmitted varies, as there could be many different factors (i.e., latency affecting throughput). The latency is clearly low at the edge due to the proximity to the IoT devices and increases as we move to the cloud.

While computation and memory capabilities can increment as the analytical tasks are executing from the edge to the cloud, reachability must be always considered for an analytical task. Reachability is a critical dimension that requires analytical tasks to return results in a timely way, dependently of computational resources.

Analytical Capability Streaming analytics are used to extract value from IoT data streams at the edge, the fog or the cloud. The aim is to generate new insights as demanded by an IoT application in order to answer the questions:

- "What is happening in the real-world?" (Streaming Descriptive Analytics);
- "Why is it happening?" (Streaming Diagnostic Analytics);
- "What will happen?" (Streaming Predictive Analytics).

The main goal of the Analytics Everywhere model is to automate a priori known analytical tasks that will be executed at the edge, the fog, and the cloud in order to answer these questions.

The analytical tasks can have different levels of complexity and can support multiple paths of computation ranging from data cleaning and data aggregation tasks that require a continuous stream of data, to more complex tasks, for example data contextualisation and data summarisation tasks that require accumulated data streams for time-sensitive results.

Streaming descriptive analytics may be performed at the edge, the fog and the cloud, although they are often executed at the edge because IoT data streams have tiny volumes at the edge and many IoT applications are prevented data from being moved to a cloud due to privacy and cost concerns.

Streaming diagnostic analytics can be executed near to or far from an IoT device, depending on where it is possible to install the necessary computational resources. Streaming diagnostic analytical tasks are usually supported by a few on-line algorithms, stream clustering algorithms, ad-hoc queries, and continuous queries. Fog and cloud resources are used to perform streaming diagnostic analytics, since they provide computation, storage and accelerator resources that are more suitable than edge nodes to perform the streaming tasks. Fog and cloud computing can improve the accuracy and reduce the computational complexity of the automated tasks in near real-time.

Streaming predictive analytics requires on-demand analytical tasks with high availability and rapid elasticity through the virtually unlimited resources of the cloud; predictive analytical tasks use a huge amount of historical IoT data that need to be processed according to the nature of IoT applications.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Data Life-Cycle A data life-cycle can be either stateful or stateless depending on the requirements of the application. A stateless data life-cycle treats each analytical task independently and creates output data tuples depending only on the input data tuples of that analytical task. On the contrary, stateful data life-cycles combine different analytical tasks together and create the output data tuples based on multiple input data tuples taken from those analytical tasks. Moreover, data can be managed in the following way:

- At most once: there is no guarantee that data tuples in a stream are handled at most once by a streaming task and in case of failure at the edge, fog, or cloud nodes, no additional attempts are made to re-handle these data tuples.
- At least once: the data tuples in a stream are guaranteed to be handled at least once by all streaming tasks and if a failure takes place, additional attempts are made to re-handle these data tuples.
- Exactly once: exactly once means that data tuples are guaranteed to be handled exactly the same as they would be, also in the event of various failures.

The edge-fog-cloud continuum brings a high complexity in connecting and orchestrating designing analytics framework and the right compromise among these factors have to be found according to the framework requirements. Analytics Everywhere framework, for example, supports a stateless data life-cycle having the “at most once” approach for guaranteeing reliability and low latency for running analytical tasks of an IoT application.

The increasing momentum of big data technologies and the growing trend towards a data economy, which leverages on exploiting the untapped economic value of legacy data assets, represents an unprecedented opportunity for the Electric Power and Energy Systems (EPES) to gain significant improvements in the reliability and efficiency of electricity grid operation and planning, enable a more efficient management of the grid-owned and behind-the-meter assets connected to the grid and, at the same time, a more reliable assessment of energy efficiency and renewable energy financial investments, while taking into due consideration operational performance monitoring and comfort of citizens. BD4NRG aims at addressing these emerging challenges in big data management for energy with an open holistic solution aimed at creating a European approach to the Business-to-Business platforms able to give a competitive edge to the European stakeholders operating in this sector and to open new market opportunities.

BD4NRG vision is to unlock and exploit the economic potential of big data to enable improved operation for all the stakeholders in the energy value chain, through creating and deploying an elastic energy analytics reference framework, which includes a high-velocity flexible elastic data capture and management optimised stack coupled with effective near real time cross-domain cross-stakeholder reusable data analytics specifically tailored to the extreme-scale smart energy grids environments. In order to enable and realise the proposed vision, BD4NRG intends to deliver a reference architecture (RA) for development and piloting of big data edge analytics for smart energy grids and its seamless integration with operational framework already in place in smart grids.

The BD4NRG ecosystem envisions a number of distributed collaborative federated nodes, which may act as data providers and/or data consumers, deployed over a variety energy and non-energy data sets and/or legacy energy interoperable platforms from smart grid assets, technologies and components,

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

hence representing sources of data owned/managed by distribution system operators (DSOs), transmission system operators (TSOs), market operators, energy service companies (ESCOs), aggregators stakeholders. At each node, the following functional blocks will be implemented for the smart management and processing of data collected:

- Polyglot persistence, which will be in charge for intelligently managing distributed edge-level storage of those data which have been incorporated within the big data processing stack. Such component may leverage on the services offered by the BD4NRG Marketplace to select a number of edge resources physically located nearby data generation point.
- Elastic Data Capturing & Streaming, which aims to manage dynamically the frequency rate of the data streaming for the subsequent in-memory processing of the near real time data. Such component could be tightly connected to the performance of some analytics. For example, the sampling rate could be increased depending from an event-based detection (descriptive analytics) or anticipated prediction (predictive analytics).
- Edge Data Stream Parallelisation, will be in charge of managing and coordinating intelligently the distribution and the parallelisation of data streams along a variety of available computing nodes for both cloud and edge applications. A special attention will be given to data volume, velocity, trajectory and changing variability of data streams.
- Edge/Fog Resource/Service Management is aimed to incorporate the capability of orchestrating the deployment and operation of edge data management micro-services on a virtualised set of locally available computing resources at edge nodes.
- Big Data Pipelining will be in charge for coordinating and optimising the access and the orchestration from the technical perspective to the big data technologies and in particular optimise access to computing and storage resource, which may be provided by the underlying Edge Framework and/or by the cloud platform-as-a-service (PaaS) as offered by EGI Foundation.
- AI-based ML will be engaged to move to a rule-based matching of data to enable Big Data handling, as the appropriate scientific discipline that explores the construction and study of algorithms that can learn from data such algorithms operate by building a model from example inputs and using that to make predictions or decisions in order to exploit hidden opportunities in big data. Since one of the major problems preventing BDA to be fully exploited is the lack of a critical mass of data sets due to strong regulatory constraints and the awareness of the electricity grid as a critical infrastructure, vital for our everyday life, we will leverage on Privacy-preserving Federated Learning where rather data models will be shared at level of a subset of training models, hence bringing to a significant reduction of communication costs and fully preserving the data privacy concerns. In such a way BD4NRG will demonstrate the Edge AI. Moreover, whereas a partial data sharing is allowed or data providers are willing to share their own data, combining unsupervised deep learning/reinforcement learning models for cross-context transfer learning (e.g., cross-context AI by design) with inline adaptive self-learning will enable to improve the resulting ML models. A library of reusable AI-based ML models will be made available through the Marketplace with a view to promote quick adaptation and reuse of ML models along different contexts.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

- The semantic enrichment will facilitate the understanding of the sources of knowledge, as well as the knowledge inputs and outputs, which will be stored, communicated, and used to facilitate ML, decision support as well as functionalities of the upper layers in the BD4NRG environment.

