Integrated HEDNO Platform

Vasiliki Karamesiou f2821903 Erasmia Kornelatou f2821907

Supervisor: Damianos Chatziantoniou

ABSTRACT

Power systems have been through different challenges and technological innovations in the last years and are rapidly evolving into digital systems through the deployment of the smart grids concept. They can benefit from the application of big data analytics, which can help leveraging the optimization processes going on in power grids nowadays. To take advantage of these new opportunities and to keep pace with both the transformation of the Power Sector and changing customer needs, distribution system operators will need to adjust their current role. The objective of this thesis is initially to study and analyze in detail the techniques used in the area of Data Science and Business Analytics, for the collection, preparation and analysis of data, and the deeper understanding of them as well as the extraction of the insight and knowledge they contain, aiming at the best possible business exploitation of them. Then, we aim at helping HEDNO S.A. in the selection of technologies/platforms appropriate to their analytic processes by offering a short-review according to some categories of Big Data problems as processing, storage, data integration, analytics, data governance, and monitoring.

KEYWORDS

Data Science, Business Analytics, Machine Learning, Smart Grid, Smart Meters, Data Integration, Data Warehouse, Data Virtualization, HEDNO S.A.

Table of Contents

1. Big Data Competencies

1.1 What is Big Data?

The constantly increasing momentum of Big Data and their related technologies constitutes an unprecedented market opportunity for improving the energy efficiency and its lifecycle for better managing energy consumption. 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. The so-called 'three Vs' of Big Data [1], are described as follows:

Volume: The size of available data has been growing at an increasing rate. More sources of data are added on continuous basis. In the past, all data was generated internally by employees. Currently, the data is generated by employees, partners and customers. For a group of companies, the data is also generated by machines. More sources of data with a larger size of data tend to increase the volume of data that has to be analyzed. This is a major issue for those looking to use that data instead of letting it just disappear. Peta byte data sets are common these days and Exa byte is not far away. The possible solution to this problem is the distributed systems to store data in different locations connecting them by networks and bringing them together by software. In smart grid, which will be described in a later chapter, the widespread application of smart meters and advanced sensors technology provides huge amount of data.

Variety: Data now comes from a much greater variety of sources compared to traditional data systems. The so-called structured data (tables and other data structures of relational databases, record formats of most applications, and the character-delimited rows of many flat files) which still form a majority of data, is now joined by unstructured data (text, voice, and video) and semi-structured data(XML, JSON, RSS feeds, and hierarchical data). According to the extensive data sources in smart grid as shown in Fig. 1, the formats and dimensions of data are diverse in structure.

Velocity: The growth of data and the resulting importance of it has changed the way we see the data. Once was a time when we did not see the importance of data in the corporate world, but due to the change of how we gather it, we have come to rely on it day to day. Velocity essentially measures how fast the data is coming in. Some data will come in real-time, whereas other will come in fits and starts and sent to us in batches. With a sampling rate of 4 times per hour, 1 million smart meters installed in the smart grid would result in 35.04 billion records, equivalent to 2920 Tb data in quantification.

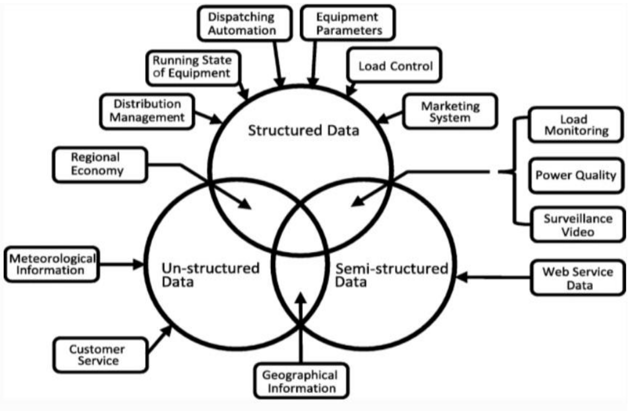

Figure 1: Data Sources of the Grids

1.2 The Data Analysis Lifecycle

Data management is becoming increasingly complex, especially with the emergence of the Big Data era. The best way to manage this data is to dispose a data lifecycle. Data life cycle management is very useful for any enterprise or application where data is being used and processed for producing results. Data's appearance for a certain period of time ensures accessibility and usability in the system. Data is generated through different sources and it is available in various forms for accessibility. A big data-based application such as the energy sector generates lots of data through sensors and other electronic devices which can be further classified into a model for report generations and predictions for various purposes for the benefits of customers and utilities, as well. The lifecycle of data [2] starts from creation to destroy in the system and applications (Fig. 2).

- ***Data Creation*** is the first phase of data life cycle management. Under data creation, various sources can be considered for data generations including data acquisition point, data entry point and sensors installed in industries. At the Data Acquisition point, data is generated by industries outside the enterprises. Data entry is an electronic point that generates lots of data for the enterprise. Sensors generate data suitable for the analytics system to filter and evaluate relevant data and remove missing and irrelevant data. Data creation can be executed through different ways including first-party data collection, third party data collection, repurposing and surveillance data.
- ***Data Store*** is another phase of data life cycle management, which is more important for storage purposes. This process requires movement, integration, enrichment, and ETL (extract, transform and load). Data store phase consists of structured, semi-structured and unstructured types of data. Storage starts with data conceptualization and collection but it never ends.

- ***Data Usability*** indicates how data are usable for industries through technologies and applications. During data usability phase, critical and generated data is reviewed, analyzed, processed and modified.
- ***Data Sharing*** is important in terms of accessibility. All components and modules of the system use data-sharing technologies. Therefore, every component in the system monitors different data, passes through different locations and systems and accesses to the different platforms. It helps to provide meaningful data at the right time.
- ***Data Archive*** since data are required over and over again in the system, it is transferred to a new location for future purposes in the data life cycle management system. In case of need, it can be brought back to an active module for execution. The data archive is a process to transfer and store data under data life cycle management.
- ***Data Destroy*** is the phase where data is no longer required. In this phase the data is deleted since it has been successfully used and now it is without added value.

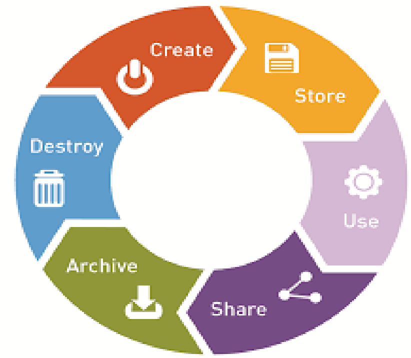

Figure 2: Data Life Cycle Management

1.3 Supervisory Control and Data Acquisition (SCADA)

SCADA is the foundation for the distribution automation system. A typically SCADA System comprises of I/O signal hardware, controllers, software, Networks & communications [3]. The SCADA System also provides a host control function for the supervisor to control and define settings. It typically implements a distributed database, commonly referred to as a tag database, which contains data elements called tag or points. A point represents a single input or output value monitored or controlled by the system.

Supervisory control is a general term for a high-level of overall control of many individual controllers or multiple control loops. It gives the operations supervisor an overview of the plant process and permits integration of operation between low-level controllers. Data acquisition is the process of sampling signals by measuring a physical property of the real world in the form of signals and converting it from analog waveform into digital numeric values so that it can be processed by computing machines. SCADA system comprises of the following components (fig. 3):

- Sensors (either digital or analog) and control relays that directly interface with the managed system.
- Remote telemetry units (RTUs). These are small-computerized units deployed in the field at specific sites and locations. RTUs serve as local collection points for gathering reports from sensors and delivering commands to control relays.
- SCADA master units. These are larger computer consoles that serve as the central processor for the SCADA system. Master units provide a human interface to the system and automatically regulate the managed system in response to sensor inputs.
- The communications network. It connects the SCADA master unit to the RTUs in the field.

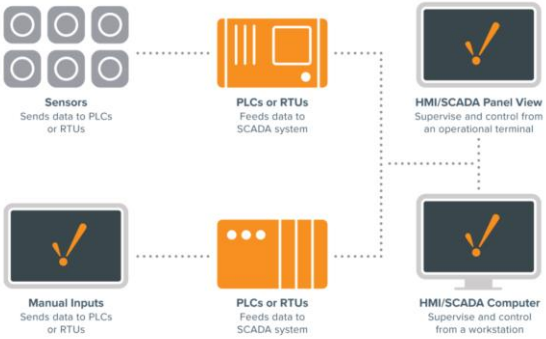
Sensors
Sends data to PLCs or RTUs
PLCs or RTUs
Feeds data to SCADA system
HMI/SCADA Panel View
Supervise and control from an operational terminal
Manual Inputs
Sends data to PLCs or RTUs
PLCs or RTUs
Feeds data to SCADA system
HMI/SCADA Computer
Supervise and control from a workstation

Figure 3: Basic SCADA Diagram

What is more, SCADA has a wide variety of functions, which are crucial to the day-to-day running of electrical power utility. These functions include identifying faults, isolating them and restoring service, circuit breaker and recloser control, switching feeder, voltage regulator, monitoring, temperature transformer, and metering. Commonly SCADA systems are used when a need arises to automate complex processes where human control is not feasible. In power system specifically, this can include:

- The system needs an uninterrupted power supply and a protected environment
- We would need to know the status of a complex power system in real-time
- We would need to monitor and control system that are in remote areas

Advantages of SCADA

SCADA systems are an extremely advantageous way to run and monitor processes [4]. They can be effectively used in large applications such as monitoring and controlling a power plant or mass transit system.

Optimizing performance: SCADA systems minimize errors by accurately measuring data and increasing the overall efficiency of the system.

Reliability and robustness: The specific development of SCADA is performed within a well-established framework that enhances reliability and robustness where power requirement is crucial.

Improve quality: Analyzes and controls the quality of the produced electric energy profile using standard SCADA functionality.

Reduce operating and maintenance costs: Less personnel and trips are required to monitor field gear in remote locations; this reduces maintenance and training costs.

Integrate with business systems: A SCADA system can be easily integrated with the business systems, leading to increased production and profitability.

1.4 Database Management Systems

A database is a collection of data or records. The database management system (DBMS) is a software package designed to define, manipulate, retrieve and manage data in a database (fig.4). A DBMS generally manipulates the data itself, the data format, field names, record structure and file structure. It also defines rules to validate this data. DBMSs are set up on specific data handling concepts, as the practice of administrating a database evolves. The earliest databases only handled individual single pieces of specially formatted data. Today's more evolved systems can handle different kinds of less formatted data and tie them together in more ways that are elaborate. Over time, the models for database management systems have changed considerably. The most common types [5] are described below:

Hierarchical Databases: In a hierarchical database management system (hierarchical DBMSs) model, data is stored in a parent-children relationship node. Besides actual data, records also contain information about their groups of parent/child relationships. Data is organized into a tree-like structure and it is stored in the form of a collection of fields where each field contains only one value. The records are linked to each other via links into a parent-children relationship. In a hierarchical database model, each child record has only one parent. A parent can have multiple children. To retrieve a field's data, we need to traverse through each tree until the record is found. While the hierarchical structure is simple, it is inflexible due to the parent-child one-to-many relationship.

Network Databases: Network database management systems (Network DBMSs) use a network structure to create a relationship between entities and they are mainly used on large digital computers. The network databases are hierarchical databases, but unlike hierarchical databases where one node can have a single parent only, a network node can have a relationship with multiple entities. A network database looks more like a cobweb or interconnected network of records. In network databases, children are called members and parents are called occupiers. The data in a network database is organized in many-to-many relationships.

Relational Databases: In relational database management systems (RDBMS), the relationship between data is relational and data is stored in tabular form of columns and rows. Each column represents an attribute, each row represents a record and each field represents a data value. Relational databases work on each table that has a key field that uniquely indicates each row. These key fields can be used to connect one table of data to another. Relational databases are the most popular and widely used databases. The language used to query RDBMS, including inserting, updating, deleting, and searching records, is Structured Query Language (SQL). Some of the popular DDBMS are Oracle, SQL Server, MySQL, SQLite, and IBM DB2.

NoSQL Databases: NoSQL databases are the databases that do not use SQL as their primary data access language. Graph database, network database, object database, and document databases are common databases of this type. Not having predefined schemas makes them a perfect candidate for rapidly changing development environments. NoSQL allows developers to make changes on the fly without affecting applications. The most popular NoSQL databases are: Cosmos DB, ArangoDB, Couchbase Server, CouchDB, Amazon DocumentDB, MongoDB, CouchBase, Elasticsearch, Informix, SAP HANA and Neo4j .

Object-Oriented Databases: Object DBMS provides full-featured database programming capabilities while containing native language compatibility. It adds the database functionality to object programming languages. This approach is the analogical of the application and database development into a constant data model and language environment. The object-oriented database derivation is the integrity of object-oriented programming language systems and consistent systems.

The power of object-oriented databases comes from the cyclical treatment of both consistent data, as found in databases, and transient data, as found in executing programs. Object-oriented databases use small, recyclable separated from software called objects. The objects themselves are stored in the object-oriented database. Each object contains a piece of data (e.g., sound, video, text, or graphics) and instructions, or software programs called methods, for what to do with the data. Some popular OODBMs are TORNADO, Gemstone, ObjectStore, GBase, VBase, InterSystems Cache, Versant Object Database, ODABA, ZODB, Poet. JADE, and Informix.

Graph Databases: Graph Databases are NoSQL databases and use a graph structure for semantic queries. The data is stored in the form of nodes, edges, and properties. In a graph database, a Node represents an entity or instance such as a customer. A node is equivalent to a record in a relational database system. An Edge represents a relationship that connects nodes. Properties are additional information added to the nodes. The Neo4j, Azure Cosmos DB, SAP HANA, Sparksee, Oracle Spatial and Graph, OrientDB, ArrangoDB, and MarkLogic are some of the popular graph databases. Graph database structure is also supported by some RDBMs including Oracle and SQL Server 2017 and later versions.

Document Databases: Document databases (Document DB) are also NoSQL databases that store data in the form of documents. Each document represents the data, its relationship between other data elements, and attributes of data. Document databases store data in a key value form and have become popular recently due to each document storage and NoSQL properties. NoSQL data storage provides faster mechanism to store and search documents. Popular NoSQL databases are Hadoop/Hbase, Cassandra, Hypertable, MapR, Hortonworks, Cloudera, Amazon SimpleDB, Apache Flink, IBM Informix, Elastic, MongoDB, and Azure DocumentDB.

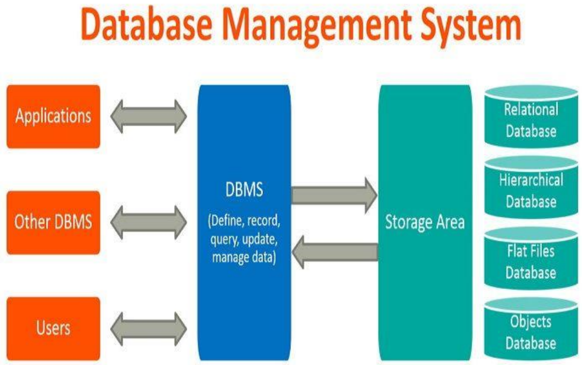

Figure 4: The Database Management System

1.5 Data Engineering

1.5.1 Extract-Transform-Load (ETL) Tools

The accuracy and timeliness of reporting, ad hoc queries and the predictive analysis depend on being able to efficiently get high-quality data into the data warehouse from operational databases and external data sources. Extract-Transform-Load (ETL) refers to a collection of tools that play a crucial role in helping discover and correct data quality issues and efficiently load large volumes of data into the warehouse (fig.5).

The first part of an ETL process [6] involves extracting the data from the source system(s). In many cases, this represents the most important aspect of ETL, since extracting data correctly sets the stage for the success of subsequent processes. Most data-warehousing projects combine data from different source systems. Each separate system may also use a different data organization and/or format. Common data-source formats include relational databases, XML, JSON and flat files, but may also include non-relational database structures or even formats fetched from outside sources. The streaming of the extracted data source and loading on-the-fly to the destination database is another way of performing ETL when no intermediate data storage is required. In general, the extraction phase aims to convert the data into a single format appropriate for transformation processing. An intrinsic part of the extraction involves data validation to confirm whether the data pulled from the sources has the correct/expected values in a given domain (such as a pattern/default or list of values). If the data fails the validation rules, it is rejected entirely or in part. The rejected data is ideally reported back to the source system for further analysis to identify and to rectify the incorrect records.

The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes/tasks:

- Filtering: loading only certain attributes into the data warehouse.
- Cleaning: filling up the NULL values with some default values, mapping, etc.
- Joining: joining multiple attributes into one.
- Splitting: splitting a single attribute into multiple attributes.
- Sorting: sorting tuples based on some attribute (generally key-attribute).

The third and final step of the ETL process is loading. This phase loads the data into the end target, which can be any data store including a simple delimited flat file or a data warehouse. Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis.

Other data warehouses (or even other parts of the same data warehouse) may add new data in a historical form at regular intervals for example, hourly. To understand this, consider a data warehouse that is required to maintain sales records of the last year of a utility company. This data warehouse overwrites any data older than a year with newer data. However, the entry of data for any one-year window is made in a historical manner. The timing and scope to replace or append are strategic design choices dependent on the time available and the business needs. More systems that are complex can maintain a history and audit trail of all changes to the data loaded in the data warehouse.

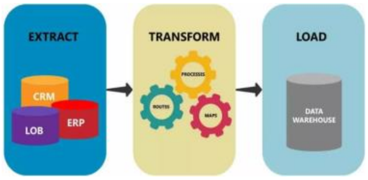

Figure 5: The ETL Process

1.5.2 Web Scraping

Web scraping [7] is a technique to extract data from the World Wide Web (WWW) and save it to a file system or database for later retrieval or analysis. Commonly, web data is scrapped utilizing Hyper, text Transfer Protocol (HTTP) or through a web browser. This is accomplished either manually by a user or automatically by a bot or web crawler. Due to the fact that an enormous amount of heterogeneous data is constantly generated on the WWW, web scraping is widely acknowledged as an efficient and powerful technique for collecting big data.

To adapt to a variety of scenarios, current web scraping techniques [8] have become customized, from smaller ad hoc, human-aided procedures to the utilization of fully automated systems that are able to convert entire websites into a well-organized data set. State-of-the-art web scraping tools are not only capable of parsing markup languages or JSON files, but also integrating with computer visual analytics and natural language processing to simulate how human users browse web content.

The process of scraping data from the Internet can be divided into two sequential steps; acquiring web resources and then extracting desired information from the acquired data. Specifically, a web-scraping program starts by composing a HTTP request to acquire resources from a targeted website. This request can be formatted in either a URL containing a GET query or a piece of HTTP message containing a POST query. Once the request is successfully received and processed by the targeted website, the requested resource will be retrieved from the website and then sent back to the given web-scraping program.

The resource can be in multiple formats, such as web pages that are built from HTML, data feeds in XML or JSON format, or multimedia data such as images, audio, or video files. After the web data is downloaded, the extraction process continues to parse, reformat, and organize the data in a structured way. There are two essential modules of a web-scraping program: a module for composing an HTTP request, such as Urllib2 or selenium and another one for parsing and extracting information from raw HTML code, such as Beautiful Soup or Pyquery.

1.6 Statistical Analysis

Statistical analysis is the science of collecting, exploring and presenting large amounts of data to discover underlying patterns and trends. Statistics are applied every day in research, industry and government in order to become more scientific about decisions that need to be made. For instance, energy companies use statistics to improve services and reduce customer churn by gaining greater insight into customer requirements.

Traditional methods for statistical analysis, from sampling data to interpreting results, have been used by scientists for thousands of years. However, today's data volumes make statistics ever more valuable and powerful. Affordable storage, powerful computers and advanced algorithms have all led to an increased use of computational statistics. Whether you are working with large data volumes or running multiple permutations of your calculations, statistical computing has become essential for today's companies. Popular statistical computing practices [9] include:

- Statistical programming: From traditional analysis of variance and linear regression to exact methods and statistical visualization techniques, statistical programming is essential for making data based decisions in every field.
- Econometrics: Modeling, forecasting and simulating business processes for improved strategic and tactical planning. This method applies statistics to economics to forecast future trends.
- Operations research: Identify the actions that will produce the best results based on many possible options and outcomes. Scheduling, simulation, and related modeling processes are used to optimize business processes and management challenges.
- Matrix programming: Powerful computer techniques for implementing your own statistical methods and exploratory data analysis using row operation algorithms.
- Statistical visualization: Fast, interactive statistical analysis and exploratory capabilities in a visual interface can be used to understand data and build models.
- Statistical quality improvement: A mathematical approach to reviewing the quality and safety characteristics for all aspects of production.

There is a whole range of software packages and tools for statistical analyses and visualization from Access or Excel to dedicated packages, such as SPSS, Stata and R for statistical analysis of quantitative data, SAS for multivariate analysis and predictive analytics, Nvivo for qualitative (textual and audio-visual) data analysis (QDA), or ArcGIS for analyzing geospatial data.

1.7 Machine Learning

Machine learning (ML) is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it to learn for themselves.

Machine Learning becomes more and more important in the energy industry [10] and it is having great potential for the future design of the energy system. Typical areas of application are electricity trading, smart grids, or the sector coupling of electricity, heat and transport. Prerequisites for an increased use of Machine Learning in the energy system are the digitalization of the energy sector and a correspondingly large set of data that is evaluable. ML helps make the energy industry more efficient and secure by analyzing and evaluating the data volumes. In particular, it is present in the field of intelligent networking of electricity consumers and generators across sector boundaries. With the increasing decentralization and digitalization of the power grid, it is becoming more difficult to manage the large number of grid participants and keep the grid in balance. This requires evaluating and analyzing a flood of data. Smart grids are another area of application. These networks transport not only electricity but also data. Especially with an increasing number of volatile power, generation plants such as solar and wind, are becoming more and more important for power generation to react intelligently to consumption.

ML can help evaluate, analyze, and control the data of the various participants (consumers, producers, storage facilities) connected to each other via the grid. A particular focus of ML in the energy industry is on the integration of electro mobility. An increase in e-cars offers opportunities and challenges. The charging of electric cars must be coordinated, but at the same time, they offer the possibility of storing electricity and stabilizing the grid, for example by adjusting the charging demand to price signals and availability. ML can help with all this by monitoring and coordinating. In addition, the ML can stabilize the power grid by, for example, detecting anomalies in consumption, or transmission in near real time, and then develop suitable solutions.

Consumers, intelligently connected in the electricity system, can contribute to a stable and green electricity grid. Smart home solutions and smart meters already exist, but they are not yet widely used. In a smart networked home, the networked devices react to prices on the electricity market and adapt to household usage patterns in order to save electricity and reduce costs. Energy supply and the entire energy system are part of this critical infrastructure. This is why cybersecurity is becoming more and more important today and in the future in order to protect the highly networked power grid from attacks and data theft from the outside. There are already strict security requirements for participants in the electricity market in the area of data protection and data security, though. Finally, in the energy industry, ML offers a multitude of suitable application scenarios that will support the energy transition and a climate-friendly energy system (fig. 6).

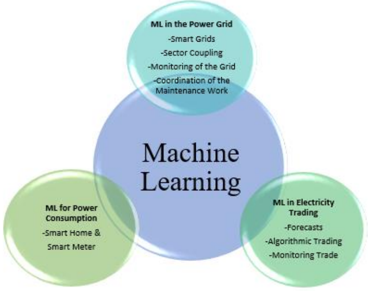

Figure 6: Machine Learning in the Energy Industry

1.8 Data Analysis

Data analysis is an important phase in the data management, which leads the way for assisted decision-making in enterprises. From an analysis perspective, the infrastructure must support both statistical analysis and deeper data mining. The purpose of data analysis is to produce a statistically significant result that can be further used by enterprises to make important decisions.

Different types of analytics [1] change as we move along the continuum of value:

- Descriptive analytics aim to provide information about what happened and it comprises the first step that tries to identify useful information/data for further processing. It might include data visualization, data mining or aggregation of reports. Diagnostic analytics aim to understand the cause of events and system behavior and tries to identify challenges and opportunities.
- Predictive analytics are used to make probabilistic predictions to identify trends with the aim to determine what might happen in the future.
- Prescriptive analytics are applied to identify the best outcome to events, given the system's parameters, and draw strategies to deal with similar events in the future. It uses tools such as simulation techniques and decision support to explore optimal strategies to best take advantage of a future opportunity or to mitigate a future risk.

1.9 Data Stakeholders

An analytics environment encompasses several roles. The distinct roles [37] identified are the following:

Data Engineers: computer scientists with knowledge on data management principles and techniques, relational or otherwise. Their main duty is to extract and transform data from multiple sources and then integrate these to a structure deemed appropriate for input to a data analysis task. They would like to have a model where they can easily and quickly map data and programs' output (e.g. models) onto it. This model could also be used to share data with other data stakeholders in a consistent and semantically proper manner.

Data Scientists/Statisticians: statisticians and/or computer scientists with knowledge on statistical modeling, and/or machine learning techniques. They usually build a model for an entity, using the features (attributes) of that entity. We cannot assume that these people are DB-literate. They would like to have at their disposal a simple conceptual model that makes clear the entities and their attributes, so they can easily select/experiment with those of interest. Possibly, they also want to transform these attributes with built-in aggregate functions or plug-in functions written in Python or R. The end-result is a table that will be used in most cases as input for a learning algorithm.

Data Contributors/End-users: this is an emerging role, though a not well-defined one yet. Under European GDPR law, data contributors (customers, suppliers, employees, users in apps, whoever an organization keeps data for, sometimes called data subjects) are entitled to their data including the results of models built using his/her data. These data have to be delivered to data contributors in a machine-readable format (i.e. somehow structured, such as in excel, JSON, relations, etc.)

Data contributors may use these data to get a better rate from a competitor, sell them to marketing companies or hand them to credit bureaus for some kind of rating. A new market will be built around data exchange, data integration and model building and will fuel the data-driven economy. In that respect, today's tech giants (GAFA) will turn to "data banks" for their users. Data contributors would like to be able to easily select (parts of) their data and export them to a data model of their choice (e.g. relational, semi-structured, etc.) In addition, they should be able to link their data sets from different organizations in a simple manner to create their 'data portfolios'. Some sort of 'self-service' data integration should be easily attainable.

Data Protection Officers: business/legal/information systems people with some good technical skills. Their primary role is to ensure that their organization processes personal data of individuals according to applicable data protection laws. They want to easily see data provenance and consent of data at a fine granular level.

1.10 Data Visualization

Data visualization makes big and small data easier for the human brain to understand, and easier to detect patterns, trends, and outliers in groups of data. Data visualization tools [11] are constantly evolving to offer more powerful features while improving accessibility and user-friendliness. Here are some of the best visual tools [12] for companies:

PowerBI: Microsoft's industry-standard business analytics platform offers some very useful and intuitive data visualization functionalities. PowerBI makes it simple for someone with no prior experience to start creating interactive dashboards and charts from their enterprise data.

Tableau: Tableau is the grandmaster of enterprise data visualization tools, but that does not mean it will be the perfect solution for everyone. Tableau has been considered as one of the leaders in data visualization. It is renowned for generating graphical representations of data.

QlikSense: It supports not just powerful data visualization, but also a full range of analytics services, from data exploration to conversational analytics and augmented intelligence. One of its strengths is the ability to ingest data from multiple sources including external data from cloud services to be combined and analyzed together thanks to its flexible Connect and API functions.

Dundas BI: It is a data visualization and business intelligence platform. An easy tool allows anyone to create attractive data visualizations and dashboards. Dundas BI can also be embedded into your own existing systems and applications, which makes it an attractive choice for those companies that want to enhance their existing tools with better data visualization and analytics capabilities.

Infogram: This is a dedicated visualization tool geared towards creating charts, presentations and dashboards for enterprise use. A simple drag-and-drop interface lets you quickly build documents or slides packed with graphical information. Its collaborative-working elements are particularly strong, enabling teams to experiment with different data representations in different ways to see what works best. If you need a dash-boarding tool specifically designed for infographics, charts, slides, and maps, to plug into an existing analytics platform, Infogram is certainly worth a look.

Furthermore, Python and R (programming languages), offer multiple great graphing libraries that come packed with lots of different features. More specifically, some of the popular plotting libraries are:

Python:

- Matplotlib: Low level, provides lots of freedom
- Pandas Visualization: Easy to use interface, built on Matplotlib
- Seaborn: High-level interface, great default styles
- ggplot: Based on R's ggplot2, uses Grammar of Graphics
- Plotly: Can create interactive plots

R:

- Ggplot2: It works with both univariate and multivariate numerical and categorical data. Thus, it is very flexible. The plot specification is at a high level of abstraction and has a complete graphics system.
- Plotly: This package creates interactive web-based plots. Its advantage is that it can build contour plots, candlestick charts, maps, and 3D charts, which cannot be created using most packages.
- Leaflet: A well-known package based on JavaScript libraries for interactive maps. It is widely used for mapping and working with the customization and design of interactive maps. Besides, Leaflet provides an opportunity to make these maps mobile-friendly.

1.11 Data Governance/ Data Protection/ Data Privacy

The General Data Protection Regulation (GDPR) [13] came into force across the European Union (EU) on 25 May 2018 and is intended to overhaul the way that companies collect and use personal data. GDPR puts the onus on companies to ensure that they have a lawful basis to collect and process personal data (fig.7). The eight basic rights of GDPR that companies must comply with are:

The right to access: This means that individuals have the right to request access to their personal data and to ask how their data is used by the company after it has been gathered. The company must provide a copy of the personal data, free of charge and in electronic format if requested.

The right to be forgotten: If consumers are no longer customers, or if they withdraw their consent from a company to use their personal data, then they have the right to have their data deleted.

The right to data portability: Individuals have a right to transfer their data from one service provider to another. Moreover, it must happen in a commonly used and machine-readable format.

The right to be informed: This covers any gathering of data by companies and individuals must be informed before data is gathered. Consumers have to opt in for their data to be gathered, and consent must be freely given rather than implied.

The right to have information corrected: this ensures that individuals can have their data updated if it is out of date or incomplete or incorrect.

The right to restrict processing: Individuals can request that their data be not used for processing. Their record can remain in place, but not be used.

The right to object: This includes the right of individuals to stop the processing of their data for direct marketing. There are no exemptions to this rule, and any processing must stop as soon as the request is received. In addition, this right must be made clear to individuals at the very start of any communication.

The right to be notified: If there has been a data breach, which compromises an individual's personal data, the individual has a right to be informed within 72 hours of first having become aware of the breach.

Compliance with the requirements of GDPR presents a particular challenge within the energy sector. One high profile example is in connection with the use of smart meters and smart grids. When smart grids are combined with smart metering systems, automatically monitor energy usage, adjust to changes in energy supply and provide real-time information on consumer energy consumption. Moreover, the energy sector faces significant challenges if it wants to both utilize and benefit from large data sets available to it, comply with GDPR and protect the rights of individuals.

Despite the challenges, the benefits of big data analytics for both the company and the individual in the energy sector mean that solutions to these issues must be considered in order to facilitate the growth of domestic demand. Side response services to manage energy consumption more efficiently, respond to changes in local usage and give customers greater visibility and control over their energy consumption.

Figure 7: General Data Protection Regulation (GDPR)

1.12 Open Data

Open data [14] is data that anyone can access, use and share. Governments, businesses and individuals can use open data to bring about social, economic and environmental benefits. It becomes usable when made available in a common, machine-readable format. What is more, open data must be licensed; its license must permit people to use the data in any way they want, including transforming, combining and sharing it with others, even commercially.

The benefits of this Data are diverse and range from improved efficiency of public administrations, economic growth in the private sector to wider social welfare.

Performance can be enhanced by Open Data and contributes to improving the efficiency of public services. Greater efficiency in processes and delivery of public services can be achieved thanks to cross-sector sharing of data, which can provide an overview of unnecessary spending.

The economy can benefit from an easier access to information, content and knowledge in turn contributing to the development of innovative services and the creation of new business models.

Social welfare can be improved as society benefits from information that is more transparent and accessible. Open Data enhances collaboration, participation and social innovation.

Open Data plays a major role in the electric sector. DSOs hold many datasets which, when made publicly available, can help other stakeholders and market parties with e.g. better decision making, create new services and promote synergies between different sectors.

On the other hand, not all data is suitable for publication due to potential breaches of security or violations of privacy regulations. It is therefore important for DSOs to have a common understanding.

European and Development Programs Division (EDPD) took a relevant step towards digitalization by creating an open platform for the provision of information associated with energy data. EDPD launched this platform with the utmost intention of promoting the involvement of society in the energy transition, namely by:

- creating a new platform to make energy data externally and freely accessible

- making data re-usable for different purposes and promoting innovative approaches
- giving external parties the possibility to create societal value based on existing data, acting in compliance with existing regulations (GDPR).

The energy data in the platform is loaded with 15-minute detailed historic information, which goes back to 2014 and is divided in two main groups: generation and consumption, separated by voltage level. The data can be consulted with daily, monthly or yearly aggregation and the stats area of the platform provides an automate comparison with the homologous period. Furthermore, the platform has information regarding the temperature of each day, so that the user can perform a correlation analysis between energy data and temperature. EDPD's open data platform was built on top of their own data handling systems and made available daily, with validated and certified information of the previous day.

1.13 Business Issues

Gradually, Web services, business process management, business performance management, business intelligence and next-generation analytics architecture will begin the transformation into real-time enterprises. During that time, business process management will play the key role among technologies enabling companies' sense and respond transformation.

The Business Process Management (BPM) functional components [15] that companies need to consider are:

- ***Process modeling:*** This component provides a graphical tool for modeling business processes in the as-is and to-be states. Models can also be tailored to depict best practices for exception handling or be prepackaged to reflect energy industry-specific needs. The visual representation (e.g., swim lane diagrams, UML models) must enable a business user (not a developer) to model the process from a business. Different tools support various business process description semantics (i.e., proprietary approaches versus emerging standards such as BPEL, BPML and BPSS). Process modeling is often bundled with a process orchestration engine.
- ***Process improvement methodology***: Aligning the energy company and its enterprise business strategy with a process improvement program is a critical success factor. Many modeling tools have incorporated support for business oriented improvement methodologies (e.g., Six Sigma, lean thinking, business process integration and management, CPI, Balanced Scorecards).
- ***Process orchestration engine:*** A process orchestration engine (POE) takes runtime instructions from a process model. To date, these engines have been fairly proprietary. Many are migrating to support emerging description and execution standards (e.g., BPML, BPEL4WS), yet few of these are commercially available.
- ***Business rules engine vendors:*** Most of the business POE vendors provide lightweight business rules engines embedded in their tools. In other words, the execution engine uses business rules as input (i.e., its runtime instructions). Business rules represent decision choices that a user or a system will make based on a set of conditions. The basic technology has evolved so that user decision points in workflows are surfaced (often via a graphical model or alerts sent to a user).
- ***Integration servers:*** These tools bind the abstracted business process to the data, documents, business logic, messages and events needed by the process. Adapters connect the integration server under the orchestration engine controller to structured data and logic in the underlying applications, unstructured data from the content management environment, search terms (vocabularies) from a taxonomy library and messages/events from message queue managers. The transformation capabilities of these tools provide semantic reconciliation across components as required. In addition, these tools provide basic transport and routing of information and events through the process.
- ***Process monitoring and analysis:*** The primary function of these tools is to enable the analysis of live data as it moves through the process. There are two aspects to this real-time activity monitoring and analysis: 1) analysis of the process itself for optimal design (completeness and bottlenecks); and 2) monitoring the operational process' performance for predefined KPIs and notifying users of out-of-tolerance limits. This provides the opportunity to define and initiate corrective actions. The tool may also provide an end-user-facing dashboard with graphics in a visual display of KPIs for decision makers.
- ***Process simulation/optimization***: These tools are used to simulate the business process through multiple options, discovering bottlenecks and creating alternatives. Allowing the previously mentioned analytics to be captured and used as input into the simulation creates a more efficient process (via both design and runtime feedback). Some products allow a process to be simulated against production conditions to provide feedback.

The volatility of today's business environment puts a premium on being fast and flexible. However, speed does not matter unless you are moving in the right direction. The ability to innovate is a top priority for companies everywhere and, as always, technology plays a vital role. Not only does technology underpin most of your business processes, it has also fundamentally shaped the way you interact with your customers not least by increasing their expectations.

An important challenge facing Energy Distribution companies will be adapting their business model to new technologies and innovation trends [16].

Inclusion of distribution automation: With the inclusion of distribution automation, energy companies will be able to make real-time adjustments to changing loads, generation and failure conditions of the distribution system without operator intervention. The system will support local power grids and ease the load across long-distance transmission lines.

Energy Storage: Energy storage systems allows efficient functioning of electrical system, which helps in improving efficiency, reducing electricity cost and less emissions. Also in the event of power outages, batteries/storage systems present itself as a reliable alternative to power generation. Advancement in energy storage systems will further improve the efficiency of electrical systems. In addition, it will play a critical role in the expansion of renewable energy systems.

Companies across the globe are conducting Research and development (R&D) to develop more advanced energy storage systems to further increase efficiency and reduce cost.

Provision of AI services to customers: Utility companies are facing an increasing demand from customers to provide artificial intelligence (AI) services in order to simplify and enhance interaction with customers. Companies currently engage with customers via channels such as phone, text, and email. However, opting in for AI services will increase customer flexibility with consumers gaining access to platforms including in-home display systems and web portals. The growth of smart homes (i.e. homes powered by IoT devices), will result in the development and integration of new smart home devices and technologies which will be included in business models allowing consumers to access information such as power restoration times, how much energy an appliance is using or how much money is left on their pre-paid accounts.

Cybersecurity: The grid is being modernized through use of advanced power management technologies and has increased interconnections and the ability to interact remotely. This has exposed distributors to the threat of cyber-attack. In order to combat cyber-attacks, these technologies require more intensive cyber security protection. Companies are working on four key areas to fight cyber threat:

- Strengthening energy sector cyber security preparedness
- Coordinating cyber incident response and recovery
- Accelerating research, development and demonstration of resilient energy delivery systems
- Evaluation of IOT and SCADA upgrades with a security first mind-set

Micro grids and Virtual Power Plants: Business models have historically operated on a top-down approach (i.e. Utility selling to consumers).However, this trend is reversing and a bottom-up model of a bi-directional network dominated by Distributed Power Generation (DPG) and consumers are becoming popular. The implementation of Microgrids (MG) and Virtual Power Plants (VPP) will be the key to integrate DPG into the network effectively. MG is an energy system comprising of interconnected loads, distributed energy resources, consumers and an optional storage that acts as a single controllable entity with respect to the grid. VPP on the other hand is a modelling concept that aggregates supply, load and distribution capacity within a specific area. DPG is an option where power is generated and transferred.

Software for grid management: Currently, utility companies are challenged with the variability in generation and the demand for electricity. They are looking for software to enable the management of the grid to ensure quick responses and optimal grid performance.

Social media automated messaging: This platform enables companies to reduce expenses incurred in operation of call centers and call center time. Utility providers are trending towards increasing the use of automated social media direct messaging to inform customers about outages and status of bill payments.

1.14 Big Data Architectures

Technologies and promises connected to big data got a lot of attention lately. Leveraging emerging big data sources extends requirements of traditional data management due to the large volume, velocity and variety of this data. At the same time, it promises to extract value from previously largely unused sources and to use insights from this data to gain a competitive advantage. To gain this value, organizations need to consider new architectures for their data management systems and new technologies to implement these architectures. We identify additional requirements that result from these new characteristics of data, design a reference architecture combining several data management components to tackle these requirements and finally discuss current technologies, which can be used to implement the reference architecture.

The design of the reference architecture takes an evolutionary approach, building from traditional enterprise data warehouse architecture and integrating additional components aimed at handling these new requirements. Big data architecture is designed to handle the following types of work:

- Batch processing of big data sources.
- Real-time processing of big data.
- Predictive analytics and machine learning.

The following different components (fig. 8) are placed in the reference architecture by taking into account inherent characteristics of these components and their interactions with one another [17].

Distributed file system: The distributed file system (DFS) layer resides at the lowest level of this architecture to store and manage large amounts of data across multiple nodes of commodity hardware. DFS is a basic file system that allows disks in a distributed environment to behave as a single virtual disk by breaking the data down into smaller pieces and distributing them throughout the cluster. DFS are commonly designed to conform with master and slave node architecture. Master nodes are responsible for managing job submissions by distributing jobs coequally to slave nodes, which manages processing and collects results from slave nodes. The main benefit of master–slave architecture is the ability to increase the number of slave nodes in the cluster to support vertical scalability. The well-known DFS is called the Hadoop Distributed File System (HDFS).

Cluster management: This architectural component is responsible for deployment, scheduling and orchestrating the jobs across the large networks of nodes to build a readily available and highly scalable computing infrastructure. Therefore, choosing a suitable cluster-management tool is vital for the overall performance of the big-data infrastructure. Apache Mesos, Apache Aurora, Genie-Netflix and Apache Helix can be listed as examples of open-source cluster-management tools for big-data analytics.

Distributed data processing & programming: Big-data use-cases may need to process significant amounts of batch data or millions of data tuples in real time to build a data analysis model and produce results in a timely manner. This significant amount of processing load cannot be handled using traditional methods with a single node solution. To this end, there is a need to constitute an efficient and scalable or distributed programming model and processing solutions, which should be able to deal with the volume and velocity characteristics of big data.

The distributed data-processing tools vary within themselves; however, there are two notable processing methods: batch processing and stream processing. MapReduce is the well-known programming model that implements parallel processing jobs for big-data sets. Distributed batch data-processing tools such as Apache Spark and Hadoop use the MapReduce programming model.

Data Store: The significant amount of data generated by the diverse and large number of data sources is not only too voluminous but also too fast and complex to be stored using traditional storage technologies. In an attempt to store this big data, distributed, scalable, schema-free and fault-tolerant big-data storage technologies that are compatible with a distributed file system are needed. These requirements trigger the development of NoSQL databases, which are increasingly being used in big-data applications. Several types of NoSQL database, which are column-based, key-value-based, document-based, graph based and time-series-based, have been proposed to support specific needs and use cases.

Visualization: As expected, big-data visualization techniques differ from traditional data visualization approaches because of the unique characteristics of big data, such as displaying a high volume of data without collapsing/condensing, dealing with continuously flowing real-time data and separating a variety of categories and structures of data seamlessly. As a result of these challenges, there is a limited number of open-source big-data visualization tools available.

Data analysis: In order to develop a successful data-analysis model, predictive modelling, querying, machine learning and deep learning are indispensable technologies. To this end, querying tools on distributed storage systems, machine learning and deep-learning libraries that support distributed processing are included under this architectural component.

Data pre-processing: An important data pre-processing challenge is about the collection of data from outside sources and transforming these data sets to load in-house data storage systems to maximize the strength of data analytics. CKAN, Apache Griffin and Data Cleaner are among the open-source big-data tools assessed in this category.

Governance & security: Big-data applications tend to present specific governance and security policy enforcements for each individual use-case about the data they have collected. Therefore, this component of our reference architecture mainly addresses open-source solutions for data governance, data security, service programming and benchmarking. For example, Apache Atlas provides a scalable and extensible set of core foundational governance services.

Apache Ranger proposes a data-security framework for monitoring and managing the security of data across the Hadoop platform. HiBench, which was developed by Intel, is a big-data benchmark suite that helps to evaluate different big-data frameworks in terms of speed, throughput and system resource utilizations. Apache Zookeeper is a service-programming tool to develop and maintain extremely reliable distributed coordination across nodes.

Data ingestion: Data ingestion tools help in transferring data from various outside data sources to internal systems in the most efficient and reliable way. They also provide a resilient and fault-tolerant data-distribution method across the architectural components. By taking into account the volume and velocity characteristics of big data, data transferring tools play a crucial role, not only in importing data into big-data platforms but also in the overall performance of the big-data applications. One of the well-known data ingestion tools is Apache Kafka, which is pioneered by LinkedIn. Sqoop, Pulsar, Gobblin and Suro are some of the data-ingestion tools.

Application: This layer mainly provides high-level abstraction to implement specific big data applications and/or present the analysis results produced by the underlying layers to end-users. For example, Nutch is a production-ready web crawler, which is also extremely extensible and scalable in the processing of big data. KillrWeather is another reference application to integrate streaming and batch data processing with well-known open-source tools such as Apache Spark for distributed stream processing, Apache Cassandra for data storage, Apache Kafka to ingest data and Akka for service programming.

Supporting tools: there exist some specific open-source big-data tools that do not fit any other components in the proposed reference architecture, such as:

- Apache Edgent for edge programming, which enables the implementation of applications for small footprint edge devices.
- Apache Knox as an application gateway tool to provide a single access point for all REST and HTTP interactions.
- Apache Tephra for transaction management to provide globally consistent transactions on top of distributed data stores.
- Apache OpenWhisk for emerging server less computing technology to execute big-data functions in response to events.
- Apache River as a networking tool to define scalable and flexible network systems.
- Apache Solr as a search-server.

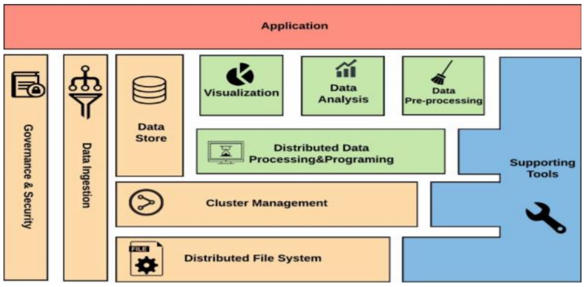

Figure 8: Architecture for Big-Data Analytics

1.15 Edge Computing / Internet of Things

Over the last few years, there has been a great increase in solutions that decentralize communications, data collection and processing, moving all those tasks to the edge. This trend has led the emergence of the Edge Computing paradigm, whose basis is enabling technologies that perform computation at the edge of the network. In this way, computing and network resources (edges) are closer to the source of the data than to the cloud data centers. Edge Computing helps in improving the performance of computer systems by lowering the latency, reducing the cost of resources and increasing responsiveness, scalability, reliability, security or privacy.

The Internet of Things (IoT) [18] is a well-known technology and a broad research topic that has become a reference for data collection and processing systems. IoT systems are formed by multiple and heterogeneous devices (sensors, vehicles, machinery, appliances, meters, etc.) connected through different communication protocols. Multiple disciplines benefit from IoT solutions such as Industry, energy efficiency, smart homes and smart cities. The management of IoT networks is challenging because of the heterogeneity of its resources, creating difficulties in communication protocols, real-time processes, data management, big data storage, security or privacy. In this regard, Edge Computing architectures offer a solution to IoT infrastructures because they are capable of managing the heterogeneous data generated by IoT devices. The basic three-layered Edge Computing [19] architecture is the following (fig. 9):

• Layer 1 - IoT + Sensors: This layer includes IoT devices (sensors, smart meters, smart plugs, etc.) as well as users. The first layer is responsible for the ingestion of data and the operations involved.

• Layer 2 - Edge Nodes: The second layer is formed by Edge nodes. These nodes are responsible for data processing, routing and computing operations.

• Layer 3 - Cloud Services: This layer is formed by multiple cloud services with higher computational requirements. It is responsible for Data Analytics, Artificial Intelligence, Machine Learning, or visualization, among other tasks.

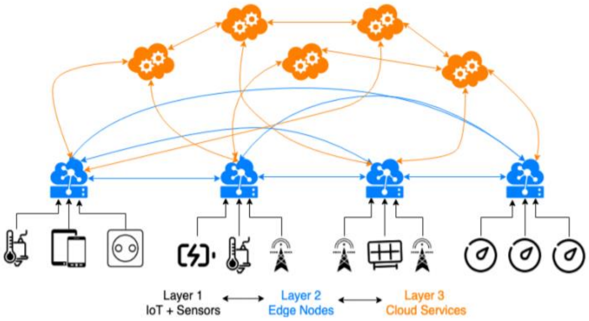

Figure 9: Three-layer Edge Computing basic architecture

Advantages of Using Edge Computing

Improved Performance: Besides collecting data [20] for transmission to the cloud, edge computing also processes, analyses, and performs necessary actions on the collected data locally. Since these processes are completed in milliseconds, it has become essential in optimizing technical data, no matter what the operations may be. Transferring large quantities of data in real-time in a cost-effective way can be a challenge, primarily when conducted from remote industrial sites. This problem is remedied by adding intelligence to devices present at the edge of the network. Edge computing brings analytics capabilities closer to the machine, which cuts out the middle-man. This setup provides for less expensive options for optimizing asset performance.

Reducing Operational Costs: In the cloud-computing model, connectivity, data migration, bandwidth, and latency features are expensive. This inefficiency is remedied by edge computing, which has a significantly less bandwidth requirement and less latency. By applying edge computing, a valuable continuum from the device to the cloud is created, which can handle the massive amounts of data generated. Costly bandwidth additions are no longer required as there is no need to transfer gigabytes of data to the cloud. It also analyses sensitive IoT data within a private network, thereby protecting sensitive data. Enterprises now tend to prefer edge computing. This is because of its optimizable operational performance, address compliance and security protocols, alongside lower costs.

As for electric sector, Edge Computing has a key role in supporting smart grid applications such as demand management and grid optimization. In some cases, edge computing can help with managing energy across enterprises. Sensors and IoT devices connected to an edge platform in factories, plants and offices are being used to monitor energy use and analyze the energy levels in real-time. By tracking and monitoring energy usage in real-time and visualizing it through dashboards, enterprises can better manage their energy consumption and implement preventative measures to limit energy usage.

Edge computing can be particularly useful for managing renewable energy, as well. Specifically, it can promote sustainable management of renewable energy resources. Edge-enabled systems would enable real-time assessment of supply and demand for limited renewable energy resources, such as solar and wind power. Edge computing could be used to provide a real-time view of the energy supply and demand levels in an area, by interacting with IoT applications at an extremely low latency. With the help of microgrids, electricity providers would then be able to supply sufficient levels of renewable energy resource to match the electricity demand of a local area.

2. Smart Meters & Smart Grids

2.1 Introduction

Smart Grid (SG) [21] development has been put into focus due to the increasing complexity of electrical power systems, growing demand on electricity, and the requirement of highly reliable, efficient and secured power supply. SG is considered the next generation power system that uses bi-directional flows of electricity and information. The ability of data integration, system monitoring, reliable data communication, secured data analysis and local and supervisory controls of the smart grid can satisfy the supplier-consumer demand requirements such as reduction in the energy consumption, energy cost and improve the system efficiency. There has been a prolific increase in the energy demand worldwide and electricity is being considered to constitute up to 40% of the total energy generation to meet the growing energy consumption demand in the world by 2040. The monitoring capability of the SG facilitates observability of the entire power network from the energy provider to the energy consumer as well as protects the network from any kind of vulnerability. The upcoming technology in the framework of SG facilitates the development and efficient interactive utilization of millions of alternative distributed energy resources (DER).

Moreover, smart meter [22] is one of the most important devices used in the SG. The smart meter is an advanced energy meter that obtains information from the end users' load devices, measures the energy consumption of the consumers and then provides added information to the utility company and/or system operator. Several sensors and control devices, supported by dedicated communication infrastructure, are utilized in a smart meter. In practice, smart meters can read energy consumption information of customers in real time, such as values of voltage, frequency, and phase angle, and then they securely communicate the information to control centers. Data collected by smart meters is a combination of parameters such as a unique meter identifier, timestamp of the data, and the electricity consumption values. Based on the information, smart meters can monitor and execute control commands for all home devices and appliances at the customer's premises remotely as well as locally.

Besides, smart meters can communicate with other meters in their reach using home area network (HAN) to collect diagnostic information about appliances at the customer as well as the distribution grid. Moreover, smart meters can be programmed such that, only power consumed from the utility grid is billed whereas the power consumed from the distributed generation sources or storage devices owned by the customers is not billed. As a result, they can limit the maximum electricity consumption, and can terminate or reconnect electricity supply to any customer remotely.

In addition, each consumer's location has to be equipped with a smart meter for monitoring and measuring the bi-directional flow of power and supervisory control and data acquisition (SCADA) systems are needed to control the grid operation. While dynamic energy management (DEM) in conventional electricity grids is a well-investigated topic, this is not the case for SGs. This is due to its much more complicated nature, since complex decision-making processes are required by the control centers. Energy management systems (EMSs) in SGs include:

i) real-time wide-area situational awareness (WASA) of grid status through advanced metering and monitoring systems

ii) consumers' participation through home EMSs (HEMS), demand response (DR) algorithms

iii) Supervisory Control through computer-based systems.

2.2 Architectural model of Smart Meter

A smart meter system [24] includes various control devices and sensors to identify parameters and situations in SG and then it transfers the collected data to the control center or provides command signals to the devices in the home of customers (fig.11). The collected electricity consumption data from all devices of customers on a regular basis helps the utility companies to manage electricity demand/response more efficiently and to provide useful information to the customers about the cost-efficient methods to use their appliances. Besides, smart meters can be programmed to maintain a schedule for operation of the home devices and control operation of other appliances accordingly. In addition, by integrating smart meters in electricity grid, utility companies can detect and identify electricity theft and unauthorized consumption in view of improving the power quality and distribution efficiency. Hence, smart meters would play an extremely important role in monitoring the performance and the energy usage characteristics of the load on the electricity distribution grid in the future.

Typically, smart meters implement two major functions, which are communication and measurement. Hence, each meter is equipped with two subsystems as communication and metrology, respectively. The communication part includes security and encryption that define the suitable data transmission approach. The metrology varies depending on multiple characters such as measured phenomenon, technical requirements, region, accuracy, applications, and level of data security. Regardless of the type or quantity of their measurement, smart meters should have six basic functionalities, which include the following [23]:

Quantitative measurement: Smart meters have to accurately measure the quantity of the medium by using various topologies, physical principles, and approaches.

Control and calibration: Smart meters should be providing ability to compensate the small variations according to each system type.

Security communication: The meters have ability receiving operational commands and sending stored data as well as upgrades for its firmware trustworthily.

Power management: Smart meters have to help the system to exactly maintain its functionality when the primary source of energy is lost.

Display: Smart meters will send and display information usage of electricity energy to customers for billing in real time. Besides, the information of real time consumption displayed on smart meters helps customers to manage their demand efficiently.

Synchronization: Typically, smart meters transmit data of customers to the collector systems or central hubs for billing and data analysis. Hence, timing synchronization is very important for reliable transmission of data, particularly in case of wireless communication.

As a result, based on smart meters, utility companies can provide highly reliable, readily accessible, flexible, and cost-effective energy services to their consumers by combining advantages of both small-distributed power generators and large centralized generators. Moreover, demand side management techniques require that these companies have to collect large quantity of data from smart meters in real time. One of key components to implement this concept is advanced metering infrastructure, which collects and analyzes data from smart meters, and gives intelligent management of various power-related applications and services based on that data.

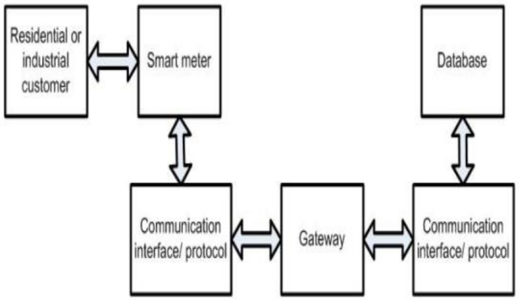
Residential or industrial customer
Smart meter
Database
Communication interface/ protocol
Gateway
Communication interface/ protocol

Figure 10: Architectural model of Smart Meter.

2.3 Smart Grid Architecture

The Fig.11 shows five interfaces between domains, marked with numbers in circles. These are places where communications and exchange of information between the Communication network and other four domains and between Smart metering domain and Customer domain take place. Sample functions at each of these points [25] are listed below:

Point 1: Between Grid domain and Communication Network: It enables the exchange of information and control signals between devices in Grid domain and the Service provider domain. The examples of SCADA and other operations are listed below:

- Remote Terminal Unit (RTU) in transmission systems to enable SCADA operations.
- Intelligent Electronic Devices (IED) in transmission systems to interact with SCADA operations in the Service provider domain.
- Plant control system interacts with SCADA and EMS (Energy Management System) in the Service provider domain.
- Plant control system interacts with Regional Transmission Organizations (RTO)/ Independent Systems Operators (ISO) wholesale market in market operations (e.g., the control signals of monitoring, reporting, and telephony between bulk storage domain and markets to enable wholesale markets operations control hence optimizing portfolios of sources).
- Information and control signals and power generation information between Grid domain (e.g., Bulk generation) and Service provider domain (e.g., control and operations).
- Grid domain (e.g., transmission sensors and measurement devices) provides information from the transmission line to the Service provider domain (e.g., transmission operation, protection and control) for transmission line maintenance information, monitoring, reporting and SCADA.
- Information exchange and coordination between Grid domain (e.g., power generation) and Service provider domain (e.g., power transmission operation and control).
- Distribution sensors and measurement devices provide distribution system information for use by Distributed Energy Resources (DER).