BD4NRG will also release a BDA Energy Toolbox, allowing energy stakeholders to design, develop, and run custom BDA applications by leveraging on reusable tools, including analytics applications and ML models/algorithms, made available by BD4NRG ecosystem members. The BD4NRG Toolbox could be deployed according to different models:

- On premise as a sandbox addressing the specific case of energy stakeholders willing to integrate such toolbox with their already existing systems. In this case, the pre-existing available local-level capabilities for sensing, gathering, integrating, and processing data through available in-house analytics platforms can be augmented with the capabilities of the proposed analytics reference toolbox.
- As-a-service, where the BD4NRG Toolbox will be operated and performed in a third-party cloud-based environment and will be made available in a servitised paradigm via interoperable interfaces to energy stakeholder users. Such option is mostly attractive for small-medium energy stakeholders, which do not have the necessary resources to internalize BD4NRG adaptation, integration and operation.

4.1 Architecture

When designing an IoT architecture that relies on the edge-fog-cloud continuum for running automated analytical tasks that have a data life-cycle with streaming data tuples as input and output, resource capabilities play an important role.

The spreading DevOps movement aims to make notable progresses in the software lifecycle. Instead of delivering monolithic, stand-alone applications, modern software engineering workbenches leverage on small, reusable, elastic, interdependent, and well-tested software components, which are built as microservices, where each service is responsible for its own small objective. Figure 9 compares the microservices architecture to the traditional monolithic design. In the monolithic design, all functional logic for handling requests runs within a single process, which implies that even a very small update made to a small segment of the application requires the entire monolith to be re-developed and then re-deployed again. In the Microservices architecture, each business capability is a self-contained service, for example, with a REST application programming interface (API); microservice running on a specific host can be easily managed at runtime that includes various operations such as deployment, instantiation, movement, replication, termination, etc.

Today organizations are using containerisation increasingly to create new applications, and to modernise existing applications for the cloud. Containerisation is the packaging of software code with just the operating system libraries and dependencies required to run the code to create a single lightweight executable—called a container—that runs consistently on any infrastructure. More portable and resource-efficient than virtual machines (VMs), containers have become the de facto compute units of modern cloud-native applications. From the packaging and deployment point of view, compared to virtual machines (VMs), the containerisation of microservices does not require an operating system to boot up. Moreover, resource usage of virtual machines (VMs) is extensive, and

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

thus typically, they cannot be easily developed on small servers or resource-constrained devices, such as Raspberry Pis or routers. On the contrary, containers can be used easily in such frameworks. Since their nature is lightweight, deployment of container-based services at runtime can be achieved faster than VMs.

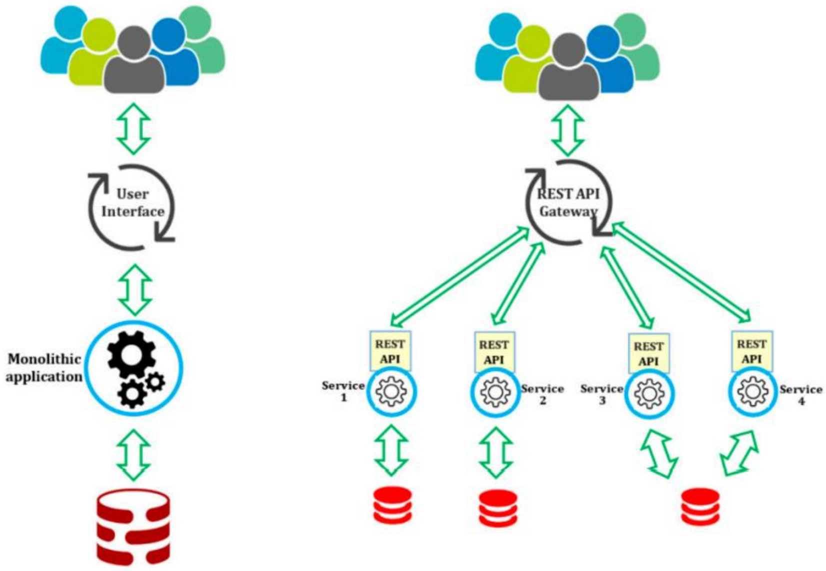

Figure 9: Monolithic architecture versus Microservices architecture[28]

Existing edge and fog computing architectures aim at achieving high quality of service (QoS) operation. Smart IoT applications utilise various streaming sensor data generally and require significant processing capacity for AI algorithms. When designing suitable edge and fog computing architectures, different aspects need to be taken into account, for example the dynamic nature of IoT devices (e.g., whether they are moving or static), big data velocity, veracity, variety, and volume, computational complexity of data analytics, etc. Furthermore, the workload can significantly change depending on various events in the runtime environment and this requires new elasticity mechanisms to be used at the Edge of the network.

Existing approaches to the definition of a suitable edge or fog computing can be categorised into three groups: (i) approaches that use a static amount of edge and fog resources as a complete replacement to cloud infrastructures; (ii) approaches that intend to discover available edge and fog computing resources during runtime, when the workload increases and elasticity must be achieved; and (iii) replicating services in all edge devices, fog nodes and cloud resources to address reliability and resilience problems.[28]

As an example, Cao et al.[27] propose a geographically distributed network of compute nodes that have a combination of modules including Admin/Control, Stream Processing & Analytics, Run Time, Provision & Orchestration, and Security & Governance (Figure 10). This IoT architecture enables microservices to run at various compute edge/fog/cloud nodes, so that each microservice can perform a specific analytical task depending on which module it belongs to. The Admin/Control module optimises the data flow in order to implement a data life-cycle that takes into account the individual requirements of an IoT application. Data management, visualisation, orchestration, and security modules have included in the architecture as well.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