Point 2: Between smart metering domain and Communication Network: It enables the exchange of metering information and interactions through operators and service providers in the Service provider domain towards customers in the Customer domain. Some examples are listed below:

- Management of meters, retrieval of aggregated meter readings from Advanced Metering Infrastructure (AMI) head-end/controller in operations and service provider in Service provider domain.
- Interacting with customer Energy Management System (EMS) to exchange pricing, data related to Demand Response (DR), including the load shedding information, and relevant information enabling automation of tasks involved in a better use of energy.
- Billing in Service provider domain that interacts with the meters in Customer domain.
- Smart meters interact with billing in Service provider domain.
- Smart meters form a metering infrastructure to ensure reliable communication to the meter head-end through this reference point.

Point 3: Between Customer domain and Communication Network domain: It enables the interactions between operators and service providers in Service provider domain and devices in Customer domain. Some examples are listed below:

- The HAN communicates over this Reference point either through a secure energy service gateway or through public network (e.g., Internet).
- Energy Services Interface (ESI)/ HAN gateway interacts with the metering/ billing / utility back office in Service provider domain (Operations).
- ESI/ HAN gateway interacts with the load management system/ demand-response management system in Service provider domain (Operations).
- Customer EMS interacts with energy service provider in Service provider domain.
- Billing in Service provider domain interacts with customers in Customer domain.
- Customer EMS interacts with distribution management system in Grid domain.
- Customer EMS interacts with aggregator/ retail energy provider in Service provider domain.
- Monitoring and controlling the information exchange for distributed generation and DER in Customer domain.

Point 4: Between Service provider domain and Communication Network domain, it enables communications between services and applications in the Service provider domain to actors in other domains to perform all Smart Grid functions illustrated above.

Point 5: Between Smart metering and Customer domain, it conducts services through ESI. Some examples are listed below:

- Smart meter interacts with devices, including customer EMS, ESI in home, customer appliances and equipment.
- Devices in Customer domain, including customer EMS, ESI in home, customer appliances and equipment interact with smart meters.

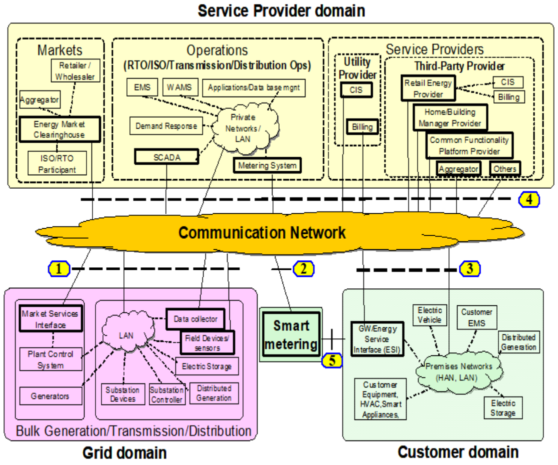

Figure 11: Architecture of Smart Grid

2.4 The Challenges of Smart Grids

Moving to a smart grid infrastructure can help utilities to address some of the challenges that they are now striving to overcome. Specifically, some of these challenges [26] are described below:

Scalable and Interoperable Computing Infrastructure: A SG is a highly distributed system. The huge amount of data is collected from every corner including energy generation, distribution, renewal energy powered vehicles and smart meters. It includes dynamic streaming and non-streaming data, structured and unstructured data. Also, there is a constant flow of the data, e.g., machine to machine, machine to human. It is very challenging to store, share and process such data. A scalable and interoperable computing infrastructure is needed.

Data volume: The amount of data being generated by electric utilities is increasing at an exponential rate. Therefore, big data challenges, such as data storage, data mining, data processing, data querying and data indexing will increase in an unprecedented manner in the future. Due to increased deployment of intelligent devices in consumer and their active engagement on different grid services, the data management expands also to the consumer level. Even at the consumer levels, data volume from various devices (smart meter, electric vehicles, inverters) will be in the order of hundreds of TB. Therefore, effective management of huge volume of data is becoming increasingly challenging issue for utilities. New innovative solutions, such as distributed and scalable computing architecture are necessary. Moreover, dimensionality reduction, a reduced representation of the data set that is much smaller in volume and yet maintaining the integrity of the original data, can significantly reduce data complexities.

Data uncertainty**:** Data uncertainty is one of the defining characteristics of real-world smart grid data and it stems basically from lack of data or an incomplete understanding of the operational context. Since data quality, which is attributed by accuracy, completeness and consistency of data, is one of the biggest concerns in the smart grid, the quality of utility decision depends entirely on the quality of data. However, since real-world data are highly susceptible to errors due to noises and missing/inconsistent data, data cannot be acquired with 100% certainty. Major causes of data uncertainties and loss of data quality stem are sensor inaccuracies and imprecision, communication latencies/delays, cyber-attack, physical damages of equipment, time unsynchronized data, missing/inconsistent data and noises. 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. Probabilistic data analytics and data mining, whereby data uncertainties are modelled as a stochastic process within certain limits, are recently been deployed to deal with data uncertainties. Similarly, data preprocessing techniques (e.g. data cleaning, data integrity, data conditioning) are often used for identifying and removing noisy data, filling in missing values, resolving redundancies, correcting inconsistencies and smoothing out noises and outliers. Data cleaning deals with the missing values, smooth out noises, identifies outliers and corrects inconsistencies within the data.

Data security: Smart grid data mostly involve consumer privacy information, commercial secrets and financial transactions. Therefore, data security (e.g. privacy, integrity, authentication) are very crucial.

- **Data privacy:** Data privacy of the user is a very critical security concern as the power consumption of consumer normally provides insights on their behavior. Data aggregation is one of the common approaches to address data privacy issues. Different techniques such as distributed aggregation, differential aggregation and aggregating with storage are recently developed to address data privacy issues.
- ***Data integrity:*** Data integrity is primarily used to prevent unauthorized modification of information. However, due to close interdependencies between power and communication infrastructure, the power industry is also susceptible to increased cyber/physical-attacks. Those integrity attacks not only deliberately modify financial transactions, but also severely mislead the utility operational decisions. Privacy-preserving data aggregation (P2DA) scheme can ensure data integrity through a digital signature or a message authentication code.
- **Data authentication**: Smart grid data requires authentication as a basis to distinguish legitimate and illegitimate identity. Data authentication is not only necessary to preserve user privacy, but also to ensure data integrity. Therefore, authentication including encryption, trust management, and intrusion detection are important security mechanisms that can prevent, detect, and mitigate network attacks. Different techniques such as data encryption and signature generation are normally used for data authentication and security management in smart grids.

Time synchronization: With the increasing need for real-time control and communication in the smart grid, time synchronization is becoming a key concern. Currently, synchro phasors or PMUs provide time synchronized data, which utilize synchronization based on radio clocks or satellite receivers. Time synchronized data allows analysts to draw meaningful connections between events and aids both forensic analysis of past events, near real-time situational awareness and informed predictive decisions. Forensic determination of a sequence of past events (e.g. what actually tripped, what was the initiating event) and real-time situational awareness of the grid's health can be very powerful to provide a preventive or remedial solution. However, communication, storage and analysis of streams of data from most of the distribution system devices and customers are currently unsynchronized. As unsynchronized data poses a potential risk of a misleading decision, data should be time synchronized with respect to the same time reference.

Data indexing: The smart grid data also possess issues on data indexing and query processing. The existing methods use generic tools such as SQL server and SAP for query purposes. However, these may not suffice from the smart grid application point of view, particularly if real-time applications are sought from the big data. Therefore, advanced data indexing and query-processing algorithms will play critical roles in smart grid big data analytics. State-of-the-art data indexing techniques including variants of R-trees, B-trees, and Quad-trees would definitely be useful for efficiently indexing the big data in smart grids.

Standards and regulation: There are a few standards information models and communication protocols for smart grid interoperability. However, none of the efforts are being yet made on interoperability among big data analytics platforms, architectures, and grid operations frameworks. Instead, different utilities are implementing big data analytics with different storage, computing, processing platforms. Such diversified use of protocols, architectures, and platforms for big data analytics will not only limit its potential but also delay the adoption of big data analytics to power grid. Therefore, to take full advantage of big data application to the smart grid, there is a need for data sharing and information exchange among different utilities and system operators. Since electric utilities usually do not share data/information with each other, a regulatory framework should be established to facilitate data sharing and unify their efforts. In order to synchronize the efforts from utility, industry and academia, there is a strong need to build standards for big data analytics architecture, platforms and interoperability.

Business models and value proposition: To successfully, deploy big data analytics in smart grids, proper business models should be developed. Even though other industries (e.g. Google, Facebook, and Amazon) disruptively transformed their business via big data analytics, electric utilities are still in the initial stage. The business models should be justified on the basis of market opportunity/volume, required investment and values to different stakeholders.

The Utility Analytics Institute has predicted that data-related costs are continuously decreasing. Over the past 30 years, the cost to store data has been cut in half every 14 months or so. The falling costs of data storage and data management are making real-time data collection and storing economically feasible, thereby providing significant opportunities for utilities to make successful business models. However, utilities require a clear understanding of where long-term economic and technical values of big data lie and should develop proper business models for all stakeholders, including utilities, system operators and customers.

2.5 The Benefits of Smart Grids

Smart grid technologies are best viewed in a system context. In such a framework, one might view Smart Grid technologies in terms of what they enable. Smart Grid technologies are expected to enable the following kinds of actions, improvements, and related benefits [27]:

Energy savings through reducing consumption: One of the advantages of smart grids is that they can tell us the consumption at an energy meter at any time, so users are better informed of their real consumption. Moreover, with better consumption monitoring, contracted power can be adjusted to meet the real need of each consumer. These two factors result in users reducing their consumption and tailoring their contracted power to their real needs.

Improve power quality: Smart grid technologies, if deployed in an integrated power grid, can improve the reliability and quality of power supply. With digital technologies increasingly ubiquitous, uninterrupted power supply with consistent voltage, frequency, and related characteristics is increasingly important to individual homes and business operations as well as the productivity of the economy as a whole.

Better customer service and more accurate bills: Another key advantage offered by tele management systems is that bills are more accurate. They always reflect the real consumption of each month instead of estimates, reducing the cost of the old system of manual energy meter readings. In addition to being able to access information about the installation remotely, problems become easier to diagnose and solutions can therefore be implemented faster, improving customer service.

Nowadays customers have to notify companies for them to take action. However, with remote management the system itself automatically reports all incidents to the electric company so it can respond faster to users.

Fraud detection and technical losses: Tele management systems can detect fraud much more accurately, as the units do not contain any parts that are subject to mechanical wear. Moreover, the new energy meters with PLC PRIME communications have systems that detect the opening of the terminal strip cover and send an automatic alert to the managers of the grid warning of potential fraud.

Units with PLC technology can perform energy balances. The system adds together the energy of all the energy meters installed and compares it to the measurement taken by a totalizer at the head of the line to see if there are any losses (or theft) at any point that the company is not aware of.

Improve system security and resilience. Smart grid designs can resist both physical and cyber-attacks. Sensing, surveillance, switching, and intelligent detection, analysis and control software can be built into grid operations to detect and respond to threats. This can make grid systems more resilient, with self-healing technologies that can respond faster and with less impact to human-made and natural incidents.

Reduced balancing cost: Smart Grids can collect much more data than the manual energy meter reading system. This permits the use of data analysis techniques and the preparation of highly realistic consumption forecasts as many more variables are taken into account.

Utilities can then better tailor their production to consumption (balances) and reduce energy surpluses.

Increased competition: Having real load curve data invites marketing companies to adjust their prices based on energy demand. When the marketing companies have more data they can make better offers that are more in line with their customers' reality, increasing competitive options through a wider variety of offers (hourly tariffs, energy packages, etc.).

This benefits consumers in that more competition leads to more competitive pricing.

Levelling of the demand curve (Peak reduction): Through the use of different pricing profiles, utilities can level out the daily demand curve to shift consumption peaks to times with lower demand, optimizing usage of the electrical network. Therefore, customers can intentionally connect loads at off-peak times when each kWh is less expensive. As an example, a customer may decide to change their consumption habits by using the washing machine during off-peak hours, at night, instead of when each kWh is more expensive, saving money and helping the utility balance consumption and avoid line saturation during peak hours. Having consistent consumption means that power plants do not have to switch on and off as many times to generate energy, which lowers generation costs.

Reduction of carbon emissions: All the benefits above involve reducing consumption, which entails a reduction in CO2 emissions. We can thus say that Smart Grids lead to a more sustainable future. All this will directly contribute to the future integration of electric vehicle charging systems in the mains. The deployment of renewable energy systems is also made easier as utilities gain greater control of their grids.

Accommodate diverse generation and storage technologies: These power generation options range from centralized power plants to distributed energy resources (DER) such as system aggregators, grid-scale power projects like wind farms, and building-scale DER such as solar PV or combined heat and power (CHP) systems. Storage systems of various kinds would also be integrated into a mature smart grid system.

2.6 Big Data Analytics Use Cases in Smart Grids

The main procedure of data analytics in smart grid is to extract valuable information from historical data for guiding the operation and maintenance with the comparison to real-time data. The huge amount of data collected from smart meters and sensors are arranged and stored with data management techniques. After preparation, the mathematical model can be established through data mining techniques based on the clean data. With the input of real-time measurements, the state status can be evaluated in the derived model, which provides the possible schemes to guide practical actions and solve potential problems. More specifically, bid data analytics use cases [28] in smart grids are the following:

Fault detection

The carbon emission reduction and sustainability of environment are the driving force and construction purpose of smart grid, which is designed in a decentralized structure. The employment of distributed generator units in modern power distribution system now provides an effective means for the utilization of widespread renewable energy such as wind and solar energy. These emerging microgrids are vital for the expectation of a low-carbon society. Moreover, the close distance between the generator and loads in microgrid improves the reliability of power delivery and reduces the power transmission loss. The ability to operate in an island mode also protects the load from damages caused by power system including voltage fluctuation, frequency deviation. However, the intermittent characteristic of renewable energy increases the uncertainty in power grid, whose typical solution is to use inverter interfaced distributed generators (IIDGs) for a better power quality. In contrast with the traditional bulk generators like large-volume thermal, nuclear or hydro generators, the much lower inertia of IIDGs is a severe potential threat when the faults in microgrids cannot be detected and cleared in a short time due to the limited current carrying capacity.

For a grid-connected microgrid, the severe weather conditions or grid blackouts may trigger an unintentional islanding accident, which threats the safety operation and causes technical issues. Artificial neural networks (ANNs) are trained with features extracted from the differential transient of the rate of change of frequency (ROCOF) signal in order to identify islanding accidents. A support vector machine (SVM) classifier is established with multiple features extracted from system variables as an islanding detection approach. The feature extraction process is implemented with a sliding window whose width is optimized for the highest detection rate.

As a real-time social sensor for the smart grid, social media like Twitter or Facebook could contain potential information indicating the occurrence and location of power outages. A probabilistic framework is devised for detecting a targeted event from the fragmented and noisy tweets. The method shows a good performance in locating accrual outage areas in experiment, which could be integrated to a social data-driven outage management.

Predictive maintenance/condition based maintenance

Distribution automation (DA) is a concept of smart grid, which focuses on the operation and system reliability at the distribution level. A successful DA has the capability to localize and isolate the faults in distribution system with a reduced restoration time and improved customer satisfaction. Under the concept of DA, increasing volume of operational data have been collected from SCADA or advanced metering infrastructure (AMI) for state monitoring and fault diagnosis.

an analyzing scheme for preventative measures to avoid or minimize the outages with the data related to pole mounted autorecloser (PMAR) is proposed. PMAR is a kind of protection intelligent electronic device installed on the overhead lines of a distribution network, which attempts several recloses after an interruption happened in the downstream of the feeder.

Thanks to the development of Information and communications technology (ICT) in power systems, a huge volume of data can be collected via AMI and communication infrastructures. Power system operating data, weather information and log data of relay protection devices are processed as the input of a one class classification system, which is a data-driven model of fault phenomena based on a hybridization of evolutionary learning and clustering techniques. This fault recognition system is validated in the medium voltage power grid. The traditional statistical methods such as linear discriminant analysis (LDA) and logistic regression are discussed for mining the relation between power system faults and the features extracted from raw data.

As a potential threat to the security of transmission systems, the galloping of power lines can cause structural and electrical failures. After analyzing the impact factors of galloping, a data-driven model based on SVM and AdaBoost bi-level classifiers is proposed for early warning. The extreme learning machine (ELM) algorithm is applied in an intelligent early-warning system for reliable online detection of risky events in power system. Since the weights in ELM training are randomly chosen and then determined through matrix computation without iterative parameter adjustment, the learning speed is much faster than conventional algorithms, which is an ideal solution in big data cases. The optimal balance between earning accuracy and warning earliness of the data-driven framework is also discussed. A method is provided to extract electrical features from high-impedance fault current and voltage signals and build an effective feature set (EFS) via a ranking algorithm. Therefore, only a small number of signal channels are required to build a statistical classifier for fault detection. Also, an effective method is provided to reduce the huge volume of PMU data while retaining the critical information for fault detection in power system.

Transient stability analysis

Transient stability is a critical issue closely related to the safety operation of power system. However, the increasing demand for electricity, growing penetration of renewable energy sources and deregulated market force power grid to operate near their secure operating limits. Facing with the challenges from a more complex system, transient stability analysis (TSA) for the study of dynamic behavior taking the electromechanical and electromagnetic process in power system taken into consideration is becoming a hot research topic. The transient process and new operating conditions need to be calculated with the TSA technique after a severe interruption in power grid for a comprehensive protection scheme. Traditional TSA based on the time-domain simulation is not able to provide universal results due to so many uncertainties.

Under the concept of smart grid, a large amount of data collected via AMI are involved in the state assessment of power systems to support the energy management, system operation and decision making. Therefore, efficient summarization techniques are required for extracting useful patterns and discovering valuable information from redundant measurements in power system. A DT-based framework is proposed for the dynamic security assessment (DSA) in power system with high penetration of DGs. Two contingency-oriented DTs are trained based on the databases generated from real-time simulations. One of the well-trained DT is fed with real-time wide-area measurements to identify potential security issues, and the other DT provides the online corresponding preventive control strategies to deal with the problems. The dominant instability generation group (DIGG) in power system is identified without time domain simulation since the features adopted for TSA are extracted from steady-state variables. An approach is proposed to classify the collected data from smart grid into two classes called vulnerable and non-vulnerable data sets with the data analytics such as multichannel singular spectrum analysis (MSSA), principal component analysis (PCA) and SVM. The large spectrum of pre-fault operating state variables and critical clearing times of several contingencies are collected to compose a dataset for pattern recognition methods. The metric which can be used for operating condition evaluation is developed through PCA.

Electric device state estimation/health monitoring

As a vital component for electrical energy conversion, a failure in power transformers may cause catastrophic blackouts in power system. Therefore, the life-cycle management of power transformers based on an accurate estimation attracts a lot of researches for a more stable and reliable power grid. The existing diagnosis methods for power transformers mainly focus on limited state parameters with the threshold-based diagnosis. To take information of system operation and meteorological conditions into state estimation analysis, three classical algorithms for association rule mining are discussed. The rule mining methods are combined with probabilistic graphical model for potential failure prediction.

In most commercial buildings, the building automation system (BAS) is designed and adopted to control the heating, ventilating and airing conditioning (HVAC) system to maintain proper temperature and humidity for the occupants. If the indoor smart grids can be monitored on a continuous or regular basis, a proper operation strategy may be proposed for the improvement of energy efficiency, fault diagnosis and system reliability. A novel health monitoring system is proposed by the fuzzy logic for abnormal operating condition detection. The fault signatures for various fault types are generated by the ANN classification technique.

As the rising number of aging assets in power system is becoming a potential threat to the safety operation, many failure models are proposed focusing on variables of aging time or conditions. A failure rate model is proposed for general electric power equipment with the lifecycle data of service age, maintainer, and health index taken into consideration. In order to make the best use of these data, the stratified proportional hazards model (PHM) is developed as a nonparametric regression method to process and classify the lifecycle data into multi-type recurrent events quantitatively. The potential risk problem and health condition can be predicted with the help of this PHM method.

Power quality monitoring

As a worldwide issue, Electric power quality (PQ) refers to the magnitude, frequency and waveform of voltage and current in power system and highly related to the safe operation of power grid as well as the satisfaction of consumers. With the increasing application of nonlinear and power electronics based loads and generators, the harmonic distortions and instable situations frequently appears in power grid. Deep learning is successfully employed for the classification of PQ events of the electricity networks. Instead of sampling the voltage data of the PQ event data like the existing analysis methods, the image files of the three-phase PQ events are processed for classification by deep learning techniques. Due to the high cost for installation of advance metering devices, the conventional electromechanical analog meters still work in some residential areas and the data analytics-based PQ analysis cannot be properly utilized. A framework is presented that collecting electricity information of from analog meters via image processing techniques. The power consumption information can then be collected to a cloud server through online data exchange. Under the consideration of balance between computation capability and the satisfactory performance of the algorithm, a compact method is presented for feature extraction from the raw data in smart grid to get information that is highly related to the field of power quality. A robust and fast processing pattern recognition algorithm is proposed in power quality events (PQE) classification. The features highly correlated to the PQE are extracted with the discrete wavelet transform-entropy and basic statistical criteria for the establishment of ELM classifier.

Topology identification

Taking the advantage of information layers in smart grid is an effective means to approach the challenges from the renewable energy sources (RES) in distribution network. The measurement, monitoring, communication and control of smart grids by advanced sensors and devices are making the complex network sensible and perceptible. The randomness of RES and uncertainty of the load are increasing the urgency and necessity for a comprehensive decision based on huge volume of data collecting and processing. The SCADA and WAMS provide voltage and power data of smart grid in near real-time sampling rate. Since the network-constrained economic dispatch problems are supposed to be solved by the real-time electricity process in a contemporary wholesale electricity market, the potential of recovering the topology of a grid is explored with market data. Another dynamic solution for online SG topology identification (TI) is proposed which is reformulated as a sparse-recovery problem. Grapy theory and probabilistic DC optimal power flow are adopted for building the network model.

With the purpose for a greener society, the low carbon technologies (LCTs) are driven by the government by application of heat pumps, photovoltaic, electric vehicles and other smart appliances in low voltage (LV) distribution networks. Therefore, the visualization of LV networks with limited metering and data acquisition equipment attracts increasing research interests. The network load profiling based on the identification of representative load profiles of LV systems is an economical alternative method. A novel three-stage network load profiling method aims to evaluate the capabilities of the current LV networks to accommodate the LCTs by clustering, classification and scaling. The first two stages are used to identify the load conditions of unmonitored LV systems with similar fixed data to those monitored LV substations. The contribution factor for each LV template is then determined by the cluster-wise weighted constrained regression algorithm.

Renewable energy forecasting

The abundant and environmental friendly RES such as wind and photovoltaic energies are supposed to be the dominant energy source for the next generation of power grid. However, the randomness and intermittent characteristics are always obstacles for a large-scale utilization of RES in a stable way. To deal with such enormous challenges and get an improved dispatch planning, maintenance scheduling as well as regulation, an accurate and reliable RES forecasting approach has become the hot spot around the world. A data mining based method consisting of k-means and neural networks is proposed.

The meteorological information in historical records are used for clustering approach to classify the days into different categories. Then the bagging algorithm based neural network is trained to get the forecasting results of wind energy. Instead of using the neural network, utilizes the support vector regression method to predict the wind speed with the time series historical wind speed processed by empirical mode decomposition into several intrinsic mode functions and residue. A short-term probabilistic wind generation forecast method is presented based on the sparse Bayesian classification and Dempster-Shafer theory as a nonparametric approach. The forecasting approach of distributed solar energy systems from macro and micro aspects is discussed in a general way with clustering of capacity and location of PV system. The data-driven forecasting approach of PV diffusion is proposed based on cellular automation in microscopic analysis. By decomposing the time-series data with discrete wavelet transform, the proposed recurrent neural network (RNN) model is developed for ultra-short-term solar power prediction.

Load forecasting

Like the RES prediction, an accurate short-term load forecasting is the essential basis for energy management, system operation and market analysis. An increase of forecasting accuracy may bring a lot of benefits and save the investments. With the emerging active role of customers in smart grid, the high efficient dynamic electricity market is also based on a good performance of electricity consumption prediction. Since electricity consumption is affected by the weather conditions to some extent, is proposed a Map/Reduce programming framework for distributed load forecasting by partitioning the geographical area according to local weather information. An extreme learning machine ensemble with a novel wavelet transform is used for electricity consumption after a conditional mutual information based feature selection. To overcome the volatility and uncertainty of load profiles, the recurrent neural network is adopted with a novel pooling layer to avoid overfitting problems. Rather than the aggregated load forecasting, the energy consumption in a single house is usually volatile and difficult to be predicted. Driven by the recent success of deep learning, a long short-term memory recurrent neural network is applied to the residential load forecasting as the latest deep learning techniques. A hidden mode Markov decision process model is developed to the forecast the customers' real-time behavior. An analysis the emerging trends and challenges in the new era of using social media through mobile apps to improve their customer engagement and load forecast is proposed.

A load-forecasting model using knowledge based expert system is proposed in for medium/long term power system planning. A significant development in the field of long term hourly load forecasting is the introduction of a new technique. The methodology establishes a relation between the forecasted hourly load and forecasted annual peak load through some load ratios; hourly, daily, weekly and monthly. A load ratio is considered as a ratio between the peak loads of two different durations of that period. As such, the forecasted hourly load becomes a function of some single valued quantities; i.e. the peaks. Any sorts of load variations over a period cannot be characterized by the peak value of that period only. Moreover, the peak loads may not always represent the actual system demand, because during some peak hours the load may be shaded due to shortage of available generation. That is, the actual peak load would be the recorded load plus the shaded one.

Load profiling

Load profiling is a way to describe the typical behavior of electric consumption, which is usually represented in time domain for load forecasting, demand-side management and capital planning. As an effective method for energy management, the tariff structure designed before is usually based on the type of activity, which is not able to indicate the electrical behavior in a comprehensive way. a two-stage clustering algorithm to classify customers according to their load curves is utilized. In the first-stage, the load patterns are clustered into different categories according to the evaluation index, and then the customers are classified according to the comprehensive load shape factors defined in the first-stage with SVM algorithm. In contrary to the time domain analysis, the DFT method is adopted to discover the information of customers' behavior, which can be accurately reconstructed using limited frequency components and still satisfy the strict requirements.

The residential electricity consumption usually can be divided into three parts: fixed, regulable and deferrable loads, which is the theoretical basis for the optimal energy management of the demand response (DR) mechanism. DR is used to initiate a change in the customers' consumption or feed-in pattern with an incentive from costs or ecological information. the spectral domain analysis methods DWT and DFT is used to decompose smart metering data with the extracted coefficients. Results show that DWT performs better than DFT in individual level while DFT is more suitable to be used in the analysis at a highly aggregated level. A learning based DR strategy combining data analytics and optimization is developed for relatable loads focusing on the residential HVAC. Because when the customers' behavior is obtained, an optimal DR technique for household HVAC unit can be designed based on weather prediction, day-ahead electricity price. Taking the advantage of the social networking to minimize the peak power consumption of the electrical appliances by proposing a 'family plan' approach, which leverages the social network topology and statistical energy usage patterns of the users.

To better understand the information behind the stochasticity and irregularity of residential energy consumption, an in-depth analysis is presented with a finite mixture model-based clustering technique. The self-organizing maps (SOM) as a type of ANN is used to reduce the dimension of collected raw data for load pattern extraction. The frequency-domain data analytics in the SOM shows a superiority over the time-domain data with a higher accuracy in new customer classification. As one of the main tasks of load profiling, a better understanding of the flexibility of customers' electricity consumption is the basis for DR, which can be used to release the pressure of distribution system in terms of thermal and voltage constrains. A multi-resolution analysis method based on wavelet analysis is proposed to extract spectral and time-domain features of load data. Different permutations of typical load profiles provide a more flexible load profiling with a reduction of computation. With the popularization of electric vehicles (EVs), learning the charging load patterns of them is becoming a key step for the stability of power grids. An unsupervised clustering algorithm is used to extract the pattern of EV charging loads with only the real power measurements. Furthermore, the flexibility of the collective EV charging demand is analyzed with Bayesian maximum likelihood. The problem brought by the huge load profile data with the popularity of smart meters installed at the household level, which poses challenges to the communication and storage of measurement data as well as the vital information extraction from massive records. K-SVD sparse representation technique is used to decompose the load profiles into several partial usage patterns for a linear SVM based method to recognize the type of customers.

Load disaggregation

Load disaggregation is also called non-intrusive load monitoring (NILM), aiming to segregate the overall load profiles at household level into the energy consumption of individual appliances. Unlike direct appliance monitoring framework, the NILM from only one smart meter installed in the house is easier to be accepted by the customers. Since different types of the household electric appliances have different potential to be involved in the DR program, the appliance-level load profiles allow the utilities to understand the customers' behavior better and helps to develop a more energy efficient strategy. The early techniques for NILM are mainly based on the detection of 'edge' in power signal to indicate the state 'on' or 'off' of a known device. More effective and complex appliance signatures are then proposed with the harmonics computation of steady-state power or current. The hidden Markov models (HMMs) are adopted with the segmented integer quadratic constraint programming to disaggregate the household power profile at an average frequency of 0.3 Hz into the appliance level. A NILM approach based on the subtractive clustering the maximum likelihood classifier is proposed for a low-sampling-rate date set of 1 Hz sampling rate. The appliances are modeled as ON/OFF states in this event-based load disaggregation algorithm. As a single channel blind source separation problem, the dictionary learning based approaches can be used in NILM. A deep learning approach with multiple layers of dictionaries trained for each device as "deep sparse coding" is utilized. Compared with HMM, the latter method is not suitable for real-time application. By combining the decision tree and nearest-neighbor algorithms, the semi-supervised machine learning is applied to the NILM problem with the signal features extracted by matching a set of net wavelets to the load classes

Last but not least, the big data analytics in smart grids is a comprehensive and complicated field, which does not only depend on the mathematic algorithms or techniques, it also depends on the operation of the systems, the behaviors of vast number of autonomous users, the ICT technologies, the expertise of the field, etc. Therefore, it needs the synergy among experts from different fields if we would like to see the benefits of it in the smart grids.

2.7 Advanced Metering Infrastructure (AMI)

Advanced Metering Infrastructure (AMI) [24] is the basic building block of Smart Grid. It is defined as a system that measure, collect, transfer and analyze energy usage and communicate with metering devices. It enables end users to participate in reducing peak demands and in contributing to energy management process. Further, meters can also capture, receive and execute remote commands like load disconnect/connect.

The main enabling features of an AMI infrastructure include smart meter, communication medium, MDAS/MDM, load monitoring, Demand response, Load control, Tamper detection, Alarm handling, Real time energy audit, Time of Day (ToD) tariff.

Utilities across North America, Europe, Africa and Asia have implemented AMI as a cost effective way to modernize their distribution system while enabling consumer participation in energy management. Various analytics such as energy consumption pattern demand response, tamper detection have benefitted these utilities in cutting non-technical losses, supporting network optimization and controlling energy consumption.

The key components of AMI are:

Smart Energy Meter: act as a source of information for consumer behavior pattern, tamper and load control. It comprises of memory to store information and communication module to transfer this information to Smart Grid Control Center.

Data Concentrator Unit: the data from cluster of smart meters are aggregated by a data concentrator unit (DCU) and then send to the Smart Grid Control Centre. It also sends messages/signals received from the utility/consumer for a particular/all meters to the intended recipient.

Smart Grid Control Centre: Meter Data Acquisition System (MDAS) and Servers are located at Smart Grid Control Centre to perform periodic collection of information from smart meters. Logics and validation rules are defined in Meter Data Management System (MDM) to sanitize the data.

MDAS is a server based meter data head end system compatible with multiple standard based protocols as well as proprietary protocols. MDAS exchange meter data to meter data management systems coupled with analytics on standard data exchange model.

AMI communications infrastructure

In AMI (fig.12), the smart meter can identify power consumption in much more detail than a conventional meter and periodically send the collected information back to the utility company for load monitoring and billing purposes. In addition, the data from smart meter readings are also critical for the control center to implement demand response mechanisms. Using smart meters, customers can control their power consumption and manage how much power they are using, particularly managing the peak load. Hence, through customer participation, the utility companies can likely provide electricity at a lower rate for all their customers, and the consequent carbon dioxide emission will be decreased. Typically, existing AMIs collect data from smart meters and sensors with intervals of 15 min. The collected data are huge and important, and it is estimated that a moderately sized city with 2 million homes could generate 22 GB of meter data every day. In particularly, MDMS with the analytical tools is considered the central module of the management system. Besides, MDMS has to ensure complete and accurate Big Data from customer to the management modules under possible interruptions at lower layers by performing validation, estimation and editing on the AMI data. Moreover, the distribution network automation system, which collects up to 30 samples per second per sensor for real-time control of SG, third-party systems, such as storages or distributed energy resources, connected to the grid, and asset management system responsible for communication among central command are also sources created Big Data in SG. As a result, the communication backbone networks should be reliable, secure, scalable, and cost-effective enough to meet the requirements in terms of bandwidth and latency to communicate the data.

By deploying an AMI [29], reliability, operational efficiency, and customer satisfaction can be achieved. In particular, the AMI communication models include thousands of smart meters, multiple access points, and a mesh network, which is formed between smart meters for data routing purposes using industrial, scientific, and medical (ISM) frequency bands. Meanwhile, the aggregated data are routed to the utility company by access points mostly using licensed bands. The reliability and security of data communications between AMI components suffer from crowded and noisy ISM bands in urban areas. Packet losses, performance degradation, latency, and signal interferences are some of the consequences of heterogeneous spectrum characteristics of the crowded wireless communications. Moreover, the use of licensed bands to communicate the data between access points and utility companies requiring extra costs, which is another obstacle to deploy AMI in SG. Consequently, providing a robust communication backbone is sometimes hardly achievable, and it comes with some obstacles for implementation of AMI in SG. Several works investigated integrated communication technologies for the communication backbone of AMI.

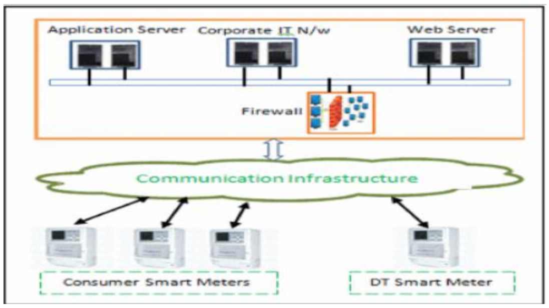

Figure 12: Typical Architecture of AMI

3. Data Integration

The process of combining data from different sources into a single, unified view is called data integration. Data integration is occurring with increasing frequency as the volume and the need to share existing data continues to grow. As a strategy, it is the first step toward transforming data into meaningful and valuable information. The main purpose of data integration is to collect data from different sources and create a system operating at a "higher level" that allows end users, through some query language (or other method), to perform queries and extract data from this unified database without having to go to different sources and perform complex procedures. It is essential in the business field, since it allows them to combine data residing in different sources to provide users with a real-time view of business performance.

3.1 The Benefits of Data Integration

The benefits of data integration [30] are:

1) ***Improve decision making:*** Access to real-time data presented in an easy to digest format will provide you with invaluable acumen, helping you to be proactive, uncovering opportunities and identifying potential bottlenecks before they occur.
2) ***Improve customer experience:*** Siloed data sets prevent a complete view of customers which impacts on sales and ultimately revenue. Only when you have access to real-time customer information as well as the historical data can you target your customers with the right message at the right time to improve your customer experience, loyalty and ultimately revenue.
3) ***Streamline operations:*** From procurement and supply chain to manufacturing and product management, real-time access is useful for improving processes, increasing production, and lowering costs across various departments, including sales, production, distribution and more.
4) ***Increase productivity:*** If you have to constantly move between numerous systems to gather insight, your productivity is significantly reduced. With automatic data integration, your data from all the different sources is pooled into a single customer view, allowing you to be more productive.
5) ***Predict the future:*** Combine historical data with sales pipeline information to make forecasts and anticipate customer demands. This will help you evaluate your products and services, while providing you with the ability to remain ahead of the competition.

3.2 The Challenges of Data Integration

In a big data environment, the data integration can lead to many challenges in real-time implementation, which has the direct impact on projects. Organizations tend to implement new ways to integrate this data to derive meaningful insights at a bigger picture. Some of the challenges in data integration are discussed below:

Accommodate scope of data: Accommodating the sheer scope of data and creating newer domains in the organization are a challenge. This can be addressed by implementing a high performance, computing environment and advanced data storage devices. It possesses better performance levels with reduced latency, high reliability, and quick access to the data. Therefore, it helps accumulate large datasets from all the sources. Another way of addressing this challenge can be through discovery of common operational methodologies between the domains for integrating the query operations, which stands as a better environment to address the challenges for large data entities.

Data inconsistency: It refers to the imbalances in data types, structures, and levels. Although the structured data provides the scope for query operations through relational approach so that the data can be analyzed and used by the organization, unstructured data takes a lead always in larger data entities and this comes as a challenge for organizations. Addressing the data inconsistency can be achieved using the tag and sort methods, which allow searching the data using keywords. The new big data tool Hadoop provides the solution for modulating and converting the data through MapReduce and Yarn. Although Hive in Hadoop does not support the online transactions, they can be implemented for file conversions and batch processing.

Query optimization: In real-time data integration, the large data entities require the query optimization at micro levels, which could involve mapping components to the existing or a new schema, which impacts on the existing structures. To address this challenge, the number of queries can be reduced by implementing the joins, strings, and grouping functions. Also, the query operations are performed on individual data threads which can reduce the latency and responsiveness. Using the distributed joins like merge, hash, and sort can be an alternative in this scenario but requires more resources. Implementing the grouping, aggregation, and joins can be the best approach to address this challenge.

Inadequate resources and implementing support system: Lack of resources haunts every organization at certain point, and this has the direct impact on the project. Limited or inadequate resources for creating data integration jobs, lack of skilled labor that do not specialize in data integration, and costs incurred during the implementation of data integration tools can be some of the challenges faced by organizations in real time. This challenge can be addressed by constant resource monitoring within the organization, and limiting the standards to an extent can save the organizations from bankruptcy. Human resources play a major role in every organization, and this could pick the right professionals for the right task in a timely manner for the projects and tasks at hand.

There is a need to establish a support system for updating requirements and error handling, and reporting is required when organizations perform various data integration jobs within the domains and externally. This can be an additional cost for the organizations as setting up a training module to train the professionals and direct them toward understanding the business expectations and deploy them in a fully equipped environment. This can be termed as a good investment as every organization would implement advancements in a timely manner to stick with the growing market trends. Support system for handling errors could fetch them the reviews to analyze the negative feedback, modify the architecture as per the reviews, and update the newer versions with better functionalities.

Scalability: Organizations could face major challenge in maintaining the data accumulated from number of years of their service. This data is stored and maintained using the traditional file systems or other methodologies as per their environment. In this scenario, often the scalability issues arise when the new data from multiple resources is integrated with data from legacy systems. Changes made by the data scientists and architects could impact the functioning of legacy systems as it has to go through many updates to match the standards and requirements of new technologies to perform a successful data integration. In recent times, mainframe stands as one of the best example for legacy system. For a better data operation environment and rapid access to the data, Hadoop has been implemented to handle the batch processing unit. This follows a typical ETL (Extract, Trans-form, and Load) approach to extract the data from number of resources and load them into Hadoop environment for the batch processing.

3.3 Approaches to Integration

In this section, an architectural perspective is applied in order to give an overview of the different ways to address the integration problem. Information systems can be described using a layered architecture, as shown in Fig. 13. On the topmost layer, users access data and services through various interfaces that run on top of different applications. Applications may use middleware transaction processing (TP) monitors, message oriented middleware (MOM), SQL-middleware to access data via a data access layer. The data itself is managed by a data storage system. Usually, database management systems (DBMS) are used to combine the data access and storage layer.

In general, the integration problem can be addressed on each of the presented system layers. For this, the following principal approaches [30] as illustrated in Fig. 13 are available:

Manual Integration: Here, users directly interact with all relevant information systems and manually integrate selected data. That is, users have to deal with different user interfaces and query languages. Additionally, users need to have detailed knowledge on location, logical data representation, and data semantics.

Common User Interface: In this case, the user is supplied with a common user interface (e.g., a web browser) that provides a uniform look and feel. Data from relevant information systems is still separately presented so that homogenization and integration of data yet has to be done by the users (for instance, as in search engines).

Integration by Applications: This approach uses integration applications that access various data sources and return integrated results to the user. This solution is practical for a small number of component systems. However, applications become increasingly fat as the number of system interfaces and data formats to homogenize and integrate grows.

Integration by Middleware: Middleware provides reusable functionality that is generally used to solve dedicated aspects of the integration problem, e.g., as done by SQLmiddleware. While applications are relieved from implementing common integration functionality, integration efforts are still needed in applications. Additionally, different middleware tools usually have to be combined to build integrated systems.

Uniform Data Access: In this case, a logical integration of data is accomplished at the data access level. Global applications are provided with a unified global view of physically distributed data, though only virtual data is available on this level. Local information systems keep their autonomy and can support additional data access layers for other applications. However, global provision of physically integrated data can be time-consuming since data access, homogenization, and integration have to be done at runtime.

Common Data Storage: Here, physical data integration is performed by transferring data to a new data storage; local sources can either be retired or remain operational. In general, physical data integration provides fast data access. However, if local data sources are retired, applications that access them have to be migrated to the new data storage as well. In case local data sources remain operational, periodical refreshing of the common data storage needs to be considered. In practice, concrete integration solutions are realized based on the presented six general integration approaches. Important examples include:

- Mediated query systems represent a uniform data access solution by providing a single point for read-only querying access to various data sources. A mediator that contains a global query processor is employed to send subqueries to local data sources; returned local query results are then combined.
- Portals as another form of uniform data access are personalized doorways to the internet or intranet where each user is provided with information according to his detected information needs. Usually, web mining is applied to determine user-profiles by click-stream analysis; thereby, information the user might be interested in can be retrieved and presented.

Data warehouses realize a common data storage approach to integration. Data from several operational sources (on-line transaction processing systems, OLTP) are extracted, transformed, and loaded (ETL) into a data warehouse. Then, analysis, such as online analytical processing (OLAP), can be performed on cubes of integrated and aggregated data. **Operational data store** is a second example of a common data storage. Here, a "warehouse with fresh data" is built by immediately propagating updates in local data sources to the data store. Thus, up-to-date integrated data is available for decision support. Unlike in data warehouses, data is neither cleansed nor aggregated nor are data histories supported. **Federated database systems (FDBMS**) achieve a uniform data access solution by logically integrating data from underlying local DBMS. Federated database systems are fully-fledged DBMS; that is, they implement their own data model, support global queries, global transactions, and global access control.

Workflow management systems (WFMS) allow to implement business processes where each single step is executed by a different application or user. Generally, WFMS support modeling, execution, and maintenance of processes that are comprised of interactions between applications and human users. WFMS represent an integration-by-application approach.

Integration by web services performs integration through software components that support machine-to-machine interaction over a network by XML-based messages that are conveyed by internet protocols. Depending on their offered integration functionality, web services either represent a uniform data access approach or a common data access interface for later manual or application-based integration. **Model management** introduces high-level operations between models (such as database schemas, UML models, and software configurations) and model mappings; such operations include matching, merging, selection, and composition. Using a schema algebra that encompasses all these operations, it is intended to reduce the amount of hand-crafted code required for transformations of models and mappings as needed for schema integration. Model management falls into the category of manual integration.

Peer-to-peer (P2P) integration is a decentralized approach to integration between distributed, autonomous peers where data can be mutually shared and integrated through mappings between local schemas of peers. P2P integration constitutes, depending on the provided integration functionality, either a uniform data access approach or a data access interface for subsequent manual or application-based integration.

Grid data integration provides the basis for hypotheses testing and pattern detection in large amounts of data in grid environments, i.e., interconnected computing resources being used for high-throughput computing. Here, often unpredictable and highly dynamic amounts of data have to be dealt with to provide an integrated view over large (scientific) data sets. Grid data integration represents an integration by middleware approach.

Personal data integration systems are a special form of manual integration. Here, tailored integrated views are defined either by users themselves or by dedicated integration engineers. Each integrated view precisely matches the information needs of a user by encompassing all relevant entities with real-world semantics as intended by the particular user; thereby, the integrated view reflects the user's personal way to perceive his application domain of interest.

Collaborative integration another special form of manual integration, is based on the idea to have users to contribute to a data integration system for using it. Here, initial partial schema mappings are presented to users who answer questions concerning the mappings; these answers are then taken to refine the mappings and to expand the system capabilities. Similar to folksonomies, where data is collaboratively labelled for later retrieval, the task of schema mapping is distributed over participating users.

In Dataspace systems co-existence of all data (i.e., both structured and unstructured) is propagated rather than full integration. A dataspace system is used to provide the same basic functionality, e.g., search facilities, over all data sources independently of their degree of integration. Only when more sophisticated services are needed, such as relational-style queries, additional efforts are made to integrate the required data sources more closely. In general, dataspace systems may simultaneously use every one of the presented six general integration approaches.

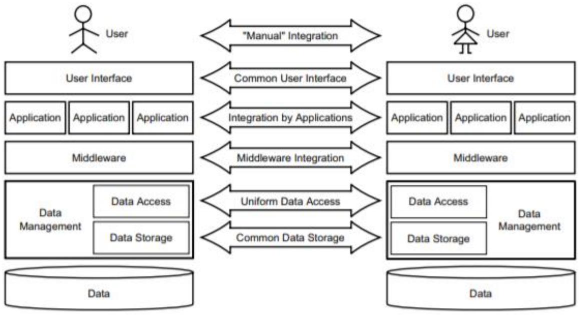

Figure 13: General Integration Approaches on Different Architectural Levels

4. Data Warehouse

4.1 Introduction

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker to make better and faster decisions. Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis).

A typical data warehouse architecture [31] for supporting Business Intelligence (BI) within an enterprise is shown in Figure 14. The data over which BI tasks are performed often comes from different sources, typically from multiple operational databases across departments within the organization, as well as external vendors. Different sources contain data of varying quality, use inconsistent representations, codes, and formats, which have to be reconciled. Thus, the problems of integrating, cleansing, and standardizing data in preparation for BI tasks can be rather challenging. Efficient data loading is imperative for BI. Moreover, BI tasks usually need to be performed incrementally as new data arrives, for example, last month's bills data. This makes efficient and scalable data loading and refresh capabilities imperative for enterprise BI. These back-end technologies for preparing the data for BI are collectively referred to as Extract-Transform-Load (ETL) tools. Increasingly there is a need to support BI tasks in near real time, that is, make business decisions based on the operational data itself.

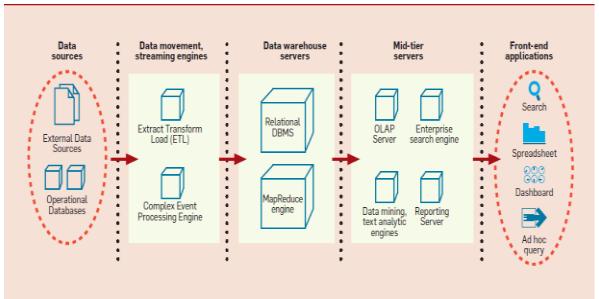

Figure 14: Typical Data Warehousing Architecture.

4.2 Technologies

A data warehouse is a subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision-making. Typically, the data warehouse is maintained separately from the organization's operational databases. There are many reasons for doing this. The data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases.

OLTP applications [32] typically automate clerical data processing tasks such as order entry and banking transactions that are the bread-and-butter day-to-day operations of an organization. These tasks are structured and repetitive, and consist of short, atomic, isolated transactions. Operational databases tend to be hundreds of megabytes to gigabytes in size. Consequently, the database is designed to reflect the operational semantics of known applications, and, in particular, to minimize concurrency conflicts. Data warehouses, in contrast, are targeted for decision support. Historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be orders of magnitude larger than operational databases; enterprise data warehouses are projected to be hundreds of gigabytes to terabytes in size. The workloads are query intensive with mostly ad hoc, complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. To facilitate complex analyses and visualization, the data in a warehouse is typically modeled multidimensionality.

Typical OLAP operations include rollup (increasing the level of aggregation) and drill-down (decreasing the level of aggregation or increasing detail) along one or more dimension hierarchies, slice_and_dice (selection and projection), and pivot (re-orienting the multidimensional view of data). Given that operational databases are finely tuned to support known OLTP workloads, trying to execute complex OLAP queries against the operational databases would result in unacceptable performance. Furthermore, decision support requires data that might be missing from the operational databases; for instance, understanding trends or making predictions requires historical data, whereas operational databases store only current data. Decision support usually requires consolidating data from many heterogeneous sources. The different sources might contain data of varying quality, or use inconsistent representations, codes and formats, which have to be reconciled. Finally, supporting the multidimensional data models and operations typical of OLAP requires special data organization, access methods, and implementation methods, not generally provided by commercial DBMSs targeted for OLTP. It is for all these reasons that data warehouses are implemented separately from operational databases.

Data warehouses might be implemented on standard or extended relational DBMSs called Relational OLAP (ROLAP) servers. These servers assume that data is stored in relational databases, and they support extensions to SQL and special access and implementation methods to efficiently implement the multidimensional data model and operations. In contrast, multidimensional OLAP (MOLAP) servers are servers that directly store multidimensional data in special data structures (e.g., arrays) and implement the OLAP operations over these special data structures.

There is more to building and maintaining a data warehouse than selecting an OLAP server and defining a schema and some complex queries for the warehouse. Different architectural alternatives exist. Many organizations want to implement an integrated enterprise warehouse that collects information about all subjects spanning the whole organization. However, building an enterprise warehouse is a long and complex process, requiring extensive business modeling, and may take many years to succeed. Some organizations are settling for data marts instead, which are departmental subsets focused on selected subjects. These data marts enable faster roll out, since they do not require enterprise-wide consensus, but they may lead to complex integration problems in the long run, if a complete business model is not developed.

4.3 Distributed Systems using Map-Reduce Paradigm

Large-scale data processing engines based on the Map-Reduce paradigm were originally developed to analyze Web documents, query logs, and click-through information for index generation and for improving Web search quality. Platforms based on a distributed file system and using the MapReduce runtime (or its variants) have been successfully deployed on clusters with an order of magnitude more nodes than traditional parallel DBMSs. Also, unlike parallel DBMSs where the data must first be loaded into a table with a predefined schema before it can be queried, a MapReduce job can directly be executed on schema-less input files. Furthermore, these data platforms are able to automatically handle important issues such as data partitioning, node failures, managing the flow of data across nodes, and heterogeneity of nodes.

Another factor that makes such platforms attractive is the ability to support analytics on unstructured data such as text documents (including Web crawls), image and sensor data by enabling execution of custom Map and Reduce functions in a scalable manner. Recently, these engines have been extended to support features necessary for enterprise adoption (Cloudera8). While serious enterprise adoption is still in early stages compared to mature parallel RDBMS systems, exploration using such platforms is growing rapidly, aided by the availability of the open source Hadoop ecosystem. Driven by the goal of improving programmer productivity while still exploiting the advantages noted here, there have been recent efforts to develop engines that can take a SQL-like query, and automatically compile it to a sequence of jobs on a MapReduce engine. The emergence of analytic engines based on MapReduce is having an impact on parallel DBMS products and research. For example, some parallel DBMS vendors allow invocation of MapReduce functions over data stored in the database as part of a SQL query. The MapReduce function appears in the query as a table that allows its results to be composed with other SQL operators in the query. Many other DBMS vendors provide utilities to move data between MapReduce based engines and their relational data engines. A primary use of such a bridge is to ease the movement of structured data distilled from the data analysis on the MapReduce platform into the SQL system.

4.4 Benefits of a Data Warehouse

The benefits [33] of a data warehouse include improved data analytics, greater revenue and the ability to compete more strategically in the marketplace.

Enables Historical Insight: No business can survive without a large and accurate storehouse of historical data, from sales and inventory data to personnel and intellectual property records. If an electric company suddenly needs to know the bills of the customers 24 months ago, the rich historical data provided by a data warehouse make this possible. Also, a data warehouse can add context to this historical data by listing all the key performance trends that surround this retrospective research. This kind of efficiency cannot be matched by a legacy database.

Enhances Conformity and Quality of Data: Business generates data in myriad different forms, including structured and unstructured data. A data warehouse converts this data into the consistent formats required by analytics platforms. Moreover, by ensuring this conformity, a data warehouse ensures that the data produced by different business divisions is at the same quality and standard allowing a more efficient feed for analytics.

Boosts Efficiency: It is very time consuming for a business user or a data scientist to have to gather data from multiple sources. It’s far more advantageous for this data to be gathered in one place, hence the benefit of a data warehouse. Additionally, if a data scientist needs data to run a fast report, he doesn’t need to get the assistance from tech support to perform this task. A data warehouse makes this data readily available in the correct format improving efficiency of the entire process.

Increase the Power and Speed of Data Analytics: Business intelligence and data analytics are the opposite of instinct and intuition. BI and analytics require high quality, standardized data on time and available for rapid data mining. A data warehouse enables this power and speed, allowing competitive advantage in key business sectors, ranging from CRM to HR to sales success to quarterly reporting.

Drives Revenue: Creating more standardized and better quality data is the key strength of a data warehouse, and this key strength translates clearly to significant revenue gains. The data warehouse formula works like this: Better business intelligence helps with better decisions, and in turn better decisions create a higher return on investment across any sector of your business. Most important, these revenue gains build on themselves over time, as better decisions strengthen the business. In short, a high quality, fully scalable data warehouse can be seen as less of a cost and more of an investment – one that adds exponential value like few other investments that businesses make.

Scalability: The top key word in the cloud era is "scalable" and a data warehouse is a critical component in driving this scale. A topflight data warehouse is itself scalable, and enables greater scalability in the business overall. That is, today's sophisticated data warehouse are built to scale, handling ever more queries as the business grows. Moreover, the efficiency in data flow enabled by a data warehouse greatly boosts a business's growth. This growth is the core of business scalability.

Interoperates with On-Premise and Cloud: Unlike the legacy databases of yesteryear, today's data warehouses are built with multicloud and hybrid cloud in mind. Many data warehouses are now fully cloud-based, and even those that are built for on premise typically will interoperate well with the cloud-based portion of a company's infrastructure. As an additional important side point: this cloud based focus also means that mobile users are better able to access the data warehouse. This is beneficial for sales reps in particular.

Data Security: A number of key advances in data warehouse have enhanced their security, which enhances the overall security of company data.

Much Higher Query Performance and Insight: The constant business intelligence queries that are part of today's business can put a major strain on an analytics infrastructure, from the legacy databases to the data marts. Having a data warehouse to more effectively handle queries removes some of the pressure on the system. Furthermore, since a data warehouse is specifically geared to handle massive levels of date and myriad complex queries, it is the high functioning core of any business's data analytics practice.

Provides Major Competitive Advantage: This is absolutely the bottom line benefit of a data warehouse. It allows a business to more effectively strategize and execute against other vendors in its sector. With the quality, speed and historical context provided by a data warehouse, the greater insight in data mining can drive decisions that create more sales, more targeted products, and faster response times.

4.5 Data Warehouse Tools

Data Warehousing tools [34] can help analyze large volumes of disparate data from varied sources to provide meaningful business insights and there are many available in the market. Following is a curated list of most popular open-source and commercial Data Warehouse tools with key features:

1) Oracle

Oracle data warehouse software is a collection of data, which is treated as a unit. The purpose of this database is to store and retrieve related information. It helps the server to reliably manage huge amounts of data so that multiple users can access the same data.

Features:

- Distributes data in the same way across disks to offer uniform performance
- Works for single-instance and real application clusters
- Offers real application testing
- Common architecture between any Private Cloud and Oracle's public cloud
- Hi-Speed Connection to move large data
- Works seamlessly with UNIX/Linux and Windows platforms
- It provides support for virtualization
- Allows connecting to the remote database, table, or view

2) Amazon RedShift

Amazon Redshift is an easy to manage, simple, and cost-effective data warehouse tool. It can analyze almost every type of data using standard SQL.