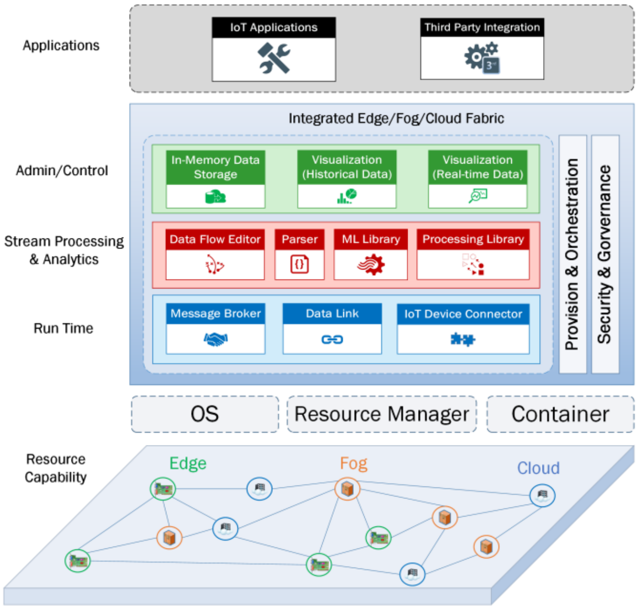

Figure 10: Internet of Things (IoT) architecture for Analytics Everywhere framework[27]

In case of limited amount of available resources, such as storage capacity or computing power on a node (e.g., an edge node), the running microservice could be offloaded from that node to another node (e.g., a fog node). This approach has been adopted, for example, by Taherizadeh et al.[28], in their architecture called Monitor-Analyze-Plan-Execute over a shared Knowledge (MAPE-K) reference model (Figure 11), capable of providing the necessary elasticity management of an application in situations where an edge node is going to be overloaded. In comparison with centralised cloud, the fog computing model is able to decrease the amount of traffic in the network, and offer low-latency response time for the service. According to this model, if a fog node is overloaded due to a drastic increase in the workload, mobility of microservices from the fog node to the cloud also needs to be considered during the execution time.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

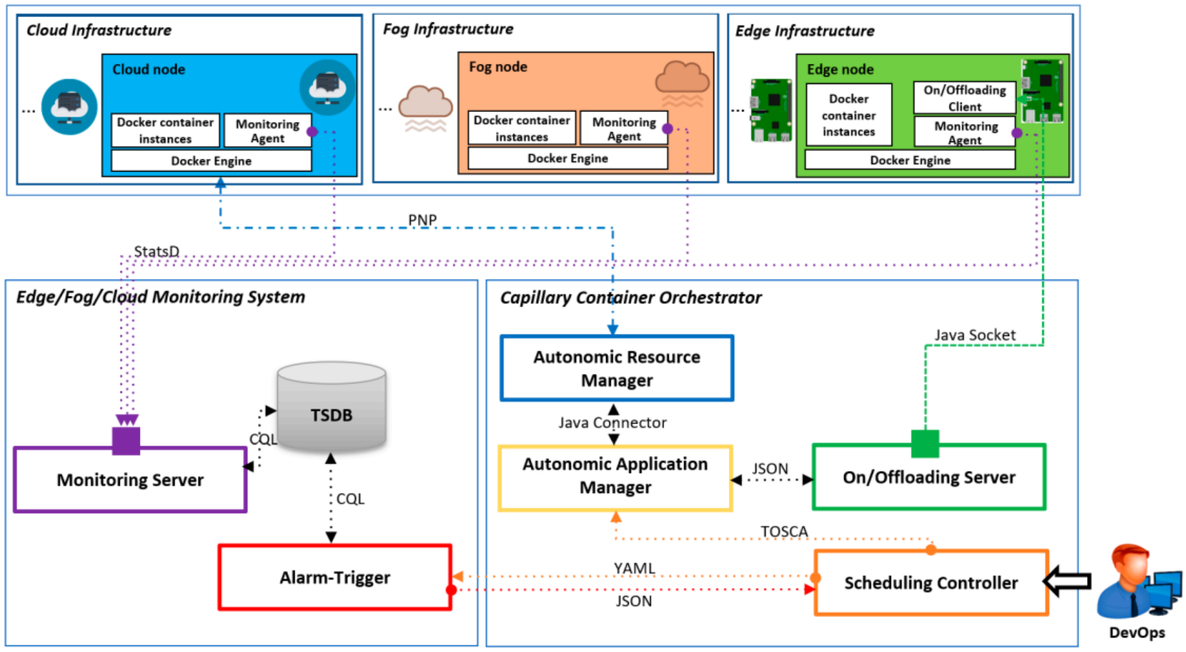

Figure 11: Capillary distributed computing architecture for smart IoT applications[28]

Figure 12 depicts the BD4NRG big data system architecture designed in the context of WP 3 to serve as a masterplan of the data location, data description, data flows and data availability. It is intended to deliver a conceptual infrastructure to support data quality, data stewardship, data integration, data provisioning as well as system collaboration among applications and data repositories. Specifically, it shows how the data flows from the large scale pilot (LSP) raw datasets, through the various layers of transformation (Interoperability & Homogenisation Functions) and quality checks (Data Quality & Compliance) to obtain the Golden Data Source, through data integration tasks (Dynamic Data Provider Functions), through operational data store (HTAP), all the way to the business intelligence, analytics, and decision support applications that use the Integrated Query Engine for querying optimised data structures for data processing. The architecture is organised in three vertical layers, which cover specific aspects along the data processing chain, and two horizontal layers, which are responsible for cross-cutting issues that have an impact on the vertical layers. Everything is built upon a common infrastructure. Similarly, to the Big Data Value Reference Model the vertical layers do not imply the adoption of a layered architecture model since data can be served to consumers/client applications without being processed.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

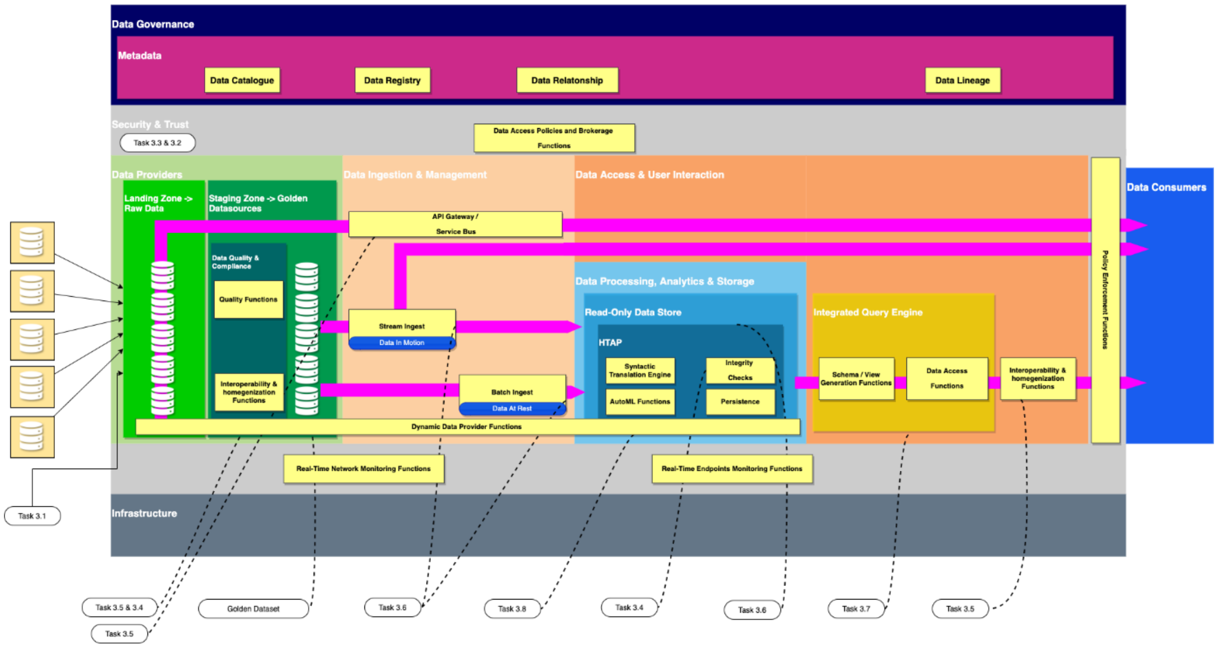

Figure 12: BD4NRG Data System[30]

Data Ingestion & Management is the layer where the data starts their journey, that is the place where data are delivered and distributed within the BD4NRG data system for immediate use or storage. Two main mechanisms are considered, namely: real-time streaming ingestion (for data in motion) and batch ingestion (for data at rest). The specific data ingestion mechanism strictly depends on the application use case.

Data Access & User Interaction layer facilitates access and navigation of the data, allowing to the generation of different business views to be represented by the data. A high-performance real-time analytics database lies at the center of this layer for supporting interactive dashboards and analytics platforms while being used as data warehouse. A query engine is also part of the layer and serves as rapid application development (RAD) to present fast logical views of data to consumer applications.

Before processing any data, the data have to be retrieved from a data source. This is a functionality provided by the Dynamic Data Provider, which interacts with the data source specifying what data are required.

Figure 13: Functional Design of Dynamic Data provider[30] depicts the functional design of the Dynamic Data Provider and the information flow and interactions with the full data pipeline. The envisioned adaptation engine is focused on tackling changes in data.

[30] BD4NRG D3.1: BD4NRG Governance – 1st Technology Release accompanying report

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

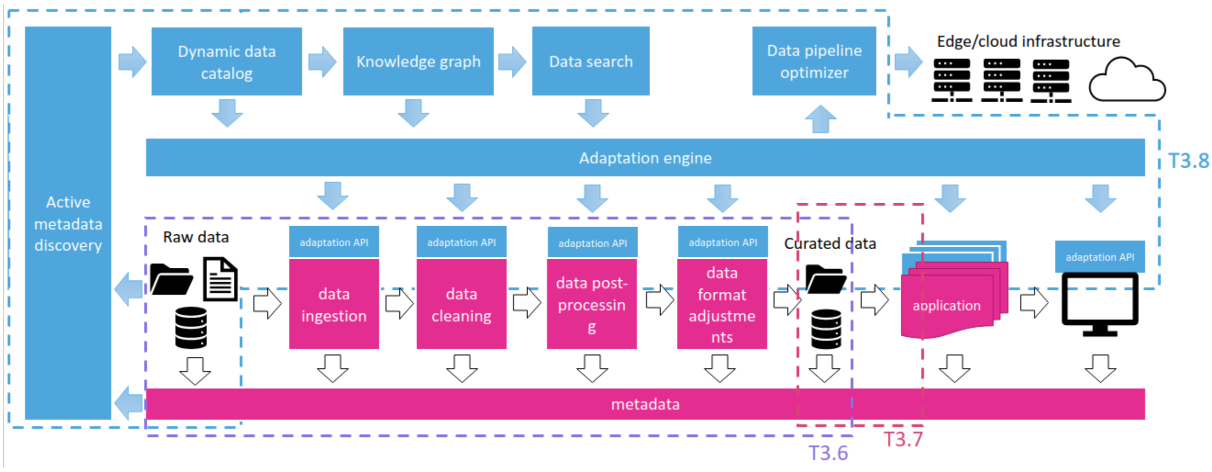

Figure 13: Functional Design of Dynamic Data provider[30]

The process of making curated data in required format, of necessary quality, and guaranteeing data integrity includes the following stages:

- Data ingestion: the mechanism to load the raw data coming from the data sources;
- Data cleaning: removing the noise from the data;
- Data postprocessing: filling in the gaps within data, filtering it, compressing, calculating means, averages and deviations, etc.;
- Data format adjustments: putting data into the required format.

After going through all these stages data becomes curated, ready for use by the applications which need these data. Curated data are stored in the database and can be pulled by the applications at any time.

Adaptation actions could happen at any of these stages. The adaptation engine should control the actions and the introduction of new models into all the phases of the transformation of data from raw to curated and its use in application.

Part of the activity of WP3 "Data Governance and Management" is the investigation of how data locality and the ability to move the computation closer to where the data reside, can improve the overall throughput of the system and minimizes network congestion. Data locality at the same time allows analysis of commercial and confidential data without sharing. The functionality that is being implemented is the ability to distribute data pipelines on the computation nodes, trying to exploit as much as possible data locality. For this purpose, the module needs to have access to the following information:

- available computation nodes on the infrastructure, possibly linked with data locality information (e.g., if a computational cluster is co-located in the same data center as the data we are trying to access);
- available datasets on the infrastructure (data catalogue), provided by one of the components to be implemented in the context of task T3.8 (Dynamic Data Catalogue).

Assuming that the computational nodes are running a container orchestrator (e.g., Kubernetes), the Data Pipeline Optimiser deploys and configures the needed data streaming technology according to

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

the pipeline to be executed. The component will provide an API for the Adaptation Engine component to interact with, in order to describe the computation and the data to be used (Figure 14). Infrastructure resource discovery needs to be implemented in order to understand the available computation nodes and if the given computation nodes is co-located with the data source, to explore data locality.

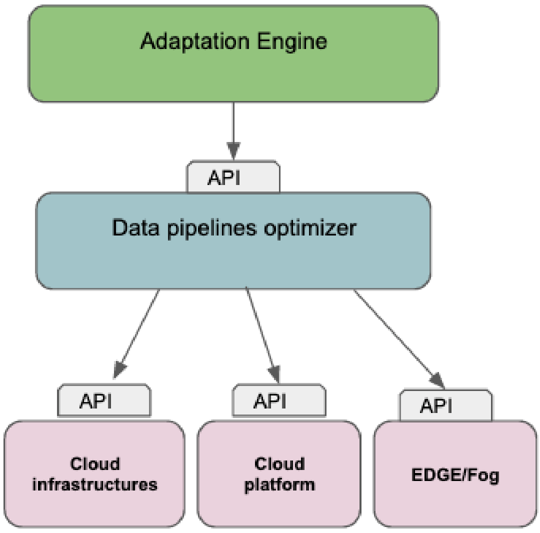

Figure 14: Data Pipeline Optimizer[30]