Features:

- No Up-Front Costs for its installation
- It allows automating most of the common administrative tasks to monitor, manage, and scale your data warehouse
- Possible to change the number or type of nodes
- Helps to enhance the reliability of the data warehouse cluster
- Every data center is fully equipped with climate control
- Continuously monitors the health of the cluster. It automatically re-replicates data from failed drives and replaces nodes when needed

3) SAP

SAP is an integrated data management platform, to maps all business processes of an organization. It is an enterprise level application suite for open client/server systems. It is one of the best data warehouse tools that has set new standards for providing the best business information management solutions.

Features:

- It provides highly flexible and most transparent business solutions
- The application developed using SAP can integrate with any system
- It follows modular concept for the easy setup and space utilization
- You can create a Database system that combines analytics and transactions. These next next-generation databases can be deployed on any device
- Provide support for On-premise or cloud deployment
- Simplified data warehouse architecture
- Integration with SAP and non-SAP applications

4) SAS

SAS is a leading data warehousing tool that allows accessing data across multiple sources. It can perform sophisticated analyses and deliver information across the organization.

Features:

- Activities managed from central locations. Hence, user can access applications remotely via the Internet
- Application delivery typically closer to a one-to-many model instead of one-to-one model
- Centralized feature updating, allows the users to download patches and upgrades.
- Allows viewing raw data files in external databases
- Manage data using tools for data entry, formatting, and conversion
- Display data using reports and statistical graphics

5) IBM – DataStage

IBM data Stage is a business intelligence tool for integrating trusted data across various enterprise systems. It leverages a high-performance parallel framework either in the cloud or on premise. This data warehousing tool supports extended metadata management and universal business connectivity.

Features:

- Support for Big Data and Hadoop
- Additional storage or services can be accessed without need to install new software and hardware
- Real time data integration
- Provide trusted ETL products data anytime, anywhere
- Solve complex big data challenges
- Optimize hardware utilization and prioritize mission-critical tasks
- Deploy on-premises or in the cloud

6) Informatica

Informatica PowerCenter, is Data Integration tool developed by Informatica Corporation. The tool offers the capability to connect and fetch data from different sources.

Features:

- It has a centralized error logging system which facilitates logging errors and rejecting data into relational tables
- Build in Intelligence to improve performance
- Limit the Session Log
- Ability to Scale up Data Integration
- Foundation for Data Architecture Modernization
- Better designs with enforced best practices on code development
- Code integration with external Software Configuration tools
- Synchronization amongst geographically distributed team members

7) MS SSIS

SQL Server Integration Services is a Data warehousing tool that used to perform ETL operations; i.e. extract, transform and load data. SQL Server Integration also includes a rich set of built-in tasks.

Features:

- Tightly integrated with Microsoft Visual Studio and SQL Server
- Easier to maintain and package configuration
- Allows removing network as a bottleneck for insertion of data
- Data can be loaded in parallel and various locations
- It can handle data from different data sources in the same package
- SSIS consumes data, which are difficult like FTP, HTTP, MSMQ, and Analysis services.
- Data can be loaded in parallel to many varied destinations.

5. Data Virtualization

5.1 Introduction

Data virtualization [35] is a relatively new business trend and is closely related to mediators and virtual databases (a form of data integration), if not a reinvention of these. Data virtualization is any approach to data management that allows an application to retrieve and manipulate data without requiring technical details about the data, such as how it is formatted at source, or where it is physically located and can provide a single customer view (or single view of any other entity) of the overall data (fig.15).

This technique can make the architectures of business intelligence systems simpler, cheaper and more flexible. The introduction of this technique does not mean, however, that the data warehouse should be abandoned but expanded. Virtualization is the solution to the need to implement the idea of operational business intelligence and to extend existing business intelligence architectures without disrupting ETL processes.

To realize the great innovation of such a discovery, it is enough to understand why it is constantly emphasized that the user can handle the data with the least possible effort and complexity. Take for example the smart meter. It represents digitally continuous power consumption data. In particular, consumers have the ability to monitor their consumption and consequently estimate the amount they will pay without waiting for the electricity bill. They can also save electricity by reducing unnecessary energy consumption.

The dashboard is a simple and everyday example of data virtualization. The user-consumer simply and quickly receives the information he needs from different data sources and in different formats. The way the information reaches the end user does not matter and the system does not require any data processing at all in order to come in a readable and understandable form. A complex data retrieval and calculation system that would require end user effort would be disastrous and completely pointless. Finally, data virtualization is a very useful solution in terms of business intelligence data federation, data warehouse expansion, big data integration, cloud data integration.

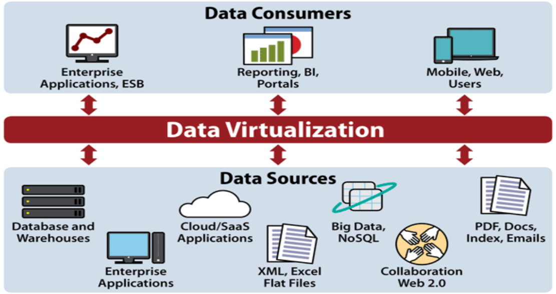

Figure 15: Data Virtualization Architecture

5.2. Advantages of Data Virtualization

Why is data virtualization beneficial? What more can it offer than simply using SQL for fast data recovery? Below are some of its advantages [36]:

- The end user can choose the data virtualization tool that allows them to export the data they want using the language they know best, for example Java, Python or other languages such as MDX, XQuery or CQL.
- Data sources can be easily replaced without affecting the end-user view.
- In order to understand the importance of this particular advantage we must first explain what it is and what is the role of cache in a computer. The cache is essentially part of the main memory and stores temporary data needed by RAM (random access memory). Simply put, it is a very fast accessible memory, ideal for big data processing. Virtualization tools allow users to process data at the memory cache level, which means that a copy of the data is transferred to the cache and the data can be accessed by also another user without interference or delays.
- The intermediate layer between data sources and data consumers (data consumers - which are the end users), ie the level of data virtualization allows the recording and definition of metadata, which makes the application simpler and less "demanding" in terms of resource consumption. When we talk about metadata, we mean the format of arrays, their attributes and the values of attributes. Because they are defined at the intermediate level that of data virtualization it is easier and simpler for data consumers the user level to retrieve this information.
- The data consolidation code is written only once (for each data source) and runs at the virtualization level for all consumers and not for each individual. Otherwise, it would make the application very cumbersome.
- The connections (joins) that need to be made are made at the level of virtualization, which makes the application more efficient and the data consumers faster.

5.3 Virtualization vs Integration

The traditional way of integrating data in business intelligence project consisted in developing a data warehouse or a collection of data marts. That involved designing new data storage and using some ETL tools (Extract, Transform and Load) to get the required data from the source systems, clean them, transform them to satisfy the constraints of the destination system and the analysis requirements and, finally, load them into the destination system. Usually, developing an enterprise data warehouse takes at least several months and involves important financial and human resources, but also a powerful hardware infrastructure to support the storage and the processing in seconds of terabytes of information. Data integration has also some specific advantages: data quality and cleaning, complex transformations.

Data virtualization software uses only metadata extracted from data sources and physical data movement is not necessary. This approach is very useful when data stores are hosted externally, by a specialized provider. It allows real-time queries, rapid data aggregation and organization, complex analysis, all this without the need of logical synchronizations or data copying. Data virtualization is recommended especially for companies that need a rapid solution but does not have the money to spend for consultants and infrastructure needed by a data warehouse implementation project. Using data virtualization the access to data is simplified and standardized and data are retrieved real-time from their original sources. The original data sources are protected as they are accessed only through integrated views of data.

However, data virtualization is not always a good choice. It is not recommendable for applications that involve large amounts of data or complex data transformation and cleaning, as those could slow down the functioning of source systems. It is not recommended if there is not a single trusted source of data. Using unproven and uncorrected data can generate analysis errors that influence decision making process and can generate important losses for the company. But, data virtualization can also complement the traditional data warehouse integration. Here are some examples of ways of combining the two technologies [38] whose practical application has proved to be very valuable:

a. Reducing the risk of reporting activity during data warehouse migration or replacement by inserting a virtual level of reporting between data warehouse and reporting systems. In this case, data virtualization makes it possible to continue using data for reporting during the migration process. Data virtualization can help reducing costs and risks by rewriting re-port queries for data virtualization soft-ware instead of old data. When the new data source is ready, the semantic layer of data virtualization tool is updated to point the new data warehouse.

b. Data preprocessing for ETL tools as are not always the best approach for loading data into warehouses. They may lack interfaces to easily access data sources (e.g. SAP or Web services). Data virtualization can bring more flexibility by developing data views and data services as inputs to the ETL batch processes and using them as any other data source. These abstractions offer the great advantage that ETL developers do not need to understand the structure of data sources and can be reused any time it is necessary. Virtual views and data services reuse lead to important cost and time savings.

c. Virtual data mart creation by data virtualization which significantly reduce the need for physical data marts. Usually, physical data marts are built around data warehouse to meet particular needs of different departments or specific functional issues. Data virtualization abstracts data warehouse data in order to meet consumer tools and users integration requirements. A combination of physical data warehouse and virtual data marts could be applied to eliminate or replace physical marts with virtual ones, such as stopping rogue data mart proliferation by providing an easier, more cost-effective virtual option.

d. Extending the existing data warehouse by data federation with additional data sourcing, also extending data warehouse schema. Complementary views are created in order to add current data to the historical data warehouse data, detailed data to the aggregated data warehouse data, external data to the internal data ware-house data.

e. Extending company master data, as data virtualization combines master data regarding company's clients, products, providers, employees and so on with detailed transactional data. This combination brings additional information to allow a more comprehensive view of company activity.

f. Multiple physical data warehouse federation as data virtualization realizes logical consolidation of data warehouses by creating federated views to abstract and rationalize schema designs differences.

g. Virtual integration of data warehouse in Enterprise Information Architectures, which represents the company's unified information architecture. Data virtualization middleware forms a level of data virtualization hosting a logical scheme that covers more consolidated and virtual sources in a consistent and complete way.

h. Rapid data warehouse prototyping as the data virtualization middleware serves as prototype development environment for a new physical data warehouse. Building a virtual data warehouse leads to time savings compared to the duration involved in developing a real data warehouse. The feedback is quick and adjustments can be made in several iterations to complete a data warehouse schema. The resulted data warehouse can be used as a complete virtual test environment.

So, data virtualization represents an alternative to physical data integration for some specific situation, but can always come and complement the traditional integration techniques. Could these solutions be somehow combined in a single one, so we get both sets of advantages? The answer is positive, if there was a way of using the semantic model from data virtualization for applying ETL quality transformations on real time data. This would mean a single integration toolset integrating both virtualization and data integration capabilities. This is the solution that integration vendors are reaching after.

5.4 Data Virtualization Requirements

The Data Virtualization requirements are described below:

Model Simplicity, Visual Manipulations: The data model should be simple, based on concepts that people understand well, such as entities and attributes (e.g. the name of a customer and the amount of a bill). It should not go beyond these, yet it should also model relationships among entities in an intuitive and transparent manner. It should represent all these [37] in a graphical manner, so users can visually add/identify/select entities and attributes. A model that is amenable to visual manipulations both in terms of schema management and in terms of query formulation is of paramount importance for data scientists, data protection officers and data contributors.

Schematic Agility/Flexibility: Data sources become rapidly available and unavailable in modern analytics environments. Data engineers should be able to easily and quickly incorporate parts of these in a virtual schema without significant semantic effort. It should also be easy to modify the virtual schema (update/delete) with no major implications to the rest of it.

Standardized, Consistent Data Sharing: In real-world environments people need to share parts (columns) of spreadsheets, flat files, json documents or relations. In most cases, this is done manually, in an ad hoc manner, by exporting to a text file and moving this file around. There is no principled way to describe what someone shares in an intermediate representation, unless if imported to a data warehouse. A virtual model should make this process easy, quick and semantically clear. It should serve as the medium for data sharing in a standardized, collaborative, distributed manner.

Easily Expressing Dataframes: Data scientists usually form dataframes in Python, R or Spark. A dataframe is a table that is built incrementally, column-by-column, usually around an entity. The first column(s) is usually the key of the entity (e.g. customer ID) and the remaining columns are related attributes. An attribute can be processed (aggregated, filtered or transformed via a user-defined function in Python or R) before 'attached' as a column to the dataframe. A dataframe usually serves as input to learning algorithms or for ad hoc reporting. It is important to facilitate this process in a simple and intuitive, visual, manner. One should easily select (possibly along a path) attributes, express conditions and define transformations, using built-in methods or plug-in functions in some programming language of his/her choice (polyglotism) to form a dataframe. Query evaluation should be efficient and based on a theoretical framework, possibly similar to relational algebra.

Model Polymorphism: One size does not fit all and a virtual schema has to be easily concretized to different logical data models. While in a traditional database design the data model is predefined and determines storage models, in a conceptual design one can create database instances in different logical data models (e.g. relational, semi-structured, multi-dimensional, etc.) These database instances can be handed to data contributors or queried by data scientists via the native query language of the model (e.g. SQL, Mongo queries, MDX).

A Linkable/Crawlable Model: In many cases, especially in today's data-rich/open data era, data engineers want to make schema (or parts of it) available to external users, or link it to other schemas external to the organization. This 'schema sharing/linking' should be easy, requiring no or little semantic effort to do it (e.g. via a link). Graph-based models (e.g. the web) is a representative example for this.

5.5 Data Virtualization Tools

There are many ways to implement the idea of data virtualization but the most appropriate one that can support analytics functions in the context of business intelligence (and not just ad hoc queries that require little computing power) is the data virtualization server. At the moment, there are several servers [39], such as:

Data Virtuality

Platform: Data Virtuality Platform

Data Virtuality accesses, manages and integrates any database and cloud service by combining data virtualization and extract, load, and transform (ELT) processes. The company offers data pipeline solutions in two iterations (self-service and managed), and Logical Data warehouse, a semantic later that allows users to access and model data from any database and API with analysis tools. Data Virtuality connects to more than 200 data sources and offers a number of data replication features based on use case.

Data Virtuality is a great fit for organizations with large data sets that cannot be easily virtualized with other solutions. Instead of just providing access through a virtual layer, the system can also replicate large data sets for faster query performance.

Key values:

- Data across multiple sources can all be queried in a standardized way using SQL (Structured Query Language).
- Creates a logical data model from different data sources including databases and other sources of data such as Google Analytics.
- Connected data sources can be made available and combined via a Business Intelligence tool front end.
- For large data sets, Data Virtuality has an integrated analytical database that can replicate data, providing a faster query rate and better overall performance.
- Works as a Java based application server that can be hosted in the cloud or operated on-premises.
- The ability to very precisely pull data from an application, for example custom fields in Salesforce is seen as key differentiator by some users.

Denodo

Platform: Denodo Platform

Denodo is a major player in the data virtualization tools marketplace. Founded in 1999 and based in Palo Alto, Denodo offers high-performance data integration and abstraction across a range of big data, enterprise, cloud, unstructured and real-time data services. Denodo also provides access to unified business data for business intelligence, data analytics, and single-view applications. The Denodo Platform is the only data virtualization solution to be provisioned as a virtual image on Amazon AWS Marketplace.

It is one of the best choices for organizations of any size looking to not just virtualize their data, but also understand what data they have.The data catalog feature that has emerged in the latest version of Denodo is a powerful feature for data virtualization users, offering the ability to not just combine and virtualize data but to identify and catalog data.

Key values:

- Data catalogue feature in Denodo 7 enables organizations to use a semantic search query capability to find data as well as providing insights into how data is used by other users and applications.

- Parallel processing with query optimization capability minimizes network traffic load and can help to improve response times for large data sets.
- Integrated data governance capabilities are also particularly useful as they can help organization to managed compliance and data privacy concerns.
- For data usage, end users can choose to use SQL or other formats, including REST and OData to consume and secure data.

IBM

Platform: IBM Cloud Pak for Data

IBM offers several distinct integration tools in both on-prem and cloud deployments, and for virtually every enterprise use case. Its on-prem data integration suite features tools for traditional (replication and batch processing) and modern integration (synchronization and data virtualization) requirements. IBM also offers a variety of prebuilt functions and connectors. The mega-vendor's cloud integration product is widely considered one of the best in the marketplace, and additional functionality is being rolled out on a perpetual basis. For organizations looking for a converged solution that handles data collection and analysis, IBM Cloud Pak for Data is a good choice.

Key values:

- IBM Cloud Pak for Data is an integrated platform that enables organization to both collect and analyze data with the same platform.
- The core-organizing concept for workflow is project based, with sophisticated controls for access and data governance for each activity.
- The user experience includes a drag and drop interface to connect data and perform complex ETL (Extract, Transform and Load) jobs to prepare data for analysis.
- IBM has also integrated an enterprise wide data catalogue to help users organize and identify the data they want to collect and analyze.

Informatica

Platform: Informatica PowerCenter

Informatica's data integration tools portfolio includes both on-prem and cloud deployments for a number of enterprise use cases. The vendor combines advanced hybrid integration and governance functionality with self-service business access for various analytic functions. Augmented integration is possible via Informatica's CLAIRE Engine, a metadata-driven AI engine that applies machine learning. Informatica touts strong interoperability between its growing list of data management software products.

For organizations looking for a leading data virtualization tool with integrated data quality tools, PowerCenter is a solid choice. PowerCenter is consistently rated as a top data integration tool from analyst firms for its powerful set of features.

Key values:

- Ease of use is a key attribute of PowerCenter, with a no-code based environment that uses a graphical user interface (GUI) to integrate nearly any type of data.
- A key element of the platform is the metadata manager which goes beyond just integrating data, to helping users with a visual editor that creates map of data flow across an environment.
- One of the most useful functions of PowerCenter is the impact analysis feature, which can identify the impact to an enterprise about a data integration effort before any changes are actually implement.
- It can happen that a data integration or data virtualization activity breaks something, which is why PowerCenter's data validation capabilities are so critical. The data validation capability is there to make sure that data has not been damaged by a move or transformation.
- Data archiving is another really useful feature, that enables organizations to move and compress data out of older applications that are not actively used.

Oracle

Platform: Oracle Data Service Integrator

Oracle offers a full spectrum of data integration tools for traditional use cases as well as modern ones, in both on-prem and cloud deployments. The company's product portfolio features technologies and services that allow organizations to full lifecycle data movement and enrichment. Oracle data integration provides pervasive and continuous access to data across heterogeneous systems via bulk data movement, transformation, bidirectional replication, metadata management, data services, and data quality for customer and product domains.

For those organizations that are already making use of other Oracle applications for data storage and analytics, Oracle Data Service Integrator is an obvious and easy choice to make for data virtualization.

Key values:

- Among the unique features is that Oracle Data Service Integrator can both read and write data from multiple sources, enabling this tool to be used for a variety of different use-cases.
- Service integration is done via a graphical modelling capability, so users do not need to code the integration manually.
- Security is a key feature with rules-based security policies to make sure that certain data elements are protect, or redacted to help meet various privacy and compliance needs.

- Going a step further on the security side, Oracle has also integrating sophisticated auditing, that keep track of users and what data was accessed and when.
- Real-time access to data is a particular strength, thanks to optimized query and data patch technology that Oracle has baked into this offering.

SAP

Platform: SAP HANA

SAP provides on prem and cloud integration functionality through two main channels. Traditional capabilities are offered through SAP Data Services, a data management platform that provides capabilities for data integration, quality, and cleansing. SAP's Cloud Platform integrates processes and data between cloud apps, 3rd party applications, and on-prem solutions.

Key values:

- Improve the quality of your data from standardization to enrichment with built-in data quality capabilities and services. Transform your data format from source to target. You can standardize, cleanse, identify, and correct duplicate records; enrich address data with geocode intelligence; and manage other data quality issues across domains and sources in a single user interface.
- Provide transparent, real-time data from other systems on demand. Give information workers instant access to information by federating queries on remote data sources including cloud-native remote sources, Hadoop, SAP Adaptive Server Enterprise (SAP ASE), and other databases. You can also retrieve relevant answers without the cost and effort of migrating data.
- Acquire business intelligence from any data source, for an integration option of your choice. Enable three integration methods: ETL, replication, and federation by moving data from any source, such as SAP or third-party databases, and seamlessly replicate data so it is always available in real-time. Moreover, open and extensible support is available to support your data volumes, data types, and data sources.

SAS

Platform: SAS Federation Server

SAS is the largest independent vendor in the data management marketplace. The company's main goal is to build a data quality platform that allows users to improve, integrate, and govern enterprise data. SAS Data Management can ingest data from legacy systems and Hadoop, and create rules once and reuse them. In addition, users can update data, tweak processes, and analyze results themselves. A built-in business glossary as well as third-party metadata management and lineage visualization capabilities allow for collaboration.

Key values:

- Federated data as a service (DaaS) simplifies data access and allows web-based query submission, server and database discovery via a robust REST interface.
- Shared metadata allows for easier integration with other SAS solutions.
- Web-based administration console for improved governance.
- Capability to enable or disable data caches.
- Supports data caching of materialized views to reduce loads on operational systems.
- Incorporates advanced data manipulations that can run inside the database, such as data cleansing, merging and scoring.

5.6 Data Virtualization Use Cases & Benefits

Data virtualization software has become a critical asset to any business looking to triumph the growing data challenges. With innovations like query pushdown, query optimization, caching, process automation, data catalog, data virtualization technology is making headway in addressing a variety of multi-source data integration pain points.

Here are a few data virtualization use cases and applications that show how it is helping businesses:

BI and Analytics

Business intelligence (BI) and data analytics are front and center for every successful company. In fact, with the growing use cases of the IoT, predictive analytics, and machine learning, data analytics often make the difference between success and failure. Besides predictive analytics based on historical data, real-time analytics has become very critical for companies that depend on real-time data to operate their day-to-day business. Electricity companies might focus on the behavioral analysis of customers, but if they are expanding overseas, they might consider cloud based analytics will be more important. While many energy companies are engaging in advanced data analytics, here are a few important considerations to keep in mind:

- Unstructured data is playing a larger role in data analytics than structured data. A contextual and prescriptive data analytics platform requires seamless integration of structured and unstructured data in real-time.
- Storing and replicating big data for analytics is both expensive and time consuming. As the role of data, scientists become more important, and scientists need sandbox environments for their analysis, physical data movement becomes a bottleneck.
- The current challenges of cloud based data analytics remain an issue, as many companies need both on-

premises and cloud based data for analytical purposes. Very few tools today can seamlessly integrate on-premises and cloud data in real time, to create a hybrid, integrated data environment.

- Securing the data across applications is another important problem to solve. With the explosion of streaming data, big data, cloud data, and more, it is not always clear which data steward is responsible for the security of which part of the enterprise data.
- Leveraging data virtualization, there are many virtualization platforms that combine the widest range of structured, semi-structured, and unstructured data without replication, making them the ideal data integration layer to feed visualization and analytical tools. They can easily integrate real-time streaming, social media, and Web data with data from legacy systems to enable analytics in the cloud, on-premises, or in a hybrid environment. Such versatility makes them an easy choice for all advanced analytical needs, enabling companies to gain competitive advantage.

Investments in the Internet of things (IoT), deep learning, machine learning, and big data analytics will not provide a competitive edge unless companies can gain highly contextual, actionable insight from those investments. Data virtualization is essential for delivering those actionable insights via:

- The real-time integration of all on-premises and off-premises data without creating physical replicas of the data.
- Self-service capabilities for business users, so they can be independent from the IT team for discovering day-to-day business insights.
- Augmenting raw data with context from social media, Web, enterprise systems, and other data sources.
- A semantic layer, making the trove of enterprise data meaningful to business users.
- Enhancing the speed of data and keeping business continuity, while creating room for continuous improvement.

Big Data

All businesses now have one thing in common; they all are data businesses. Every organization is trying to make use of their big data and streaming data by turning those into information and knowledge to fuel business growth.

Volume, variety and velocity of data continues its upward movement and seems to stay the same for the foreseeable future, based on proliferation of internet connected devices, web platforms and trends such as cognitive science, machine-learning and IoT. Adding to the complexity, IT wants to empower all big data scientists and business users through self-serviceable analytical and reporting platforms.

While big data offers a lot of promises to business growth across industries, it comes with a set of challenges that are consistent across all shapes and sizes of organizations.

- Big data is no more synonymous only with Hadoop. Spark, Hive, Presto, Kafka, Impala is crowding the big data and streaming data storage and query space. Heterogeneity brings information inconsistency among various business units within any organization, as each user group have their own wish list from big data analytics.
- Data privacy and data security concerns are more pronounced and prominent with the rise of big data. More silos of big data means more separate data privacy and data security requirement per silo.
- There is a huge proliferation of consuming applications over the years and a lot of them do not interact very well or at all with various sources of big data or streaming data.

Data virtualization technology provides an agile and cost-effective approach to combining, governing, and managing big data, and to overcoming the inherent challenges presented by big data silos. We call it big data virtualization. There are three most popular use cases of big data virtualization:

Logical Data Lake: Data virtualization bridges one or more data lakes along with traditional data warehouses, MDM systems, cloud sources and beyond. This use case improves enterprise functionality of data lakes by providing additional context with data from other enterprise sources.

Data Warehouse Offloading: Data virtualization offloads less frequently used or cold data from enterprise data warehouse to a Hadoop cluster to free up expensive enterprise computing resources.

IoT Analytics: Data virtualization combines streaming data with other sources of enterprise data to make streaming data more meaningful and useful for business users.

Virtualizing and combining big data along with other sources of enterprise or cloud data offers many benefits, so that organizations can truly reap the benefits of big data:

- Reduces expensive big data replication across the organization, at the same time offering significantly faster time-to-market.
- Creates consistent data governance, privacy and security structure across wide range of systems, both on-premise and cloud.
- Offers flexibility and agility to big data and IoT analytics by offering easy connectivity across a broad range of source and consumer systems.
- Simplifies information creation and consumption model by creating an abstraction layer so that business users are separated from underlying complexity.

Cloud Solutions

Transitioning to the cloud is never easy, whether you are engaged in cloud modernization, cloud analytics, hybrid cloud deployment or other cloud initiatives. Often, the biggest challenge is data integration in the cloud. Data integration is difficult across standalone SaaS applications and other legacy or modern data repositories, which can lead to lack of consistent information, which could in turn impede your marketing team in doing a stellar job at lead generation or lead nurture, or prevent your sales team from selling the right product to the right prospect, or cause your operations team to struggle with supply chain issues.

These are the primary three challenges to successful data integration in the cloud:

Data Silos: Your marketing team might draw data from marketing automation, CRM, collaboration tools and social media, while your sales team might source data from your deal management system, your call center, or your activity management tools in addition to the CRM tool. While moving through this unsynchronized set of applications, data becomes outdated and unusable.

Security: Users want to be able to access any application from any device and browser, but companies must balance ease of access with concerns about data breaches, violations of regulatory constraints, and unauthorized access without the protection of a unified security or data governance model.

Network Latency: As applications multiply and one-to-one communications between those applications proliferate through your network, the user experience suffers.

Data Virtualization provides an easy-to-deploy cloud solution that seamlessly integrates all SaaS, on-premises, or other cloud sources, in real-time, in the cloud. Virtualization Platforms enable you to apply a unified security setting across your entire infrastructure. Here are some benefits of cloud solutions:

- Integrate all enterprise data in real-time without any replication and provide all business users with business-critical information on the go.
- Enable migrations from on-premises to cloud sources, with no impact on day-to-day operations.
- Manage cloud and on-premises security from a single point.
- Track cloud usage by department or individual.
- Reduce the inherent latency of cloud access using the Virtualization Platform’s advanced caching capabilities.
- Scale cloud solutions as needed with a flexible pay-as-you-go model.

Data Governance

Data governance defines how enterprises manage availability, usability, integrity and security of their data with a set of predefined rules and processes in place. Data governance has never been more important. With the advent of social media, Web, big data, and cloud technologies, petabytes of data is scattered all over the place, both on-premises and off-premises. Without proper visibility and control of your organizational data, your business can not only risk revenue and productivity loss but also face existential crisis. Specifically, Industries such as electric utilities are constantly under strict government scrutiny and audit to ensure that customer data is private and protected. To make the situation even more complex, geographical regions are now setting up their own data protection laws, such as GDPR. Data governance and privacy will grow exceedingly complex, unless businesses begin to take action starting now.