BD4NRG aims to develop data analytics services capable to analyse data in a seamless and holistic fashion, across multiple data sources. Among the existing efforts analysed, ALIDA BDA platform has been selected as a starting point for the BD4NRG Analytics Toolbox and it will be extended and validated in the energy domain, according to the needs that emerged in the requirements elicitation phase of the project. ALIDA embraces the paradigm of BDA-as-a-service (BDAaaS). Besides relevant cost savings, the provision of Big Data analytical capabilities using the cloud delivery model could ease the adoption of the toolbox and could simplify useful insights with different kinds of competitive advantage. BDAaaS consists of as a set of automatic tools and methodologies that allows users lacking Big Data expertise to manage BDA and deploy big data pipeline applications ready-to-be-executed addressing their goals at edge/fog/cloud nodes. The ALIDA platform supports full orchestration of BDA application workflows and allows for composition, deployment and execution of workflows (either batch or stream) of BDA applications. Figure 15 describes the ALIDA general architecture.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

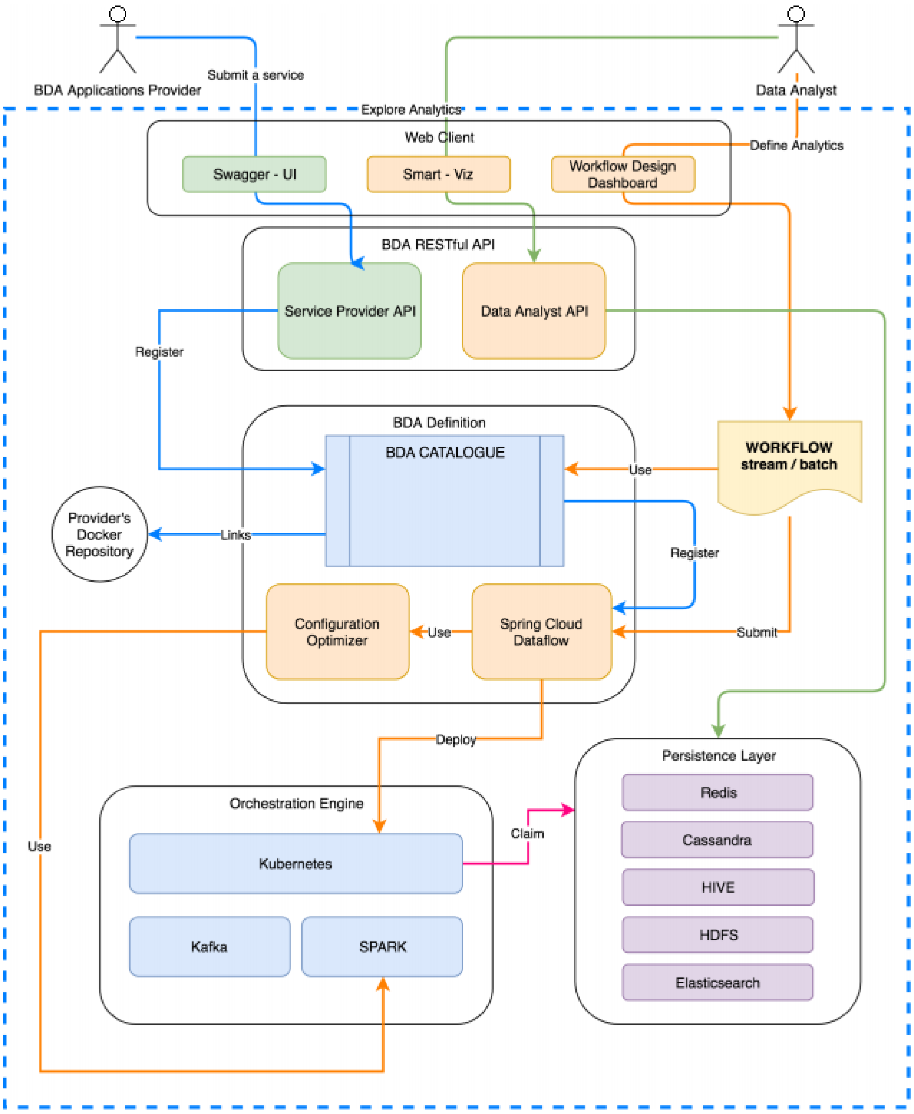

Figure 15: ALIDA general architecture[31]

BD4NRG will also offer a Marketplace, as a "virtual workbench", to incorporate a variety of assets, including data, third party services, ML models, computing resources, storage resources as tradable assets (Figure 16). The Marketplace enables and supports the BD4NRG ecosystem building and scaling up. A library of reusable AI-based ML models will be made available through the Marketplace with a view to promote quick adaptation and reuse of ML models along different contexts. ML models will be offered as a service (MLaaS). Furthermore, the Marketplace will offer access to simulated data testbeds and other external off-domain data sets (e.g., weather, geographical imagery) via third party services (Data-as-a-Service, DaaS).

[31] BD4NRG Deliverable 2.5: BD4NRG Reference Architecture First Version

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

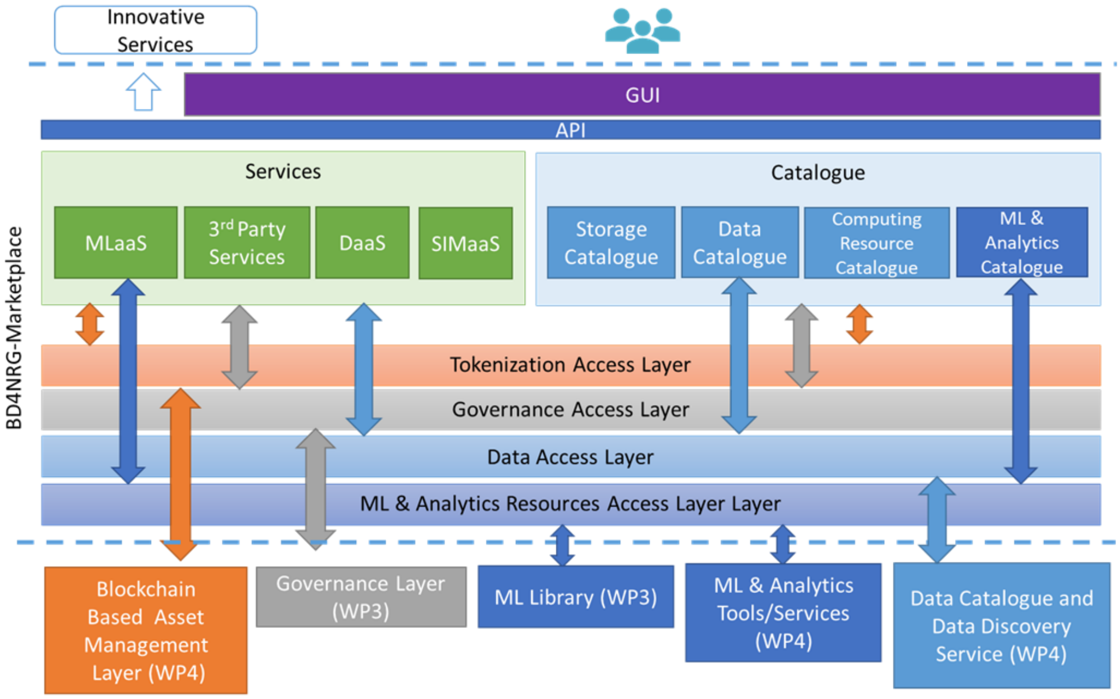

Figure 16: BD4NRG Marketplace architecture

4.2 Data Analytics and Artificial Intelligence

Data analytics is the science of analysing raw data to generate valuable insights and conclusions. Many of the techniques and processes of data analytics have been transformed into automated processes and algorithms that work over raw data to extract value from them.

Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information and, thus, can be used to optimize processes to increase the overall efficiency of a business or system. Implementing analytics into the business model can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data. A company can also use data analytics to make better business decisions and help analyse customer trends and satisfaction, which can lead to new products and services.

The process of data analysis involves several different steps:

- The first step is to determine the data requirements or how the data are grouped. Data may be separated by age, demographic, income, or gender. Data values may be numerical or be divided by category.
- The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.
- Once data are collected, it must be organised so that it can be analysed. This may take place on a spreadsheet or other form of software that can take statistical data.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

- Data are then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analysed.

Data analytics is broken down into four basic types.

- Descriptive analytics: This describes what has happened over a given period of time. Have the number of views gone up? Are sales stronger this month than last?
- Diagnostic analytics: This focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing. Did the weather affect beer sales? Did that latest marketing campaign impact sales?
- Predictive analytics: This moves to what is likely going to happen in the near term. What happened to sales the last time we had a hot summer? How many weather models predict a hot summer this year?
- Prescriptive analytics: This suggests a course of action. If the likelihood of a hot summer is measured as an average of these five weather models is above 58%, we should add an evening shift to the brewery and rent an additional tank to increase output. [32]

The data models, typical of data analytics, are often static and show limits in cases of a continuous change of the data structure. When it is necessary to correlate a lot of information, such as in the IoT, that produce millions of data, and, then, it is necessary to implement complex ML models. ML is useful when we know what we want, but does not know what all the important input variables are for making decisions. ML uses the goal imposed in the algorithm to "learn" from the data and choose which factors are important to achieve the goal.

Model training is the process by which a ML program is trained to perform the specific task at hand. The goal is to achieve the best possible performance through model training, to the point where the end result can mimic human-level performance in its specific task, or even surpass it. Most commonly, this means synthesizing useful concepts from historical data. In ML, a common task is the study and construction of algorithms that can learn from and make predictions or decision on data, through building a mathematical model from input data. These input data used to build the model are usually divided in multiple data sets. In particular, three data sets are commonly used in different stages of the creation of the model: training, validation and test sets.

The model is initially fit on a training data set, which is a set of examples used to fit the parameters (e.g., weights of connections between neurons in artificial neural networks) of the model. The model (e.g., a naive Bayes classifier) is trained on the training data, which often consists of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), where the key answer is commonly denoted as the target (or label). The current model is run with the training data set and produces a result, which is then compared with the target, for each input vector in the training data set. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation.

[32] https://www.investopedia.com/terms/d/data-analytics.asp

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Successively, the fitted model is used to predict the responses for the observations in a second data set, called "the validation data set". The validation data set provides an unbiased evaluation of a model fit on the training data set while tuning the model hyperparameters.

Finally, the test data set is a data set used to provide an unbiased evaluation of a final model fit on the training data set. If the data in the test data set has never been used in training (for example in cross-validation), the test data set is also called a "holdout data set".

BD4NRG intends to provide the necessary capabilities for pushing forward big data management and processing towards the edge, with a view to reduce some of the major barriers actually hampering big data and analytics full scale deployment in the smart energy grid domain, for example the excessive communication costs and the need for keeping generated data as much as possible closer to the generation point and the owner. In that respect, a number of technology enablers will be deployed to support:

- Edge-level AI-based learning, including privacy preserving federated learning on semantically structured data streams for mitigating the risk to share sensitive data, cross-context transfer learning combined with self-learning to promote the reuse of ML models in contexts with few or poor quality of data;
- Scalable big data management at the edge, including an IoT/edge/fog nodes distribution and coordination management, policy enforcement mechanisms for dimensionality reduction management and elastic in-memory data streaming management.
- Incremental (on-line) learning, in case of large volume of datasets that do not fit the memory, or in case of data availability issues (data coming in mini-batches).

The resulting BD4NRG Open modular Energy Analytics Toolbox will made available a huge variety of analytics ML algorithms/models to be selected among correlation, categorisation, and feature extraction to implement a variety of analytics types (descriptive, predictive, prescriptive).

4.2.1 Incremental and Self-learning Analytics

The majority of AI application are based on the traditional **batch learning** (or offline learning) method. Batch learning refers to the training of ML models in a batch manner, where the training dataset gets accumulated over a period of time and it is fed to the training module as a single batch. In this case, the model is unable to learn in a continuous way from a stream of observations becoming available gradually. Moreover, this approach is both time- and resource-consuming exploiting the available resources (CPU, memory space, etc.). The disadvantage of this approach is that when the pre-trained models need to be re-trained on new data, the models need to be fed the whole dataset (old and new data) in order to be reconfigured from scratch.

However, in the field of AI there are certain algorithms which facilitate the continuous extension of the model knowledge through continuous re-training of the model when new data are available. This approach is called **incremental learning** and it is widely applied in many cases when data are becoming available in smaller batches. The difference of incremental learning with the traditional learning approach is that through incremental learning the same model can be re-trained and incorporate new data, thus increasing its accuracy through the adoption of up-to-date information.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

Incremental learning is the most suitable approach in several problems. First and foremost, incremental learning is the best practice when the volume of the available data exceeds the system memory limits. In such cases, data are fed to the training module in an incremental way so as to fit to the system memory size and enable training based on the whole data set. Another wide application of incremental learning is when data become available in data streams (for example, through smart sensors or IoT devices). The continuous flow of data makes it necessary to train the model in an incremental way in order to incorporate the incoming streams. Generally, the reasons which necessitate the use of incremental learning are presented in the Figure 17.

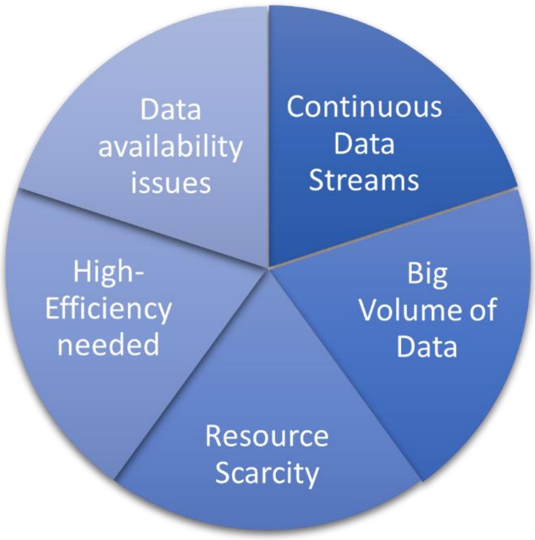

Figure 17: Problems dealt with the Incremental Learning Approach

Another widely utilised term in ML and deep learning is the concept of **online learning**, which is often confused or wrongly identified with incremental learning. Online learning refers to processing one data point at a time and updating the model directly, instead of processing the whole training set at once in order to establish the model. Therefore, in the online learning approach, the model is able to feed with new incoming data and get updated dynamically from the growing training set. In online learning, the model is continuously updated to incorporate new data and it is possible to forget previously learned attributes (catastrophic interference). On the other hand, in the incremental approach previous inferences are not always forgotten. The difference lies on the fact that online learning learns a model when the training instances arrive sequentially one by one, whereas incremental learning updates a model when a new batch of data instances arrive. Thus, online learning is always incremental learning, but incremental learning is not necessary to be online.

The process of online learning can be divided into three steps. Firstly, data must be streamed in smaller instances. Secondly, the features of these instances must be extracted through a pre-processing routine. Thirdly, these features must be fed to an incremental algorithm[33]. After these three steps the model can be used in the same way with traditional models. The first step includes the connection with a stream device that will generate the respective data in instances. This can be achieved through the

[33] https://scikit-learn.org/0.15/modules/scaling_strategies.html

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

creation of an online script that will read the output of this device in standard time steps and store data in a specific location. The second step, described as feature extraction involves the pre-processing of the data streams in order to transform them to the model input form. This might include several procedures such as cleaning the data from null values, vectorising data, performing standardisations or normalisations, and handling categorical variables.

Finally, new data must be fed to the incremental algorithm. The majority of incremental models incorporate a special function which is used for this purpose. In most documentations it is described as **partial fit** (or out-of-core fit). It is a variation of the traditional fit function, applying an incremental fit on small batches of instances. It should be called several times (each time on a different mini-batch of data), thus implementing out-of-core learning (Figure 18).

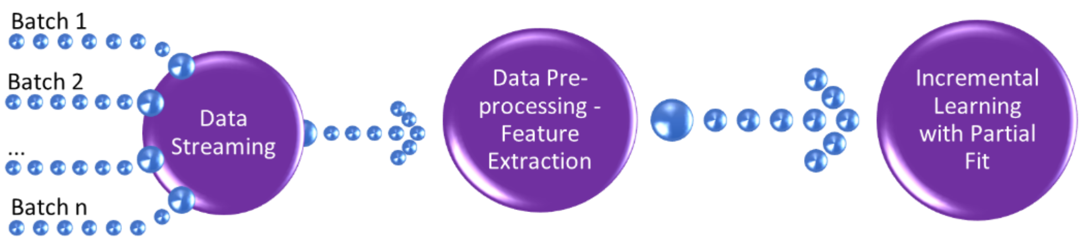

Figure 18: Incremental Learning Process

There are several incremental models which can deal with all the tasks that traditional ML algorithms treat, such as Classification, Regression and Clustering. Scikit-Learn, which is one of the most useful and robust libraries in Python, has developed a set of incremental algorithms for many different tasks, which are presented as follows:

- Classification Tasks
- Perceptron
- Multinomial & Bernoulli Naïve Bayes
- SGD Classifier
- Passive Aggressive
- Regression Tasks
- SGD Regressor
- Passive Aggressive
- Clustering Tasks
- Mini Batch K-Means

Except from Scikit Learn, there are many other libraries which support incremental learning. Creme[34] is a library tailored to implement Incremental Learning in Neural Networks. It supports learning the

[34] https://pypi.org/project/creme/

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

stream of data using different approaches and it allows models to train on one data point at a time. Xgboost[35] models (which provide an implementation of gradient boosting trees) can also be trained incrementally, saving the model after training of the first batch has been performed, and successively providing the model's train method to the new batches.

4.2.2 Federated Learning

Developments in the field of IoT, smart cities, smart infrastructures, smart industries, etc., result in ever-increasing amounts of data. Within these big data developments, there is now a lot of focus on creating large data sets, but little attention is paid to making targeted information from combinations of data sets from specific stakeholders. With the increase in available data and its size, the challenge of extracting meaningful information from combinations of data sources is increasing.

The current method for large-scale data processing is that all data are taken (copied) to one central data analysis infrastructure and processed there. Such an analysis platform can have a large-scale (distributed) character so that a lot of data can be processed simultaneously, but it is always based on the availability of all data in one central place at the start of the analysis.

However, ***Central processing is not always possible – data must remain local in many cases***.

Moving data to such a large-scale data analysis platform is not always possible and not always desirable for several reasons:

i. The amount of (streaming) data is simply too large to copy. Think of copying a petabyte of information via the network or live streaming all camera images in a city via 5G.
ii. Combining several of these sources implies infrastructure requirements that are not related to the analysis to be performed.
iii. Companies and parties do not always want to share data with everyone under the same conditions. Think of privacy legislation and purpose limitation, or company policy and non-competition clause. Laws also require that data may not be stored outside the Europe or United States. As a result, parties want to keep more control over their own data.

The types of questions are changing, and so are the requirements for the infrastructure. Setting up the infrastructure in a different way for these changing demands takes too much effort and time to make this cost-effective.

How is it possible to allow multiple data owners to train collaboratively and use a shared prediction model while keeping all the local training data private? - Yang et al.[36] describe how federated ML addresses this problem with novel solutions combining distributed ML, cryptography and security, and incentive mechanism design based on economic principles and game theory.

Federated learning aims to build a joint ML model based on the data located at multiple sites. A federated learning system may or may not involve a central coordinating computer depending on the application. In a setting involving a coordinator, the coordinator is a central aggregation server (a.k.a. the parameter server). The federated learning architecture can also be designed in a peer-to-peer

[35] https://xgboost.readthedocs.io/en/stable/

[36] Yang, Q. et al. Federated Learning (2019).

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

manner, which does not require a coordinator[37]. To summarize, federated learning can work on top of different topologies adding layers of security and some established platforms.

Federated learning was first implemented on mobile phones, enabling them to collaboratively learn a shared prediction model while keeping all the training data on the device, decoupling the ability to do ML from the need to store the data in the cloud. This goes beyond the use of local models that make predictions on mobile devices by bringing model training to the device as well. This first implementation works like this: a device downloads the current model, improves it by learning from data on a phone, and then summarizes the changes as a small focused update. Only this update to the model is sent to the cloud, using encrypted communication, where it is immediately averaged with other user updates to improve the shared model. All the training data remain on the device, and no individual updates are stored in the cloud[38].

This implementation showed the potential of federated learning to achieve the same and even better results than centralised solutions. In some Google posts, how a federated model can even pursue better results than other solutions due to the capacity of aggregating predictions that were not accessible before is explained. Mobile devices may be the first implementation, but the concept, architecture, and even tools are able to work at organisational levels.