Setting up a well-controlled data management framework and process is important for each organization across the globe, just as it is critical to establish data stewards with proper ownership over the data. However, mid-to-large scale organizations face many challenges with regard to setting up a centralized, well-controlled data management framework.

Fragmented data governance: As companies grow and datasets diversify, the number of on-premises and cloud tools and systems proliferates. Each siloed data source or consuming system has its own data governance and data security mechanism, and none are shared across the enterprise.

Data inconsistencies across departments: Many departments within mid-to-large scale organizations use their own sets of tools and systems to access the same information, which leads to problems with analyzing data lineage as well as information inconsistencies.

Data access control: With too many siloed tools and disparate data access rules spread across systems, it becomes impossible to lay out proper data access rules for internal and external users.

Geographical challenges: Almost every company that is spread across more than one country replicates its data multiple times, increasing cost, affecting data lineage and quality, and possibly violating regional data protection rules.

Data virtualization enables organizations to create central data access, data governance, and security policies across heterogeneous systems of structured and unstructured data sources.

Data virtualization is an agile, flexible data integration technology that can help organizations address the growing challenges in data governance, security, and compliance. Data virtualization can enable:

Consistency in policy enforcement: Integrate the fragmented data sources across internal and external systems and enforce consistent policies of data access and data security.

Ease of cross border information sharing: Data virtualization enables multinational companies to facilitate data integration and regional policy enforcement across borders, without moving or replicating any data.

Error free data lineage analysis: For industries that depend on regulatory compliance for success, data virtualization is becoming the most critical component of their enterprise data architecture.

Data Services

Data is gold. Often, it is a company's most valuable asset. But to get the most out of their data troves, especially when these massive volumes are comprised of myriad heterogeneous sources, companies need to establish robust services to support data delivery and consumption. Such services might include:

- Data-as-a-service (DaaS) initiatives, which present business users with data that is curated, cleaned, and organized to meet business their needs, or data services marketplaces, which deliver any type of data, even transactional data, to business and technical users over a unified interface.
- Internet of things (IoT), online transaction processing (OLTP), and data analytics. They become increasingly vital for many companies to leverage in order to gain the maximum value from data.

Data services are gaining traction, but organizations need to overcome a few key challenges to deploy them successfully:

- Data needs to be processed in real time while IT is getting burdened with a wide variety of data at high volumes, including data from social media, streaming sources, and the cloud, business users often need data the instant it changes.
- The complexity of data silos as many different forms of data sources come to life and proliferate, they tend to create their own data silos. The greater the number of silos, the harder it is to gather holistic intelligence for the enterprise.
- Costly physical data aggregation massive volumes of data from external sources such as social media platforms, the Web, and sensors, need to be stored in expensive infrastructures in order to be integrated with legacy technologies. Data also is replicated multiple times and often becomes out-of-date.
- Data security, privacy, and governance as organizations make data available to a wider range of users, it becomes of paramount importance to provide effective encryption and proper access controls for personally identifiable information (PII), as well as to ensure that they always deliver the most trustworthy, accurate data.

Most data integration technologies physically move data from multiple sources to a central repository, but data virtualization offers a novel approach. It creates real-time, consolidated views of the data without moving the data.

What is more, it integrates data from transactional systems and cloud systems in real time, which is just not possible with legacy data integration solutions.

Master Data Management

Master data management (MDM) refers to the strategies that enterprises use for establishing authoritative sources of critical data so that stakeholders can obtain a single view of key entities. For many years, MDM has been a critical, integral part of the enterprise architecture. Your organization is probably using an MDM tool to manage customer data, product data, supplier data, or all of it. MDM tools are great for delivering a single view of your data or even the relationship between customer, product, or supplier entities. But when it comes to gaining a view into the various transactions that take place among these business entities, standalone MDM systems fall short because they do not store the associated transactions. Modern enterprises need a complete view of their master data that also includes transactions and interactions from social, streaming, and other forms of modern data.

A true, complete view of master data has three components: A single view of the master data, for example knowing that 'John Smith' and 'J. Smith' are the same person. A 360° view of the master data relationships, for example knowing that 'John Smith' and 'P. Smith' are part of the same family. A view into each master data entity's transactions and interactions, for example knowing that 'John Smith' has called customer service three times in the last hour regarding a recent purchase.

Master data management systems are good at creating the first two views, but they do not store the associated transactional data.

MDM, when augmented with data virtualization, provides the complete view of the master data. The combination of technologies support three common usage patterns:

Master data management for analytics: In this scenario, the MDM system draws master data from disparate data sources, reconciles discrepancies, and creates golden records, and the data warehouse draws transactional data from a similar range of transactional sources. Data virtualization creates a unified view, composed of data drawn from both the MDM system and the data warehouse.

Master data management for operations: In this scenario, there is no data warehouse, as the data does not need to be stored for historical or analytical purposes. The data virtualization layer combines the master data from the MDM system with the associated transactional data drawn from the operational systems.

Virtual master data management: Traditional MDM systems, which store golden records in a separate database, often cannot be used in industries with heavy restrictions on data replication; industries such as public sector. In such cases, the data virtualization layer itself provides many of the functions of a traditional MDM system, while also creating a virtual view of the transactional data across myriad sources, as in the operationally focused architecture.

Data virtualization offers the following benefits, when augmented by a master data management system:

- A complete view of the master data, including a single view of the customer, a 360° view of customer relationships, and a complete view of customer transactions and interactions.
- The ability to combine master data with any other data, throughout the enterprise; data virtualization can connect to MDM systems and myriad other data sources.
- Real-time access to the complete view, for any stakeholder in the organization.
- Reduced replication and its associated costs and risks. Data virtualization provides data access without replication.
- A short implementation timeframe; a robust data virtualization layer can be developed and deployed in a matter of weeks.

6. Cases Studies

6.1 Enterprise Analytics Data Platform on AWS to generate business insights

A power generation company was experiencing difficulty in gaining value from its data and was dependent on external service providers to generate analytics insights.

The company realized there was a need to establish in-house analytics capability to enable business to make data driven decisions and leverage the vast amount of data that they already possess.

The key objectives for the project included:

- Developing a data & analytics strategy to define future analytics vision, analytics use cases, architecture and implementation roadmap
- Building a scalable and extensible data and analytics platform upon which to grow and build out company's vision
- Demonstrating value early in the development of the platform to achieve business buy-in through the delivery of use cases
- Pivot analytics focus to address rapidly evolving business needs

Accenture assembled a team consisting of Utility SMA for analytics, Utilities expert Data Scientists, AWS cloud architects, Databricks, Spark and other Big Data skill experts to build end to end data integration supply chain, create analytics model and generate business driven dashboards. The core project components include:

- Elaboration and selection of analytics use cases to generate business value
- Designing enterprise data platform reference architecture
- Technical design for enterprise data platform, data supply chain pipeline, metadata management, storage directory structure and establish governance process
- Setup the analytics environment (AWS, Databricks, Tableau Online, Collibra) and define the policies and process structure
- Building data processing pipeline for data transformation, curated 35+ data sources
- Designing and building analytics dashboards for business insights
- Designing the overall analytics use case execution framework for enterprise

Value Created

- Developing new data insights to aid traders in better articulating market dynamics and to identify and validate relationships among variables.
- Clustered customers by load profile; use to identify load management opportunities and strategies
- Addressed growing Data Privacy regulations through strict Data Governance and improved Security and Privacy controls
- Eliminated 30% to 40% of time on data preparation by analytics teams by pushing that responsibility down into the enterprise Data Hub
- Increased analytics capacity by re-focusing on data science vs. data preparation
- Established the basis for a new energy load forecasting capability
- Provide data scientists with the ability to shop for data
- Enabled data search via Data Governance framework
- Developed key COVID-19 case tracking, load trending and energy generation correlation dashboards within a week for executive review

Technologies Used

- ✓ AWS cloud platform
- ✓ Databricks managed platform
- ✓ Databricks Delta Lake
- ✓ Apache Spark, Python
- ✓ S3, SNS
- ✓ Collibra
- ✓ ElasticSearch
- ✓ Lambda
- ✓ Visualization: Tableau Online
- ✓ Athena
- ✓ Security: AWS KMS

Solution Highlights

- ✓ Reduce storage through removal of "n" copies of data, build an enterprise unified data platform as single source of data and cohesive storage of multi-faceted data
- ✓ Enabling framework to support streaming analytics, censor data and data consumption

- ✓ Configuration driven data transformation rules for different set of disparate data sources
- ✓ Standardized metadata and physical data attributes to match with their business taxonomies
- ✓ Provision fit for purpose data and integrate with downstream applications using scalable data distribution services
- ✓ Generate analytics insights using in-house platform
- ✓ Enabled data analytics readiness check

6.2 Enabling Self-Serve Analytics using Big Data Platform for Major Utilities Company

A utilities company established an analytics environment, enabled through big data technologies to deliver capabilities and services that:

- ▪ Established the single source of data
- ▪ Empowered different business units with self-serve analytics capability using a governed data lake storing data that they can trust

The company had no single source of truth for reporting and analytics. It was also plagued with various siloed analytics efforts by different business units. Traditional customer data warehouse (CDW) lacked sufficient integration with different source systems, consisted of legacy stored procedures which were difficult to maintain and rebuild, posed a scalability challenge with regards to storage, and a definitive lack of governance. In addition to this, owing to a regulatory requirement of implementing a new Time of Use (TOU) rate, the company wanted dashboards to provide visibility into:

- ▪ Marketing & Communications
- ▪ Customer Enrollment & Tracking
- ▪ Customer Contacts
- ▪ Customer Extended Attributes
- ▪ Customer Billing and Billing Exceptions
- ▪ Master Data Exceptions
- ▪ Work Distribution for Billing Operations Management team to manage exception bills, assignment and resource management

Accenture's Role and Solution

- ▪ Assessed the Hortonworks Data Platform (Hadoop distribution) deployed on client's environment for
- ▪ Production, QA and Development clusters
- ▪ Recommended setting up Hadoop clusters with needed ecosystem tools and designed integration patterns with required target systems
- ▪ Designed analytical data models, and defined data movement on the Hadoop distribution
- ▪ Completed one time historical data load of required source system entities into Hadoop data lake
- ▪ Designed wireframes pertaining to dashboard requirements

The solution achieved the following**:**

- ▪ Enabled self-serve analytics in a big data environment aligning with industry best practices
- ▪ Defined a roadmap for integration of new source systems and future state capabilities with respect to new data entities
- ▪ 7 Individual dashboards with 340+ individual graphs and widgets supporting dashboard requirements

Value Created

- ▪ Delivered a self-service analytics platform that effortlessly scales and provides a single source of data and analytics.
- ▪ The solution enabled the client to standardize a canonical enterprise data model, data repository and improved the ingestion of data from internal and external sources.
- ▪ Governed data lake provided intuitive graphical data lineage increasing trust on data, efficient data discovery for analytics users and data scientists
- ▪ Offloading enough processing from expensive proprietary commercial platforms is expected to have a long-term effect on total cost by cutting down licensing and infrastructure expenses.
- ▪ Reduced the time of discovering Customer Billing Exceptions from several days to a daily refresh with an accuracy of 98% (dashboard widgets)
- ▪ Helped improve marketing campaign execution for ~1.3M customers based on insights from the Marketing & Communications dashboard reducing campaign execution errors by 80%
- ▪ Master Data exceptions revealed insights on data sync issues between the billing and CRM systems

Technologies Used

- ✓ Hadoop distribution (Hortonworks Data Platform)
- ✓ Storage: HDFS, HBase
- ✓ Processing: Spark, Tez
- ✓ Hadoop Data warehouse: Hive
- ✓ ETL pipeline & scheduling: Talend Big Data
- ✓ RDBMS: PostgreSQL
- ✓ Governance: Apache Atlas
- ✓ Monitoring: Ambari, Grafana
- ✓ Data Science: Zeppelin
- ✓ Security: Kerberos, Knox, Ranger
- ✓ Visualization: SAP Lumira

Solution Highlights

- ✓ Provided a single source of data and cohesive storage of multi-faceted data into a single repository
- ✓ Efficient use of Hive table properties based on identified data access patterns enabled low-latency querying capability
- ✓ PostgreSQL was used as a staging layer for data marts supporting quick slicing and dicing by business users on Lumira dashboards

- ✓ Atlas enabled data discovery for data scientists and business users, effectively speeding up reporting and insights as well as providing lineage of data entities
- ✓ Push methodology (flat files) for ingestion enabled stronger trust on data as source system SMEs have the functional & technical knowledge of the data they own

6.3 Gotland's Smart Grid

Delivering climate-smart energy [40] and future proofing the electricity grid for 60,000 customers in Gotland, Sweden is the purpose of this project. This case study explains Vattenfall's work on the island of Gotland. Vattenfall upgraded the electricity grid to provide increased capacity for new renewable energy generation, incorporated large-scale energy storage, monitoring and self-healing networks and a new energy services marketplace as part of a comprehensive smart-grid.

By 2010, Gotland's aging electrical network was not able to keep up with the growing amount of wind power being delivered into the grid. The island had an aging high voltage DC connection with mainland Sweden, which wasn't built for constant import and export, and often caused a total power failure across the whole of Gotland.

In 2011 the Swedish Energy Authority tasked Vattenfall with addressing Gotland's energy requirements and, over a six year period, the company upgraded and evolved the island's electrical infrastructure into a modern, reliable smart grid with a road-map for future development through to 2035. A plan was developed which would see large-scale energy storage deployed, alongside a new energy services trading platform, used to align supply and demand.

Part of the proposed solution for Gotland's electricity grid involved an energy storage system to work alongside the variable wind power generation and local electricity demand.

The task of the energy storage system is mainly to bridge short-term faults, by means of frequency control or for short-term island operation. During normal operation, the energy store is used for three purposes in the following order of priority:

- Being prepared for frequency control in the event of short-term island operation or to support the remaining cable
- Reducing the number of polarity changes so that ageing of HVDC cables is reduced
- Solving capacity problems in the marketplace

In the event of a fault in any of the cables on the island, there was a risk of an imbalance between generation and consumption, which could cause a total power outage. In order to reduce the risk of an imbalance, an energy storage system that could supply a capacity output in the range of 25–50 MW, with an energy storage supply of 25 MWh to cover the time required to remedy a fault, was proposed. The energy storage facility, in conjunction with variable wind energy generation and controllable electricity consumption, would create an opportunity for an open marketplace for energy services.

Vattenfall also introduced monitoring on the low voltage grid to provide real-time energy updates to customers. Reclosers were installed as an alternative to new grid reinforcements, which isolate sections of a power line that is experiencing a fault or overload and minimize the number of customers without service.

Security of supply was also increased by the incorporation of remote controlled breakers, managed via a distribution management system (DMS) which assess faults, searches for alternative routes within the network and then actions the required changes to restore power, providing a self-healing network.

In order to double the renewable electricity capacity on Gotland, the existing 70kV electricity grid on the island would have needed to be strengthened and replaced with a higher voltage, 130kV grid. Upgrading the entire grid on the island would have been prohibitively expensive so alternative solutions were required.

As a first step to incorporate more wind power into Gotland's grid, an additional 80MW of renewable generation was made possible by introducing active network management of the grid, bringing the maximum export capacity up to 130MW. By repurposing the existing infrastructure, full redundancy was maintained and active control of the grid avoided the increased costs of building a larger energy storage facility.

In order to cater for growing demand, it would be possible to increase wind generation by a further 70MW with the addition of a larger 50MW energy storage facility. But, any system faults at this level of power consumption would require energy production to be managed through frequency control in order to maintain a balance between electricity generation and consumption. If a cable is disconnected in the event of a fault, an energy storage system of 25MW can handle demands for 95% of the hours in the year. With a 50MW energy storage, demands could be managed over 99% of the hours in the year.

Local energy marketplaces introduce dynamic pricing and encourage energy generators and consumers to regulate their production and consumption to help stabilize the grid. A local energy marketplace on Gotland is being designed to create a flexible solution to balance supply and demand on the island in both normal and disrupted operation.

Gotland's Smart Grid - Results:

- Increased grid capacity
- Improved quality in rural grid
- Improved customer satisfaction
- Significant reductions in downtimes
- Cost efficient vs new networks
- Self-healing network testing

Smart Grid- Recommendations: The size of an energy storage system should be scalable with a capacity range in order to cover operational disruptions of up to one hour. Smart grids should introduce requirements for frequency control for new electricity generation connections in grid regulations and connection agreements. Introducing a local energy marketplace provides a cost-effective method to utilize the existing electricity grid.

6.4 ACON Smart Grid

The ACON (Again Connected Networks) [41] Smart Grids project refers to the effort that will deepen and facilitate the cross-border cooperation between the Czech Republic and the Slovak Republic at the Distribution system operator (DSO) level.

The main objective of the ACON SG project is to improve the existing power distribution grid primarily in the border areas of both countries concerned, but the project activities will also impact on other parts of project promoters' distribution areas. This will create greater capacity for the development and connection of distributed electricity production and adequate space for possible connection of new distribution grid users in the region. Moreover, the distribution grid will be modernized through implementation of smart elements and new IT framework in order to create the "smart grid" energy network within the project impact area.

Above-mentioned goals will be fulfilled primarily through modernization and reinforcement of the existing 110 kV as well as the 22 kV voltage level interconnections. Medium voltage level interconnection has a supportive effect on the security of supply of the adjacent regions in the border areas. This is important particularly during outage incidents on the infrastructure of the transmission level. The base of the project will be built on the historical interconnection and cooperation between both countries.

The project consists of six major activities:

1. Cross-Border Interconnection Improvement
2. Management of Distribution Grid in new Conditions
3. Distribution Grid Communication Elements
4. Smart Grids IT Solutions
5. Communication and Dissemination
6. Action Management

The project realization consists of several sub-activities, which aim to implement smart elements into the current distribution grid, such as:

- Border areas and cross-border connections improvement with focus on improvement of existing distribution grid in the border areas of the Czech Republic and Slovak Republic, which will include operation change of 2x110 kV High Voltage level interconnection, together with reconstruction and automation of Medium Voltage feeders and construction of new 22 kV lines connecting Holíč (SK) and Hodonín (CZ) substations. Taking such steps will create technical backup to enable cross-border cooperation in case of accidents or other safety threatening operational situations.
- Construction and improvement of existing distribution grid backbone leading to increased reliability of electricity supply and more flexible connection of additional points of delivery.
- Deployment of technologies improving the reliability of electricity supply, leading to increased added value of applied equipment.
- Applying advanced communication and diagnostic methods with the aim to increase the convenience of customers receiving energy services.
- Deployment of distribution grid communication (smart) elements which will enable more efficient management of distribution grid through remote access, transmission of data regarding failures, information on system load and remote switching (deployment of optical cables on selected existing overhead lines and implementation of GPRS (LTE) and BPL communication technology).
- Implementation and integration of smart grids IT solutions, which will allow DSOs to gather larger volumes of data as a major enabler for more accurate data analysis and more addressed decisions towards the requirements of distribution grid during the whole lifecycle of distribution assets (direct impact on distribution grid management, process management, optimization of distribution grid operation or distribution grid maintenance and renewal planning).

6.5 Establishing the Smart Grid in Austria

This case study [42] encompasses the various phases of LINZ NETZ smart metering project using the NES System. The initial phases of the project focused on providing a variety of smart metering benefits including flexible tariff models, improved billing process, customer availability of energy consumption data, energy conservation, remote connections and disconnections, ripple control replacement, load management, and street lighting management.

LINZ NETZ GmbH (LN) is a distribution network operator in the greater Linz area and parts of Upper Austria and is a legal entity in the Linz AG group of companies. LINZ NETZ is responsible for the future-proof expansion, operation and maintenance of its electricity and natural gas distribution network within its officially defined supply area in Austria. The electricity network covers 1,652.5 km^2 of supply area and supplies approximately 440,000 people in 81 Upper Austrian municipalities, as well as the city of Linz. The energy is distributed to customers via 27 substations along with 3,000 transformer stations. LINZ NETZ has about 284,000 meters installed.

In 2007, LINZ NETZ made a decision to deploy smart meters with PLC communications for remotely reading measurements, disconnecting and reconnecting service, and load switching.

They decided to move forward with smart metering, because they wanted to improve their business processes and recognized, that regulations were changing and would require smart metering functionality. The procurement was carried out in several stages based on a public tender. S&T Smart Energy GmbH (formerly Ubitronix) emerged as the best bidder for the project, providing smart meters, data concentrators (DCNs) and Head-End System (HES) from Networked Energy Services (NES) as well as load switching devices from S&T Smart Energy. It was also provided a Meter Data Management (MDM) for connecting the HES to the Enterprise Resource Planning System from LINZ NETZ, which is implemented with software from SAP. In addition to the MDM, downstream systems and work force management system, LINZ NETZ also took responsibility for planning and installation activities. Communications between the NES Smart Meters and NES DCNs uses power line communications (PLC) based on Open Smart Grid Protocol (OSGP), and communications between the NES DCNs and the NES Head-End System (HES) utilize a combination of mobile communication 2G/3G and Fiber-Ethernet.

The NES smart meters securely record and save energy consumption in the meter up to every 15 minutes including a storage period of 60 days maximum based on legal requirements. The readings are used with consent of the customer to fulfill obligations from a delivery contract. Consumption values and related information are transferred into the SAP system, which is used for automatic billings.

Other utility systems (Network Information System) receive relevant meter data. For example, LINZ NETZ can view where the meters are installed, as well as the meters' current status/data. There is a Ticket/Event system, which receives and shows meter events.

This information is used to assign technicians when needed to go visit a customer and fix a problem. In addition, power quality measurement data and meter event information are also provided by the meters. Based on the present and future legal requirements, the importance of this operational information continues to grow.

LINZ NETZ is using the NES System along with Load Management Module to switch a portion of the street lighting and different devices, such as Warm Water Heating Boilers and Floor Heating. They are able to define schedules for various times of the year for each feeder line, and they can also reconfigure the switching schedule when needed. Using centralized rule base and general switching LINZ NETZ is able to manage the consumption for an overall street.

The NES System is integrated with the end customer using a web portal to help consumers obtain energy consumption data and manage their consumption. Utilizing a connection to NES smart meters via their local interface (MEP - multi-purpose expansion port), consumers can have access to consumption values in real time. This functionality is used to engage customers more and helps them view and control their energy consumption. The MEP interface also enables sub-metering of non-electricity meters (e.g. gas, heat and water). And the ripple control system has been replaced by the S&T Load Management Modules and the capabilities within the NES System.

The NES System (fig.16) is based on a 3 tier architecture: the solution includes utility data center software, field distributed application control nodes and grid devices and sensors, such as single phase, poly phase and CT smart meters, and Open Smart Grid Protocol (OSGP) compliant communication devices enabled by Control Point Modules. Middleware goes to the SAP enterprise integration layer. All software architecture/web services on all layers are integrated in a very easy way.

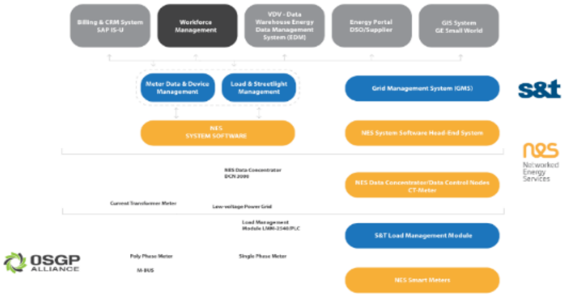
Billing & CRM System
SAP IS-U
Workforce Management
VDV - Data Warehouse Energy Data Management System (EDM)
Energy Portal DSO/Supplier
GIS System GE Small World
Meter Data & Device Management
Load & Streetlight Management
Grid Management System (GMS)
s&t
NES SYSTEM SOFTWARE
NES System Software Head-End System
nes Networked Energy Services
NES Data Concentrator DCN 3000
NES Data Concentrator/Data Control Nodes CT-Meter
Current Transformer Meter
Low-voltage Power Grid
Load Management Module LMM-3540/PLC
S&T Load Management Module
OSGP ALLIANCE
Poly Phase Meter
Single Phase Meter
M-BUS
NES Smart Meters

Figure 16: The NES System Architecture

6.6 Agder Energi Smart Grid

Due to the cold weather in the Nordics, Microsoft [43] teamed up with the locals to help electricity grid cope with demand.
Microsoft has flashed its green credentials with the news that it has teamed up with Agder Energi and Powel AS for an energy project.

The partnership between the three will see them develop a smart electrical grid to help certain Nordic regions cope with the challenge of peak demand during certain periods. The firms noted that whilst the past decade has seen some remarkable advances in energy technologies, unfortunately power grids have remained fairly constant and still require tremendous pre planning and huge investments to keep pace with growing energy demand.

Power grids nowadays also need to be able to integrate distributed energy producing systems from rooftop solar panels, batteries, and smart homes, all of which can help reduce capital investment and ease the power burden during peak times. To this end, Microsoft offered its Azure intelligent cloud, PowerBI and Azure IoT Hub, and combined it with the expertise from energy services specialist Powel. The tools from these two were used by the Norwegian power utility Agder Energi to improve the dispatch of new energy resources, including device controls and predictive forecasting for situational awareness. Agder Energi used new technology to make the power grid more efficient, more predictable and more flexible. It went from being energy generator to energy partner, with a more active role for its customers. Together with its partners at Microsoft and Powel, it used innovation to solve the challenges facing the grid.

The project aimed to show how power companies could implement intelligent cloud-based smart grid solutions to unlock a host of energy and sustainability benefits. According to Powel's CEO Bård Benum, it applied its strong domain expertise and acting as the integration and forecasting partner in this project to seamlessly integrate Agder's SCADA to the Azure IoT Hub, provide demand and production forecasts for wind and solar in the region.

The idea was that the project would be operating from an Agder Energi substation that was operating at 120 percent of its capacity a number of times throughout the year. The project-helped utilities better predict demand and engage distributed resources such as solar panels, in order to ease the demand on the substation and save money that would otherwise be needed to upgrade it. It performed automatic load balancing of renewable energy and peaked load control in near real time, as well.

Renewable energy resources and advancements in intelligent cloud technology are driving a digital transformation of grid operations to explore new business models. Enabling solutions like Agder's will accelerate widespread renewable and distributed power generation. It is a bright future for the utility industry and for a sustainable world.

6.7 CleanSpark Project

CleanSpark [44] is dedicated to helping customers optimize their energy usage through the use of microgrid technology. Since its founding in 2012, CleanSpark has worked with companies all over the world, and from all industries, to design, develop, install and maintain custom microgrids. Though CleanSpark is primarily a software company, it also prides itself on offering exceptional service throughout the energy optimization process. CleanSpark is headquartered in San Diego, California and employs approximately 50 people.

Business challenge story

At its most basic, the grid is an interconnected network that delivers energy from a central power source to end users. In most homes and businesses, connecting to the grid is as simple as sliding a plug into a wall socket. The grid is generally reliable, but it can be subject to both blackouts and brownouts, and it can be more expensive and less clean than alternative sources of energy like wind and solar. These alternative sources are becoming more and more popular, but they, too, have their limits when it comes to availability and storage.

Microgrids are designed to help companies meet power needs with a combination of traditional grid, solar, wind, fuel cell and other energy technologies. Ideally, they balance load requirements among the different sources, providing customers with steady, clean and cost-effective energy. CleanSpark was established in 2012 to help organizations develop their own microgrids.

Microgrids are not one-size-fits-all endeavors. Whether they are meant to address the energy needs in a single household or in a massive, interconnected and geographically dispersed series of buildings, microgrids are complex, and the number of factors that go into optimizing them can be mind boggling. Rich Inman, Director of Data Analytics at CleanSpark, estimates that the microgrid optimization process evaluates tens of millions of variables in a sparse matrix that may contain as many as elements.

After a careful cost-benefit analysis, CleanSpark selected ILOG CPLEX Optimization Studio software of IBM to solve the increasingly complex challenges that microgrid design presents. This software, provides significant time savings as a result of direct integration with the company's existing technology stack. It also, differentiates CleanSpark as a leader in the increasingly crowded field of microgrid development, and delivers vast amounts of highly valuable data that positions the company to continue improving microgrid technology.

6.8 Streamlining optimization and data science workflows to help system operation for the Canary Islands

The Canary Islands' [45] electricity network is isolated from the mainland grid, so Red Eléctrica de España needs to balance local generation with demand to provide a reliable supply to homes and businesses. Red Eléctrica de España is responsible for the supply of electricity to people and corporations across Spain. Founded in 1985, the company employs more than 1,800 people and its network stretches across 44,000 kilometers (27,000 miles).

When demand for electricity peaks or subsides, utilities typically have the option of trading electricity with the grid to maintain a reliable service. That's not an option on the Canary Islands, which lie far out at sea and are isolated from mainland Spain's electric transmission network. Unable to trade capacity, the islands must meet their entire electricity demand through local production. If they schedule more electricity than required, the excess capacity is spilled and money is wasted. Worse still, if demand exceeds production, black-outs and power-cuts can follow. To maintain a reliable electricity supply for Canarian homes and businesses while minimizing cost, Red Eléctrica de España must accurately forecast demand and schedule sufficient generation to match it.

Along with other utilities, Red Eléctrica de España is attempting to reduce the area's carbon footprint by integrating renewable energy sources. Compared to fossil fuels, wind and solar power are sustainable and highly efficient. However, they are also unreliable, because no-one can say for certain when the wind will blow or the sun will shine. To provide a reliable supply, Red Eléctrica de España must therefore strike the right balance between renewable energy sources and fossil fuels.

To predict energy consumption and optimize the energy mix, Red Eléctrica de España performs highly sophisticated predictive and prescriptive analytics. The company currently conducts these analytics processes using a custom-built solution that it developed in-house many years ago. While the process and models are fairly accurate, computations are heavy and the software is difficult to maintain. For example, each time Red Eléctrica de España develops a new prescriptive optimization model, the update must be installed individually on each of its 19 machines, a time consuming process. This limitation makes it difficult for the company to develop and roll out enhancements for the solution to refine its models and improve its reporting capabilities. To escape these shackles, the company was eager to test a more modern data science toolset that would be easier to manage.

In a recent proof of concept, Red Eléctrica de España trialed moving its analytics models to IBM Watson Studio Local. Watson Studio is a collaborative platform for data scientists that combines proprietary and open-source data science tools into a coherent, integrated and controlled environment. Within the IBM Watson Studio Local portfolio, Red Eléctrica de España used IBM Decision Optimization, including IBM CPLEX Optimizer, and IBM SPSS Modeler Stream Canvas.

In the first stage of the project, IBM helped Red Eléctrica de España port its existing model for forecasting electricity demand to IBM Watson Studio Local. Next, the team transferred the energy generation model, which optimizes how to produce enough energy each hour to meet demand while maximizing use of renewable energy and minimizing use of fossil fuels, and which is written and solved with IBM CPLEX Optimizer. The energy generation model is highly detailed, showing when to turn particular generators on and off, and how much reserve capacity the company should maintain in its transmission network to accommodate the uncertainty of wind and minimize the risk of blackouts. The outputs of the forecasting and energy generation models enable Red Eléctrica de España to plan and simulate generation one year ahead.

In the second phase of the project, IBM and Red Eléctrica de España used IBM Watson Studio Local to examine its long-term demand-forecasting model, which looks at factors such as climate change and population growth. An updated version of this model could potentially help Red Eléctrica de España decide where and when the system needs new capacity, and the impact of building new wind turbines, solar panels or fuel-powered plants. For Red Eléctrica de España, the proof of concept demonstrated that IBM Watson Studio Local would enable closer connections between its demand forecasting and energy generation environments. The proof of concept demonstrated the viability of implementing Red Eléctrica de España's existing models in IBM Watson Studio Local reducing the burden of system management. In the future, Red Eléctrica de España could achieve even greater benefits from IBM Watson Studio Local by taking further advantage of its machine learning capabilities. For instance, imagine that a certain combination of weather patterns across the country means that the optimal result is always to use certain generators and a particular fuel type. Today, Red Eléctrica de España would need to re-run the optimization model each time that situation arises. But by using a machine learning model to identify and encode that rule, the company would be able to identify the correct solution much faster, and save the cost of computation.

6.9 Turning up cost efficiency and output for windfarms with predictive maintenance solutions in the IBM Cloud

In the energy sector, optimizing asset performance and deploying maintenance teams efficiently can have a huge impact on output. Performance for Assets (P4A) teamed up with the IBM [46] Garage consultancy to develop an advanced monitoring system for wind turbines in the IBM Cloud, emerging with an initial version in just eight weeks. P4A serves the energy and industry sectors with actionable insights that enable companies to optimize production, increase uptime and save energy. Established in 2017, it is a spin-out company based in Belgium, built on a collaboration between its shareholders Maintenance Partners, Vincotte, Icare and SRIW.

In recent years, the European Union (EU) aims to boost the proportion of energy generated from renewable sources. In 2016, wind overtook coal as the second largest form of power generation capacity in the EU, and it's fast catching up to the leader: gas. As demand for wind energy grows, producers have the chance to make huge gains—if they can tackle the issues that limit production levels.

Apart from weather conditions, which wind energy producers cannot control, the other main factors that affect output are asset performance and availability. Until now, most wind turbine owners have had little to no insight into the condition of their machines. Even though their end-of-design lifetime is usually close to 20 years, original equipment manufacturers (OEMs) for wind turbine components will typically guarantee operation throughout a 12 to 15-year long-term service agreement (LTSA). Once the LTSA comes to an end, wind energy companies struggle to find insurance for components. At the same time, owners lack visibility of how well assets are performing, meaning they could be operating at well below maximum throughput for years at a time.

P4A saw an opportunity to extract much greater value from wind turbine assets. It created Wintell, an advanced monitoring system that combines sophisticated analytics with field experts' process knowledge in preventive, corrective and predictive maintenance, condition monitoring, testing inspection and certification and data mining to provide actionable intelligence. P4A chose to focus on wind farms as an initial use case. In developing the concept, the P4A team realized that it needed support from technology experts and a flexible, scalable cloud platform to bring Wintell to market successfully. Specifically, it chose to work with the IBM Garage in Nice to bring its cloud-based advanced monitoring system for wind farms to life. The worldwide network of IBM Garage locations is designed to make IBM Cloud industry knowledge and higher value technologies accessible to enterprises globally. P4A engaged I-Pulses, experts in business intelligence and analytics solutions, for help with the project.

To kick off the project, P4A and I-Pulses participated in a three-day Design Thinking workshop to refine the Wintell concept. During the session, the team defined the target users, their precise needs and how the solution could most help them in practice. Next, the Garage team employed agile development and continuous delivery techniques to help P4A and I-Pulses create an MVP in an IBM Cloud Foundry environment. The solution collects and processes data from wind turbine sensors and combines it with weather forecast information using IBM Watson IoT™ Platform solutions hosted in the IBM Cloud. Data is stored in IBM Informix® on Cloud, a fast, scalable database, delivered as an automated infrastructure-as-a-service solution. P4A data scientists worked with IBM and I-Pulses to build hybrid machine learning models that detect and diagnose issues with wind turbine components, and provide an accurate prognosis to indicate how long each asset will perform well. With the IBM Watson® Machine Learning service, the company can deploy these models into production at scale, taking advantage of automated, collaborative workflows to streamline development.

With support from the IBM Garage technical experts, P4A and I-Pulses created a user interface that allows field technicians and managers to browse analytics results and visualize wind turbine data. Finally, it also incorporated a virtual assistant called Nestor using IBM Watson Assistant, which enables users to ask questions about recommended next actions, the status of different components and prioritization of jobs according to their impact on cost, output and availability.

6.10 Keeping the lights on and power grids stable in Belgium

As more homeowners adopt renewable energy sources like solar panels and select energy providers that have “green” power options, utility companies must evaluate the readiness of the grid. While traditional energy sources like hydro, nuclear, gas and coal provide consistent energy, renewable energy sources vary based on weather conditions or time of day. And if these power sources are not managed properly, they can affect the stability and quality of electricity and even damage power lines and other grid equipment. To address these factors and meet the European Union’s goals to increase energy efficiency, Eandis implemented SAS Visual Analytics to better visualize the volumes of grid data available for analysis. The largest distribution operator in Belgium now has the tools to better manage a new smart grid that meets consumer demands while modernizing its infrastructure to accommodate renewable energy sources.

For decades, Eandis had a classic distribution model for electricity and natural gas, which it provides to 229 cities and municipalities in Belgium. Its system used cables and gas pipelines to transport energy from the point of generation to the point of consumption. But that dynamic is changing, thanks to solar and wind power at the consumer side of the grid.

Rather than a one-direction flow, energy goes both ways and this is a big challenge for the distribution grid, because there is a lot of pressure on it from a technical standpoint. The flow can deteriorate the power lines and affect the stability of the electricity grid. For example, the lights might flicker in a house because there is some variance on the voltage. This increases if the grid is not managed properly. To address this decentralized production of energy, Eandis is upgrading its power grid to manage the two-way flow of energy. The company is also developing a smart grid that uses digital technology and sensors to continuously track energy usage, detect abnormalities and have better insight on consumers’ electricity use.

In the past, Eandis used traditional methods of reporting. If a business unit needed a custom report, the IT department captured the requirements, developed the extractions, filled the data warehouse and built the reports. The entire process was slow and frustrating, and the results were not easy to interpret. At the heart of the problem, the old system was set up for hindsight, to learn more about past events.

With SAS, Eandis now has a sandbox environment where users can test out models and the types of visualizations like graphs and charts that might be useful to them. When the analysts are satisfied with the candidate model, IT helps them move the predictive model into production.

Not only is SAS Visual Analytics easy to use, but it allows us to explore big data and make decisions much faster than before. For example, every year Eandis measures the in-feed data from all the transmission stations supplying power to its grid, creating 10 million lines of measurements – enough to fill nine Excel spreadsheets. The analysis of that data took three to six months. With SAS Visual Analytics [47], the analyst was able to complete that same job within minutes, and in a matter of seconds he could visualize and manipulate the data, add columns and do calculations. The time gained can be used to analyze the data and make different alternative models, allowing faster and well-founded investment decisions, such as when to build new transmission stations. Another benefit is that reports produced in SAS Visual Analytics are easy to interpret and support better decisions. In particular, they are visually compelling, user-friendly and refresh faster and analysts can make sense of the complex data and ultimately make sound business decisions. Finally, Eandis’ recent analysis shows that green energy production is increasing, but overall energy consumption is still going up. It is looking to reverse that trend by educating consumers and further developing its smart grid project.

6.11 Combating energy theft with analytics

While fraudulent activity in most industries leads to lost revenue or increased risk exposure, energy theft is different. It is not just about the revenue lost or the drain on the energy grid. Energy theft can be illegal and hazardous. Tampering with equipment such as meters, pipes and wires, exposes people to the risk of electric shock, explosion, injury or possibly death.

Before using SAS Enterprise Miner, utilities could find only 35 percent of fraudsters during the inspection. Now they can detect more than 50 percent. In Brazil, the level of energy theft was a concern for the National Electric Energy Agency (Aneel). It estimates that nontechnical loss has a financial impact of US$1.2 billion a year in the country. If the risk to life and limb will not thwart thieves, perhaps analytics will.

Utilities provider Cemig (Companhia Energética de Minas Gerais) uses SAS Analytics to increase accuracy in locating power deviations and technical faults in meters scattered around the metropolitan region of Belo Horizonte, Minas Gerais. Cemig is a major electric utilities provider in Brazil with more than 7.5 million consumers in 774 municipalities. The company is also a power generator, operating 70 plants with an installed capacity of 7,295 megawatts. Because physical inspection of all energy customers is not practical, analytical prioritization enables Cemig to identify and prioritize candidates.

Since implementing SAS Enterprise Miner [48], it saves about $420,000 a month by detecting energy theft. In particular, it uses the technology for statistical modeling, helping analyze data about consumption history, socio-demographic information and geographic potential of each household. After analyzing the data, Cemig uses SAS to generate a score for the household or establishment. The score shows the probability that the location is the site of energy theft. When the investigator visits the site, he already knows in advance which installed devices are more likely to have failures and deviations. He also can identify the consumer units with higher volume offset. These are the ones that cause significant harm to the company and to the end consumer.

6.12 How Uttar Pradesh is meeting the challenges of a smart meter rollout

India has launched an ambitious modernization of its electrical grid, replacing 250 million conventional meters with smart meters over the next 18 months. It is a daunting task to switch out a quarter billion meters across a huge country, and the start of this project is focused in two Indian states, Uttar Pradesh [49] and Haryana. Uttar Pradesh Power Corporation, UPPCL, the state-owned utility serving the northern India state, announced in March 2020 the goal of 4 million smart meters installed within a year.

One and a half million smart meters have already been rolled out and there is a partnership with leading technology vendors like Oracle, under the leadership of the Energy Efficiency Services Limited (ESSL). The ESSL, a government joint venture, aims to bring low-cost and low-carbon energy to the Indian population. The ESSL works to reduce India's carbon footprint while increasing its energy efficiency and the smart meter program is a critical tool in these goals.

While installing millions of smart meters is a formidable job itself, doing it in a pandemic year adds even more challenges. On top of that, India has millions of pre-paid meters to incorporate into its modernized system. And then, once the installation is complete, here comes the massive amount of customer data that needs to be processed.

With a high percentage of pre-paid meters, and widespread electricity theft, India's electric power sector leaders believe that smart meters, and a generally modernized grid, will help solve some of the inefficiencies plaguing the country's utility industry for decades. Oracle provides cloud-native products and better grid management tools to support for every single step of the customer's journey increasing billing efficiency, completing adoption of remote billing, automating outage reporting, and using the data collected in real time.

The long-term benefits in upgrading India's 250 million conventional meters is equally as important. Utility leaders looking in to the future see greater reliability, and resiliency for the electric grid, increased energy savings, and a decrease in electricity theft. Moreover, it's not to say there aren't costs, in capital and resources that the sector must overcome during the conversion. In addition to the hard costs of purchasing and installing smart meters, there is the challenge of integrating this complex technology through numerous systems, including outage and work management systems, and making it compatible with the SCADA network. But, no matter the challenges, or the bumps along the way, the utility, the electric sector, and the government are dedicated to the electrification of the Indian electric system. In other words, they understand the clear and present context for a smart grid and its integral linkage to demand-side management. However, the responsiveness of the grid to the concurrent demand situation under every setting can only happen with a strong digital backend.

6.13 ODEC accurately forecasts energy needs, keeps rates low for members with SAS Energy Forecasting

When you flip on a light switch, you expect the light to turn on. In addition, except in rare cases, it does. That is because utility companies plan your energy needs months ahead, using data from previous years to forecast demands on the grid. Better forecasting not only means more reliable service, but it can also help companies keep costs down. Providers like Old Dominion Electric Cooperative (ODEC) [50] can use analytics to predict when to buy energy in advance. This translates into cost savings and lower energy costs for member companies. ODEC provides wholesale power to 11 non-profit member distribution cooperatives in Virginia, Maryland and Delaware, serving 1.4 million customers in the rural and suburban areas of those states. To give its members the best value, ODEC uses SAS Energy Forecasting to manage and analyze a wealth of data on weather, demand and population growth. The results are substantial. Within two years of implementation, ODEC reduced its rates four times, saving members millions of dollars along the way.

Forecasting helps utilities plan days, months and even decades ahead. They use a variety of internal and external data points to anticipate how much energy they need to buy and which investments to make, like the $3 billion natural gas-fueled electric generation facility ODEC is building in 2017.

The ability to forecast for future demand is critical for cooperatives like ODEC that buy a majority of their power from the energy market. Currently, ODEC produces 47 percent of the power it provides. It needs to purchase the remaining 53 percent months in advance – or operate at the mercy of the market.

SAS Energy Forecasting allows ODEC to forecast more efficiently and accurately. Depending on the prices, ODEC buys contracts from six months to three years in advance. The company seeks to buy blocks of power at times when the prices are low. In the past, ODEC was using an off-the-shelf software tool along with Microsoft Excel to perform energy forecasting. The inefficiencies of this method added to the data validation challenges. More importantly, it could lead to second-guessing from regulators and auditors.

For ODEC, the solution was simple: make better decisions from the data already available. Now SAS is the single source for medium- and long-term. Since the forecasting is automated, analysts are no longer required to be programmers to run the forecasting process. The system automatically builds the most appropriate model for the data, allowing ODEC to forecast with greater accuracy. In the last two years, SAS is used to produce ODEC forecast budgets within 1 percent of accuracy. Forecast department can account for variables like anticipated demand and population growth, also allowing for unforeseen changes – like an unexpectedly cold winter – to alter the forecast. Increasingly, ODEC is also using SAS to ensure grid reliability as renewable sources like solar panels, electric cars and wind farms become more mainstream. ODEC also is developing data marts to drive its models with pre-verified variable streams. These massive data processing activities allow ODEC to build and update forecasts in hours versus days. Additionally, it is looking into using SAS Analytics for short-term forecasts to determine peak days and specifically the hour of the peak and the estimated load of that hour. This will help reduce load during peak conditions to better control transmissions and costs. With this increased efficiency, ODEC can build models faster, share information better and continue to keep rates low for its members decades to come.

6.14 Southwest Power Pool

Southwest Power Pool, (SPP), [51] manages the electric grid and wholesale power market for the central United States. As a regional transmission organization, the nonprofit corporation is mandated by the Federal Energy Regulatory Commission to ensure reliable supplies of power, adequate transmission infrastructure and competitive wholesale electricity prices. SPP and its diverse group of member companies coordinate the flow of electricity across approximately 60,000 miles of high-voltage transmission lines spanning 14 states. The company is headquartered in Little Rock, Arkansas.

SPP plans to save its members $100 million annually via a real-time “Integrated Marketplace” that balances regional energy supply and demand, and enables wholesale energy to be bought and sold in both a day-ahead and real-time market SPP needed to replicate data collected by production transactional systems from 400+ source systems (including energy produced and consumed and power line data) into an analytics data source, making it available for analysis and reporting in the Integrated Marketplace. The scale of transactional data made it prohibitive to manage in the production database because Integrated Marketplace requires optimal performance, stability and scalability to operate effectively.

The Informatica Platform and data replication functionality facilitates near real-time data replication into an analytics data source that is optimized for analysis and reporting. The solution reduced typical analysis time from one day to 20 minutes, resulting in significant uplift in simulated scenarios for improved energy service provision. In particular, Informatica PowerCenter, Informatica PowerExchange, Informatica Data Replication and Informatica Data Explorer contributed to the completion of this project.

6.15 Enverus Pumps Data-driven Applications Faster Using Denodo’s Data Virtualization Platform

Enverus’s [52] business growth drove the need for the company to build next generation products to support key energy market segments. These products include applications to support well production and energy field services workflows, geo services for map analysis, a Geology, Geophysical and Engineering (GG&E) platform for interpretations and visualization, as well as a soon to be released mineral interest analysis application. Rapid time-to-market for these products and applications was crucial and this implied that the Data Tech team needed to deliver a data platform that supported the internal application development team quicker than they had been doing in the past. Also, rapid delivery of data directly to the customers was needed as well. However, the Data Tech team was challenged with integrating the data across the data warehouse, other data sources and providing it to the data consumers quickly. The product development team’s delivery timelines were routinely at risk due to data availability and data consistency issues. As a result, the developers were directly accessing the data sources and in other cases suffering from severe delivery delays. To solve this issue, one option was to use conventional ETL (extract, transform, and load), but that would take several weeks. A more timely way of meeting the needs of the product development team was needed. The Data Tech team needed a solution that would rapidly expose data sources through a variety of data services (primarily RESTful web services and JDBC) to the development team and their customers. The data services would be cataloged around eleven of their core line of business entities such as wells, leases, permits, production, rigs, and so on.

Due to strength of the Denodo Platform to rapidly expose underlying data from source systems as data services, Enverus decided to use the solution to manage and quickly provision all of the data to the product development team and its customers. Enverus ETLs the regulatory agencies data into an internal data store that powers the DI Classic product. The production data is stored in another system called DI Desktop. The data from DI Classic, DI Desktop, as well as geo-spatial and Optical Character Recognition (OCR) data stored in other systems are then ETLs into their data warehouse. Enverus has created a virtual data abstraction layer using the Denodo Platform above the DI Classic, DI Desktop, and data warehouse. The Denodo Platform connects to these data sources, combines the data and publishes the resultant virtual views as data services, which are consumed internally by the application development team, analytics and decision support applications, and application data marts, as well as externally by their customers. Enverus uses Denodo's caching mechanism to store data about business entities such as wells, completions, producing entities, permits, and so on, which are exposed as data services, analytics services, and map services. These services are then used by their internal application developers as well as customers to build applications. One of the important products built by Enverus is the Production Workspace application that enables analysis of global producing entities (wells, leases, completions). This application uses the data services provided by the Denodo Platform to access the virtually integrated data. Also, using the data services provisioned by the Denodo Platform, Enverus has built a whole new product called Royalty-info (general availability to be determined), which will enable customers to understand the interest market, its resource locations, and other competitive information. Enverus has so far built 24 different data services around 11 of their core line of business entities.

Building usable web services for developing applications used to take 1 – 2 weeks with Denodo's data virtualization solution, this process now takes less than 1 day. Enverus now has just one full time developer and a part time virtualization admin managing the entire data virtualization process. About 20 – 30 internal developers and seven external customers are using the data services to build data-driven applications. This has saved Enverus precious time and resources to achieve the primary benefit of rapid time-to-market for their products.

6.16 AI analysis of big data prepares Denmark for a greener future

In Denmark, around 50 percent of electricity comes from renewable sources, mostly wind power. The mandate is to increase that to 100 percent by 2030. This creates some challenges for Energinet, Denmark's electric transmission systems operator, because renewable energy always fluctuates. Energinet has to manage the grid carefully to maintain the security of supply. It has developed tools to manage the amount of renewable energy we have today, but as it increases, it needs new and better tools. Otherwise, it will likely have to make costly infrastructure investments or face brownouts and blackouts. Its current control room tools are good at modeling the grid to simulate error conditions, but the simulations and live data feeds generate big data that remains untapped. That led it to wonder whether an analytical tool could discover insights to improve grid management.

To test the concept, it collaborated with IBM_Services [53] on a pilot project. The result was a real technological leap for Energinet- a multicloud solution that gleans operational predictions from big data using AI. Accessing the system from a web interface, operators get help answering questions like, "What would happen if we took equipment out of service at this time?" or "Based on past experience, which assets are at risk of failing?" It's a huge step forward in decision support.

An important use case is helping operators evaluate planned maintenance. If the maintenance team wants to take down a line or transformer, operators need to assess the risks. The system's predictions are likely to be more accurate than their intuitions. Other uses include assessing grid operations, understanding system bottlenecks and suggesting cost-effective investments.

Energinet personnel had the idea for the solution, but participating in design thinking sessions helped IBM understand what is possible and how to do it. Then, with an agile approach IBM developed the proof of concept in just three months. That's very fast and cost effective compared to traditional systems development for the control room.

Key to the analytical power is preparing the big data for AI. Systems running on the Microsoft Azure cloud first create simulation and real-time datasets. IBM Cloud Pak for Data on Azure allows users to query the system and AI generates the analysis.

Of course, the usefulness depends on operators trusting the AI. The pilot addressed this by offering explanations for its predictions. IBM tested the capability by simulating outages with known causes and remedies. Experienced operators easily recognized what to do and why, and then compared their thinking to the AI analysis. The fact that they generally agreed increased trust in the system.

In conceiving the solution, IBM team aimed to help operators understand the risks of removing equipment from the grid. The project proved that possibility and more.

In the future, IBM plans to suggest actions that prevent a cascade of problems that might come later. Such AI capabilities can help assure a secure and cost-effective renewable energy supply.

6.17 Revitalize Data Infrastructure with EDB POSTGRES and Virtualization

Organizations regularly take the opportunity to reevaluate their relational database management system (RDBMS) when planning a hardware refresh or new virtualization strategy. One Enterprise DB (EDB) [54] customer, an energy utility company located in the Midwestern United States, provides a glimpse into how infrastructure projects often lead to new DBMS decisions.

The utility provides electricity and natural gas to millions of customers. As it approached a major hardware refresh, the IT team decided to use it as an opportunity to gain greater operational efficiencies through virtualization. The company wanted to reduce costs as well, and correctly suspected their RDBMS could be a source of savings. The team decided to look for an alternative to Oracle, their greatest source of expense. They selected EDB Postgres as an Oracle replacement and migrated Microsoft SQL Server to EDB Postgres as well, amplifying their savings. Reevaluating RDBMS can prompt IT teams to evaluate their overall architecture and critical elements of their infrastructure. A key reason to reevaluate architecture and infrastructure is that DBMS licenses for many traditional vendors place limits on hardware configurations and deployment environments. This can potentially drive up costs for modernization projects. Any change in infrastructure might impact RDBMS usage, driving up those costs as well. Organizations in these circumstances often consider open source alternatives to lower database costs.

EDB enabled the energy utility to accomplish the goals of its new infrastructure plan:

Migration: The company wanted to reduce RDBMS spend, and open source-based EDB Postgres reduces costs by as much as 80% compared to Oracle. As for SQL Server, the company estimated that based on list prices, migrating to EDB Postgres would reduce costs for those databases by 66%. EDB Postgres has built-in compatibility for Oracle and provides the EDB Postgres Migration Toolkit, which migrates tables, data, stored procedures, and custom developed packages from the Oracle database to EDB Postgres. Because of some customized code unique to the company, EDB engineers developed scripts for SQL Server migration, as well.

Modernization: Ultimately, EDB's subscription model was cost-effective in supporting the new configuration the company wanted for its new data infrastructure. Like the energy utility, organizations seeking to modernize their infrastructures with virtualization learn about the limitations or increased costs of their traditional DBMS vendors. By contrast, EDB Postgres is available by a flexible subscription model that allows organizations to deploy EDB Postgres on-premises, virtualized, or in the cloud or containers, and freely move the licenses between environments when needed.

EDB's experience with database migrations proved invaluable during the company's pilot test with a Microsoft SQL Server migration that was accomplished in a single day. The company decided to run a pilot project with a small application controlling building security at its headquarters involving a 30GB to 50GB database. While the EDB Postgres Migration Toolkit can be used to migrate Microsoft SQL Server to the EDB Postgres Advanced Server database, the company had developed its own procedural language and utilized a great deal of customized code. Instead, EDB engineers developed a script to ensure the database schema was compliant with the EDB Postgres Advanced Server database, and were able to avoid making any changes to the application.

6.18 Cisco Data Virtualization helps Long Island Power Authority gain productivity and reduce costs while modernizing IT infrastructure

Business Challenge Hurricane Sandy left roughly 90 percent of Long Island Power Authority's (LIPA's) 1.1 million customers without power. The recovery has been the slowest on Long Island. Many customers were without electricity for weeks after power was restored to most of New York City and other parts of the metropolitan area. As a result, customers, municipalities, and the business stakeholders demanded faster, more responsive engagement with accurate information. To better, serve its customers, LIPA [55] needed to develop a plan for a new storm process with a supporting power outage management system. At the heart of this effort was the transformation of the IT infrastructure. To implement the new process, the project team needed to upgrade dozens of interfaces from multiple generations of technology. Mainframe applications were over 20 years old. Countless copies of data left users wondering what information was accurate. Hurricane Sandy revealed the weakness in this complexity. When the power went out, LIPA experienced significant issues delivering outage information due to middleware and interface performance and reliability during the stresses of the storm. Connecting hundreds of mismatched components and data models, not to mention licensing costs and unsupported software, was complicating architectures and support plans in the new data centers. LIPA needed to modernize its IT infrastructure and deliver a transformational storm process. Network Solution Modernizing and restructuring the infrastructure with an enterprise approach (compared to a silo approach) was the only way to meet the business requirements.