All types of analytics can be placed in computing infrastructure based on federated learning: from simple to complex models, from AI data-driven models to physics-based models. It depends from the use-case, which specific learning technique is chosen to be suitable for that specific situation.

A solution that can communicate multiple organisations at large scale while processing requests, pulling, and running models is a challenge considered by a specific type of architectures, such as federated learning architecture.

The generic federated learning architecture allows the separation of local personalised models on the premises where data reside, and central model, which receives the input from the local models, studies from them, and sends the update to the personalised models in a case of need (Figure 19).

[37] Yang, Q. et al. Federated Learning (2019).

[38] Zhao, J., Xie, X., Xu, X. & Sun, S. Multi-view learning overview: Recent progress and new challenges. Inf. Fusion 38, 43– 54, DOI: https://doi.org/10.1016/j.inffus.2017.02.007 (2017).

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

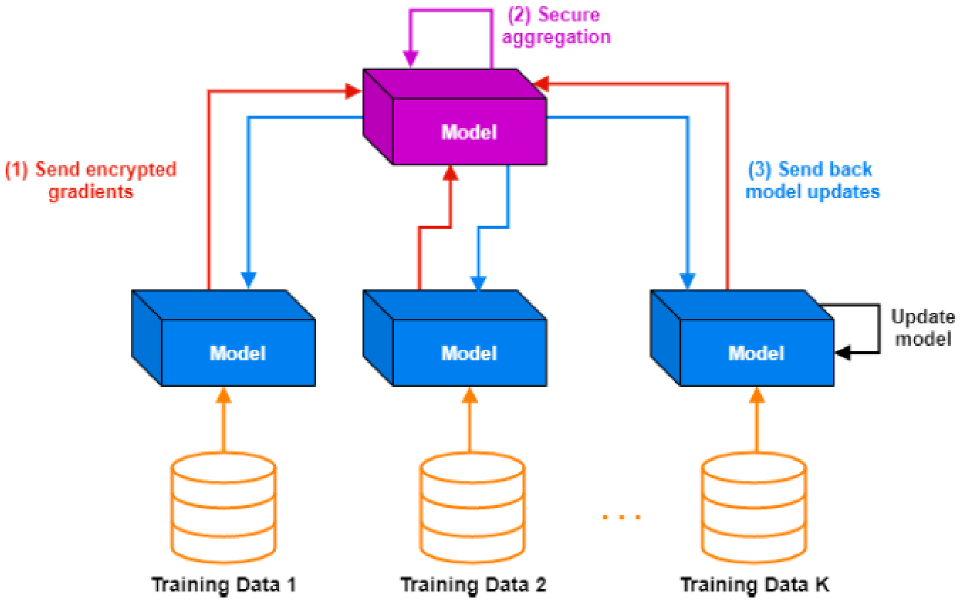

Figure 19: Generic architecture of Federated learning local models

The models in blue color are local models which run on the devices where data are produced, and violet model is a centralised model which receives the status updates from different local models in secure mode. That model trains itself on the basis of multiple inputs, and sends the update back to all local personalised models. That exchange happens in a secure manner.

This architecture is used in BD4NRG to solve the data analysis problem in the case when data should remain in the local premises of the data owner party (including sensitive data), and the model should travel towards the node close to the data source. Such issues, for example, exist in LSP 5. There, the Aggregator parties constantly generate data on their own assets, and they need the local models to run on their data on their premises. However, DSO needs that model as well to be able to run the simulations for multiple operators at the same time. Therefore, it is needed to provide the inputs to the DSO, so that the DSO can run the simulations on the basis of global model which takes the input from local models (Figure 20).

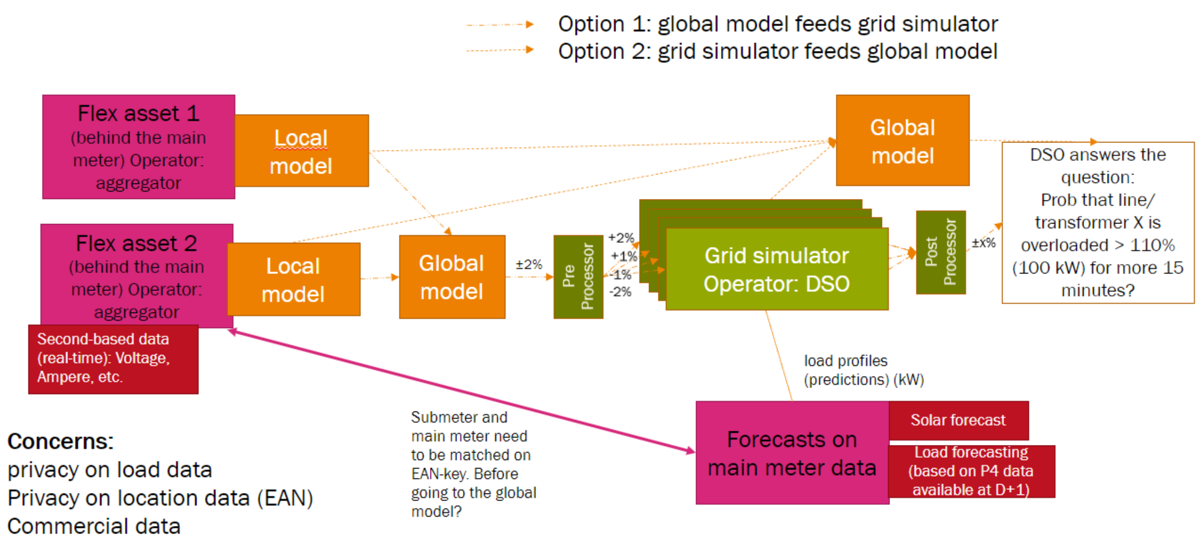

Figure 20: Federated learning architecture for LSP5 Use-case 2

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.

5 Conclusions

With the advent of the Internet of Things and the consequent deluge of data generated from it, smart grids need to reform towards utilising those massive data for operations and services. IoT offers massive societal and economic opportunities, while at the same time poses significant challenges, such as the delivery and management of the technical infrastructure underpinning it, ensuring privacy and security, and capturing value from the huge amount of data.

BD4NRG intends to exploit those opportunities by addressing the emerging challenges in big data management in the energy sector by providing an open holistic solution that proposes a European approach to the business-to-business platforms able to provide competitive edge services to the stakeholders involved in the energy value chain and to create new market opportunities.

In this document, after an introduction of the state of the art, challenges, opportunities, and key researches for BDA for smart grid, how BD4NRG intends to take opportunities in this context has been described in Chapter 2. How BD4NRG is addressing high-performance big data processing has been discussed in Chapter 3. Finally, in Chapter 4, the edge/fog/cloud paradigm and the opportunities, both in terms of architectures and data processing approaches for exploiting it have been presented, as well as how BD4NRG exploits the edge/fog/cloud continuum.

BD4NRG project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 872613.