However, this meant that LIPA had to find productivity gains and cost savings to stay within its budget. LIPA selected the Cisco Information Server, the foundation of the Cisco Data Virtualization suite, for its data federation, query optimization, and enterprise data-sharing capabilities. With Cisco technology, the team can query all types of data across its network as if it were in one place. LIPA used Cisco Data Virtualization technology to streamline interfaces and data movement. Following this enterprise approach, legacy reporting and data analysis applications were replaced with an enterprise business intelligence system featuring self-service capabilities for customers. The team used Cisco Plan and Build Services for Data Virtualization for services, such as training and migration. This Cisco solution provided confidence and productivity gains to accelerate the integration of new technologies and make the project come together on budget and on schedule. Business Results Selecting Cisco Data Virtualization greatly reduced system complexity and improved performance and reliability. With the Cisco support, LIPA stayed within an incredibly compressed deliverable timeline. The real-time predictive modeling updates give LIPA the technological advantage. LIPA's old mainframe was a batch-oriented model, dependent on crews surveying affected areas with operators transferring the power outage information to the system. Now, LIPA is backed by a geospatial electric connectivity network, which brings actionable knowledge of outage extent and repair in real time, improving customer service. The combination of Cisco technology and support enabled a transformational business and IT effort. Highlights include:

- Enterprise-level system that meets the needs of operational and regulatory reporting, as well as serving the real-time needs of customers
- 3 year project completed in less than 12 months
- 150 percent improvement in outage location accuracy
- 50 percent faster data integration

6.19 TransAlta Case Study: A Cloud Modernization Story

TransAlta [56], founded in 1909, is an electric power generator company, headquartered in Calgary, Alberta. It has grown to become an energy producer with over 2300 employees, approximately $3 billion in annual revenue, more than $9 billion in assets, and over 60 gas, hydro, solar, wind, and coal facilities across the U.S., Canada, and Australia. TransAlta's IT department initiated "Zero Data Center" project to move their entire data layer to the cloud for flexibility, agility and lower TCO. Data virtualization technology played a central role in TransAlta's real-time data integration, while helping them move to the cloud with zero downtime.

TransAlta may be a deeply established organization with a deep heritage, but the company wanted to do something radical: Reduce its datacenter footprint to zero and move all data to the cloud. This would serve the dual purpose of reducing infrastructure costs while also enabling the flexibility to scale up or down with the demands of business.

As TransAlta began this transition, the company found that not all of its applications could be moved to the cloud; some relied on authentication procedures developed in the on-premises environment, and they could not be easily reconfigured to enable cloud access. In addition, since some data centers were farther away than others, geographically, they introduced latency into the migration process.

To help with the migration, TransAlta leveraged the Denodo Platform for Microsoft Azure. As the company was already using Azure for all of its IaaS and PaaS solutions, deploying the Denodo Platform on Azure was straightforward. Once installed, the Denodo Platform established a real-time data access layer between the datacenters and the cloud. This abstracted TransAlta from the complexities of accessing the datacenters, including geographical distance, to facilitate a smooth transition to the cloud. Due to the fact that the Denodo Platform was established as a central layer, it provided a central point for managing authentication, making it seamless for users to access the migrated applications.

Besides the cloud migration, the Denodo Platform provided TransAlta with additional ongoing benefits. TransAlta leveraged the Denodo Platform as a virtual data mart in support of the company's energy trading system. This enabled TransAlta to avoid having to implement costly ETL processes while also providing near-real-time data access for energy trading. With the Denodo Platform in place, TransAlta's Wind Operations Group can make real-time decisions about mitigating icing risk by pulling machines offline versus letting them produce more power. The Denodo Platform also enabled TransAlta to create an HR dashboard that draws on data from SAP BW, HANA, SQL Server, and Active Directory to produce unique, integrated, 360-degree views of every manager.

7. HEDNO S.A. Case - Current Landscape/ Infrastructure

7.1 Introduction/ Regulatory Framework

7.1.1 Introduction

HEDNO S.A. (Hellenic Electricity Distribution Network Operator S.A.) [57] was formed by the separation of the Distribution Department from PPC S.A. (Public Power Corporation S.A.), according to L.4001/2011 and in compliance with 2009/72/EC EU Directive relative to the electricity market organization with the goal to undertake the tasks of the Hellenic Electricity Distribution Network Operator. It is a 100% subsidiary of PPC S.A., however, it is independent in operation and management retaining all the independence requirements that are incorporated within the above mentioned legislative framework.

The tasks of HEDNO S.A. include the operation, maintenance and development of the power distribution network in Greece, as well as the assurance of a transparent and impartial access of consumers and of all network users in general. It aims at providing reliable power supply to their customers, quality of electricity voltage and constant improvement of quality in services.

Their goal is to substantially contribute to the development of Greece as well as the welfare and improvement of the citizens' quality of life in the whole territory with reliable and economically efficient power supply respecting people and environment. Their vision is to establish a company-model in the field of power supply, which will provide excellent services to citizens, operate and develop the network in accordance with the standards of the most advanced countries and assure that the network users, employees, associates, shareholders and society in general are totally satisfied.

7.1.2 Regulatory Framework

A. Regulatory Framework governing competencies concerning the management of the HEDNO

HEDNO S.A. is responsible for the development, operation and maintenance of the HEDN under economically advantageous terms and in accordance with the Management License, so as to ensure its reliable, efficient and safe operation. Taking due account of the environment and energy efficiency, HEDN ensures access to it for users (consumers, providers) and suppliers in the most cost-effective, transparent, direct and non-discriminatory way in order to carry out their business operations.

The operation of distribution network management is a natural monopoly in the area where it runs as there is no competition. For this reason, these business operations are supervised and regulated by the independent Regulatory Authority for Energy (RAE). Regulating is achieved by approving the revenue that is allowed from such operation, while objectives are set for the improvement of both customer service and the efficiency of the company's operation, providing incentives for their achievement.

In addition to Law 4001/2011, which outlines the operation, development, maintenance and access of users to HEDN, the main Regulatory text which defines the above is the "Hellenic Electricity Distribution Network Code", which was approved by virtue of Decision 395/2016 reached by RAE. The content of the Code regulates the rights and obligations of the Hellenic Electricity Distribution Network Operator, as well as the rights and obligations of Network Users and Providers in addition to issues related to the development, operation, network access, the services provided by the Network Operator and the financial reward there of.

The details of the implementation of the provisions of the above Code, as well as the necessary procedures and calculation methodologies required for its implementation, are set out in the Application Manuals which are an integral part of the Code. The Manuals already approved by the RAE, which have been published and are in force, are the "Manual of Power Theft", which was approved in pursuance of Decision 236/2017 and the "Manual on Measurement Management and Regular Clearing of Network Providers", which was approved by Decision 404/2015.

In the context of improving the services provided to consumers by HEDNO, the RAE has approved a Program called "Guaranteed Services to Consumers" by virtue of HEDNO's Decision 165/2014. At the same time, the regulatory framework is also governed by other important regulatory texts issued by the Regulatory Authority for Energy by means of relevant Decisions:

The "Management License of the Hellenic Electricity Distribution Network Operator (HEDN Management License)" granted to HEDNO S.A. by virtue of Decision 83/2014. This license also covers any future extension of the HEDN.

The "Terms and limitations of the Exclusive Ownership License of the Electricity Distribution Network (Ownership License) of PPC S.A." granted by virtue of Decision 82/2014 to PPC S.A. because the ownership of HEDN remains solely with PPC. This license also covers any future extension of the HEDN.

The "Approval of the Annual Cost 2017 for the Hellenic Electricity Distribution Network Operator" by virtue of Decision 454/2016.

The "Use Tariffs of the Hellenic Electricity Distribution Network (HEDN)" by virtue Decision 455/2016 approving Use Rates (Unit Charges) based on the 2017 Required Income.

B. Regulatory Framework governing the responsibilities of HEDNO SA as Non-Interconnected Islands Operator (NII)

Non-Interconnected Islands (NIIs) are the islands of Greece whose Electricity Distribution Network is not connected to the Transmission System or the Distribution Network of the Mainland.

The management of the Electricity Systems of the Non-Interconnected Islands, which includes the management of the production, market operation and the systems of these islands, is the responsibility of HEDNO S.A. and is carried out in accordance with the "Non-Interconnected Islands Electrical System Management Code" provided for in Article 130 of Law 4001/2011.

With regard to the implementation of the Non-Interconnected Islands Electrical System Management Code, RAE has approved the "HEDNO S.A. Infrastructure Implementation Action Plan in accordance with Decision 2014/536/EC/14.08.2014 of the European Commission".

In addition, the regulatory framework is also governed by other texts issued by the Regulatory Authority for Energy concerning the average variable cost of conventional production units in the NIIs, Utilities in the NIIs, prices of hybrid power stations in the NIIs, etc.

C. HEDNO S.A. Compliance Program

Pursuant to Law 4001/2011, HEDNO S.A. has drawn up and is implementing a Compliance Program, which has been approved by RAE by virtue of Decision 678/2014. This

Program lists measures taken to exclude any discriminatory behavior, discriminatory corporate practices and distortion of competition in the exercise of its responsibilities.

7.2 Operations, Business Processes & Stakeholders

7.2.1 Operations

A major element of the company's strategic planning for its development is its transformation through the modernization of its structures, systems and processes. The main objective is to respond continuously and effectively to the challenges of an ever-changing electricity market both in Greece and at European level, taking advantage of technology and innovation while investing in knowledge based on sustainability and social development.

HEDNO S.A. is required to fulfill the role and mission assigned to it by the existing institutional framework for the exercise of the regulated business activity of Electricity Distribution in a modern, complex, and constantly evolving environment. This environment consists of a significant number of participants in the value chain and the electricity market, where network users, electricity providers, network managers and market operators are traditionally involved and where new entities are emerging by developing new business models and standards. In the light of these changes, HEDNO is required to ensure on a timely basis the transparent and impartial access of the users to the network, aiming at the continuous improvement of the quality of its services. Its main business planning objectives are the following:

- Improving energy quality
- Upgrading service quality through alternative service networks, limiting the need for transition and physical presence at customer service points
- Improving and extending our "Guaranteed Services" program, which reflects our commitment to delivering our core services within specific time limits
- Focusing on the quality and performance of corporate operations
- Continuously adapting to the needs of customers and all stakeholders through systematic research
- Protecting the environment and minimizing any extra charges that may be levied by networks
- Improving the Company’s economic fundamentals further, although HEDNO is among the most economical ones in the European Union
- Developing network capacity, optimizing infrastructure exploitation and reliability, and improving energy efficiency
- Complying with the ever-changing environment
- Actively contributing to the efficient operation of the Electricity Market
- Focusing on innovative development and smart grids

In order to achieve these objectives, operational planning for the next five years has provided for significant investments in the implementation of key modernization, technological upgrading and operational transformation projects. These projects are expected to dramatically improve the level of services provided and upgrade the company's operation and role in the electricity market by reinforcing its role and highlighting the benefits to all network users and the community at large.

Strategic Projects

The company has created a portfolio of special projects called Strategic Projects that are strategically important since they cover a wide range of important modernization undertakings such as smart grids, consumer telemetry, remote customer service and automation in many internal operations, always keeping an eye on all the new trends in the field of electricity. These strategic projects are a key pillar of the company's Business Plan. In order for HEDNO to achieve the intended benefits for its customers and the entire Electricity Market, it will focus mainly on the company’s appropriate and targeted modernization, including among other things the following:

- Remote management of the country's main electricity grids and decisive promotion of smart grids
- Remote supervision of the production of non-interconnected islands (where it has the responsibility of managing the Market)
- Decisive promotion of consumer telemetry
- Geographical mapping of networks, which will be the digital basis for network & service operations
- Creation of modern customer service channels (telephone & Internet) and their management centers
- Installation of modern information systems and further exploitation of the company’s new ERP

7.2.2 Business Processes/ Stakeholders

The major business processes and stakeholders of HEDNO are the following:

General Activities: These are crucial activities that address general issues and affect all services and all the staff and / or the Administrative Bodies of HEDNO. These issues concern the independence of Administrative Bodies, the separation of corporate identity and its communication, its promotion, the protection of confidential information, the commercial and financial relations with the parent company and the Staff training.

Consumer Management Activities: This section includes significant consumer-related activities that is, the services provided by HEDNO to him whether he is an existing consumer or for consumer requesting a new connection. Special reference is made to issues related to their management measurements and cash, the provisions to be taken for contractors used by HEDNO for the execution of its work, the management of vulnerable consumers and the Charter Consumer Liabilities.

Product Management Activities: This includes the important activities related to the services provided by HEDNO to RES producers, regarding their connection to the Network. Due to the significant degree criticality of these items, all activities related to the existing one are included process, from the reception of the request to the electrification and monitoring of changes during the life cycle of these stations.

Supplier Management Activities: This section includes the critical HEDNO activities related to Services provided to suppliers. Representation change activities, measurement data management and market clearance are included.

Non-Interconnected Islands Management activities: This includes activities related to the management of the Electrical Systems of Non-Interconnected Islands. The activities here are limited in view of the publication of the Code Management of Non-Interconnected Islands.

Activities for Organization and implementation of the Compliance Program: This section includes activities related to the Program Compliance and ensures its adaptation to the new Legislative Regulations and the continuous improvement of it. It also provides the management of the Compliance Program in times of crisis and emergency and the way of organizing for its implementation (with the establishment of Representatives Compliance in the Management and Assistant Compliance Representatives in the HEDNO Areas) and its control**.**

7.3 Departments & Systems

7.3.1 Departments

At the top of the organization chart is the board of directors. A board of directors is a group of people, who jointly supervise the activities of HEDNO. The powers, duties, and responsibilities of a board of directors are determined by government regulations (including the jurisdiction's corporate law) and the organization's own constitution and bylaws. The members of the board of directors and its administrative bodies do not hold any professional position or responsibility, nor have any interest or business relationship and receive a financial benefit, directly or indirectly, related to a company or body of the Hellenic State active in the field of production, supply or transmission of electricity or with PPC SA or any Affiliated company other than HEDNO. Responsible for the independent compliance control of HEDNO to the Compliance Program, is the HEDNO Compliance Officer appointed by the HEDNO board of directors.

Under the board of directors is the chief executive officer (CEO). A CEO is one of a number of corporates executives in charge of managing HEDNO. He typically reports to the board of directors and aims at achieving outcomes related to the HEDNO's mission, which is the development and operation of the Electricity Distribution Network and the electricity systems of the non-interconnected islands as well as the assurance of equal access to them by all consumers, producers and suppliers with transparency and objectivity. The CEO collaborates with the internal audit department and the executive office. Specifically, the executive office, as the Responsible Step for the development of the Corporate Identity, ensures the approval and publication of a relevant guide that will include the specifications for the logo, the brand name, the system written and electronic communication, the signage system of buildings, offices & services, the signage vehicles and technical equipment. The planning and recording of audit findings is carried out by the internal audit department.

The general counsel refers to the Chief Executive Officer, heads the legal department, legally supports its Management Bodies company and participates without a vote on the board at the discretion of the institution.

At the next level are the department heads. The general manager of strategy and transformation manages the following departments: strategy and transformation, research and innovation and the regulatory affairs. The general manager of finance is responsible for the budget and control, the accounting and tax and the cash flow. As for the manager of human resources and organization, he oversees the human resources, the health and safety and the premises unit.

The general manager of supply chain and digitalization manages the following departments: the materials, procurement and transformations, the information technology and telecommunications, the operational, improvement and digitalization and the data security unit, as well. The general manager of network development and operation manages the below departments: the network users, the Attica region, the islands network operation, the Macedonia-Thrace region, the network, the Peloponnese-Epirus region, the network major installations, the central Greece region, the monitoring of subsidized network projects and the islands region (fig. 16).

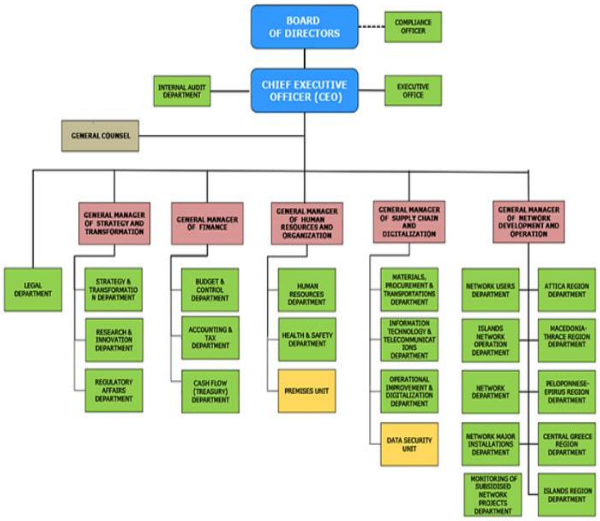
BOARD OF DIRECTORS
COMPLIANCE OFFICER
INTERNAL AUDIT DEPARTMENT
CHIEF EXECUTIVE OFFICER (CEO)
EXECUTIVE OFFICE
GENERAL COUNSEL
LEGAL DEPARTMENT
GENERAL MANAGER OF STRATEGY AND TRANSFORMATION
STRATEGY & TRANSFORMATIO N DEPARTMENT
RESEARCH & INNOVATION DEPARTMENT
REGULATORY AFFAIRS DEPARTMENT
GENERAL MANAGER OF FINANCE
BUDGET & CONTROL DEPARTMENT
ACCOUNTING & TAX DEPARTMENT
CASH FLOW (TREASURY) DEPARTMENT
GENERAL MANAGER OF HUMAN RESOURCES AND ORGANIZATION
HUMAN RESOURCES DEPARTMENT
HEALTH & SAFETY DEPARTMENT
PREMISES UNIT
GENERAL MANAGER OF SUPPLY CHAIN AND DIGITALIZATION
MATERIALS, PROCUREMENT & TRANSPORTATIONS DEPARTMENT
INFORMATION TECHNOLOGY & TELECOMMUNICAT IONS DEPARTMENT
OPERATIONAL IMPROVEMENT & DIGITALIZATION DEPARTMENT
DATA SECURITY UNIT
GENERAL MANAGER OF NETWORK DEVELOPMENT AND OPERATION
NETWORK USERS DEPARTMENT
ISLANDS NETWORK OPERATION DEPARTMENT
NETWORK DEPARTMENT
NETWORK MAJOR INSTALLATIONS DEPARTMENT
MONITORING OF SUBSIDISED NETWORK PROJECTS DEPARTMENT
ATTICA REGION DEPARTMENT
MACEDONIA-THRACE REGION DEPARTMENT
PELOPONNESE-EPIRUS REGION DEPARTMENT
CENTRAL GREECE REGION DEPARTMENT
ISLANDS REGION DEPARTMENT

Figure 16: HEDNO Organizational Structure.

7.3.2 Systems

The systems that are used by HEDNO are the following (fig. 17):

Artemis: Control of telemetry data (agreement of accumulated consumption with load curve per flow), production of invoice data per flow and formatting of medium voltage bills for printing and sending them to the printer. At the given time, the Artemis application serves both the parent company and HEDNO. The development of an independent application and database that will serve exclusively HEDNO, which will receive the hourly certified measurements from the Telemetry center and will make the necessary checks and corrections, is in progress by the Information and Communication division. The "Certified" indications will now be sent to the respective suppliers (including PPC).

ERP–SAP: The ERP - SAP information system supports the basic business processes of the company and is structured in "functional subsystems" (functional modules). The basic procedures supported are summarized below:

Module	Business Need Covered
SAP-ERP FI-SD	Financial Management
SAP-ERP MM-FM	Supply Chain
SAP-ERP PS-MM_EEX	Project Management
SAP-ERP HR	Human Resources Management
SAP-ERP BI	Business Information (MIS)
SAP-ERP CO	Costing
SAP-ERP SRM	Supplier Relationship Management
SAP-ERP PM	Vehicle Management
SAP-ERP BPC	Budget Preparation and Monitoring

New Payroll: New modern payroll system that fully covers all the specifics of the business. The new improved features compared to the old system are the following:
Technical Features:

- Database (not files)
- All procedures from ONLINE (not BATCH)
- Web Interface
- Ability to transfer to any platform
- Encryption of sensitive data by unauthorized users
- Separation and access to data according to user authorizations
- Ability to interface with external systems and communicate via Web Service
- Modern Development tool

Functional Features:

- Time Commitments for announcements do not exist until the calculation
- Complete Customization
- Flexibility in future directive modifications without code intervention
- Visualization of situation before and after the announcement
- Automated TSMEDE Archive (Engineers and Public Contractors Pension Fund)
- Automated APD (detailed periodic statement)
- Automated Accounts Registrations
- Use by end users without the involvement of DPLT (Information and Communication Division)
- Multiple Control Modes
- Automatic programming and sending of files to the competent services (IKA (Social Insurance Institute), banks)
- Automatic submission of announcements through files by the end user.
- Automatic calculation of all types of payroll: Tactic, Retrospectively, Compensation for unlicensed License, Gifts, Vacation allowance
- Possibility of budget payroll with scenarios

Prometheus (Employment Forms): Prometheus is the system for recording and managing the Employment Cards of the PPC and HEDNO Staff.

Hermes (customer service): It is the system of customer service and local government, accounting monitoring and pricing of electricity as well as monitoring of electricity meters. It serves approximately 7,000,000 electrical supplies through 236 offices throughout Greece for commercial and part of the technical service. The ERMIS system consists of a set of online and batch programs that run on a daily basis to meet the daily workflow of the Distribution Network. In summary, it includes the following functions:
Screens

- Customer / supply search screens
- Customer / supply / meter information screens
- Cash / invoice management screens
- Customer / owner data management screens
- New power supplies, increases, shifts, variants
- Monitoring of stages / tasks
- Participation calculation
- Issuing orders for various tasks-finals, successions
- Displays for entering emergency displays (final, successive, cash replacement, corrective)
- Counting plan management and monitoring screens
- Accounting screens-charges / credits with K.K. Network (real time and batch update)
- Cash Subsystem
- Price List Subsystem
- Common Panel Subsystem (fixed support)
- Email Subsystem
- Online monopoly charges calculation screen

Bulk Processing (Night Batch)

- Management of changes in the counting plan

- Path opening
- Preparation of indications for download in Artemis, Portable Registers, IVR, estimated pricing
- Processing of batch accounting functions
- Preparation and pre-checks of regular and emergency indications
- Invoicing of Monopoly Charges
- Closing a route
- Management and processing of Cuts
- Interfaces with other systems
- Create accounting articles and submit to ERP
- System balance
- Remote statement printing

On a non-regular basis, updates are made with the data of DT (municipal fee), DF (municipal tax), TAP (real estate tax) as well as other mass updates and data exports.

On a monthly basis, liquidations are performed (eg Third party amounts), monthly closing and on-request publication of consolidated files and statistical statements.

HEDNO now has its own independent mainframe system (IBM Mainframe z Series infrastructure) in which the ERMIS information system environment (operating system, database) is installed, adapted to its needs.

Email: Management system and provision of e-mail service with the ability to receive mail via the Internet, protection against viruses and spam. The E-Mail service is provided centrally to HEDNO users who belong to the new fully separate domain of the company.

Internet: Unified Internet access system with access filtering depending on the content and type of different sites. The service is provided through the corporate intranet. The Internet service is provided centrally to HEDNO users who belong to the new completely separate sector of the company.

Zeus: It is a service application of the HEDNO Units, regarding the management of technical works (studies, constructions, maintenance, failures, requests).

Gordios (contractors' certifications): Gordios is an application that manages the Repetitive Contracts of the Distribution. It serves all Distribution Areas and Regions with 200 users and manages 4000 measurements and 45,000 analytical tasks per month. Gordios is a pricing application for distribution contractors. In order to operate, it is supplied with the price lists of the works and materials per contract. The registration of the work and the materials consumed is done by the contractor himself via internet and then the control and the accounting are done by the local managers of the company. The application also supports price revisions for each contract.

IDE Distribution: The IDE Distribution Application supports the Daily Work Sheets of the Distribution Units required to support the Distribution Business Plans. Specifically, the hours of employment of each employee per item are recorded in detail per employee and type of work.

Metering: It is a system for importing and managing electricity of consumption metrics. With this system, the indications of the electricity meters are received on a quarterly and monthly basis by the meters that are in charge of the task of receiving the indications, while in addition, information about the status of the supply is recorded (eg. suspected of power theft, exact meter position). It is a decentralized system with 158 servers throughout Greece with local application and 1000 special laptops that receive the indications, distributed in 207 PPC offices. The number of indicators entered daily on these laptops reaches 85,000. It is interconnected with the HERMES system on the one hand to send the counts received and which will lead to the invoicing and issuance of clearing accounts and on the other hand to receive the data of the next benefits to be counted.

Iris: This is a portal of HEDNO in which the electricity suppliers (including PPC) have access with specific rights in order to receive the files of indications, bills, municipal fees, municipal taxes and TAP that concern them. The above files are produced daily after the batch of ERMIS.

Estia (recruitment of temporary staff): Application that allows the registration and processing of data for the recruitment of temporary staff of HEDNO.

Cut-Reconnection Request Management System: Management of load cash deactivation and reactivation requests (debt cut-off and reset) disconnections, reconnections and rechecks; Performance in HEDNO or contractor workshops and recording of results. The new application of Cut-offs is connected to THALIS II. Therefore, the users of cut-offs have direct knowledge of the commands given to THALIS. The new application of cut-offs has the following advantages over the old one (Skiron):

- Maintains complete historicity (the historicity of the cuts for each supply number and the historicity of the actions of the workshops for each required cut)
- Establishment of performance slips (with registration of the contractor and workshop where the slip was delivered)
- Delivery of a daily report to a workshop
- Daily Bulletin Reporting procedure with the results
- Ability to operate automatic reconnection - re-check
- Effective treatment of recalls
- Ability to automatically clear the contractor
- Ability to send tickets to contractors by mail
- Check entries - Reference tables
- Ability to operate from the Internet
- Performance statistics per contractor - workshop

- Communicates with THALIS who accepts the requests and informs about the results

Customer Indication Reception System (SYEP) (Available through the corporate SITE): This system enables the consumer to participate in the measurement of his consumption. Thus, consumers can themselves give their meter reading every four months in order to cancel the scheduled automatic process of estimating their consumption, through the computer system of HEDNO. The application contains all the necessary security provisions against any attempt of unauthorized access to the System data. The ways that a consumer can register his indication using the above information system are the following: By telephone, calling 10410, the consumer is guided by the voice command recognition system which is waiting for instructions, each time proposing the appropriate word to continue the process. The consumer can also choose to be served by a HEDNO representative by saying the word REPRESENTATIVE. Electronically through the corporate website (www.deddie.gr) From his smartphone (iPhone or android).

Scheduled Power Outages (Available through the corporate SITE): Application that allows the general public to be notified of scheduled power outages.

Position announcements: Application that allows the submission of curriculum vitae in case of vacancies for HEDNO executives.

Ptolemy: This application serves the needs of the Information and Communication division in terms of user management of all central applications, management of user requests for access to central applications, management of IT equipment and the issuance of statistics.

Management of audit reports: Planning and recording of audit findings carried out by the Internal Audit Ladder.

Travelogues: Registration of off-site travel within the temporary and regular staff of the organization and preparation of a report for the accounting of expenses.
It operates for 3 years in the Transport Sector of Aspropyrgos and it manages the entire travel Circuit:

- Issuance of a movement order
- Proof of deposit
- Approval of movement
- Clearance of KOE (Travel Expenses Statement)
- KOE control
- Accounting (for automated registrations in ERP-SAP)

The movement order and the liquidation of KOE will take place in the regions and the BOK. The control of the KOE and the accounting will be done in the management of the regions.

Thalis: Thalis application welcomes the following requests from all suppliers (including PPC):

- New representation (and representation of construction site supply)
- Power outage (voluntary by customer)
- Change items
- Meter check
- Load Meter off command (debt Cut)
- Load counter deactivation order (cancel cancellation or reconnection after debt cessation)
- Re-checking deactivated load meter (after debt cuts)
- Change of user and use
- Cessation of representation (unilaterally by supplier) Transition to PKY (default supplier) after pause
- Power outage after pause

Thalis is directly connected to the cut-and-reconnection requests management system and to HERMES of HEDNO.

Management of KOT applications - Vulnerable Consumers: The KOT (social household tariff) - Vulnerable application is available via the internet: any household invoice holder can connect to the HEDNO site and apply for a social invoice. In order to check the data of the applications from the competent bodies (OAED (Hellenic Manpower Employment Organization), GSPS (General Secretariat of Information Systems)), weekly data streams have been developed that are sent to the competent bodies using FTP in predefined files.

Nemo Q: Priority system of work at HEDNO service offices and service statistics.

Complaints System: System for managing complaints and requests of consumers and third parties to the services of HEDNO. Monitoring the progress of requests and response times.

- Manages customer complaints and requests (eg information)
- Monitors the process of moving the issue from Service to Service with time
- Monitors and notifies users about time constraints

Municipal Application: It concerns the changes of the calculation data of the municipal fees, municipal tax and TAP by the municipalities. It updates the HERMES database and it is obliged to transfer any changes resulting from the municipalities, and the alternative providers who issue invoices based on the above data. The competent HEDNO Offices are involved in the process because all the decisions for change of rates and zone prices are sent by the municipalities to the HEDNO Offices and from there with an internal document are sent to the Information and Communication division for information. The above procedure has arisen to avoid mistakes from the accumulation of a large volume of decisions by the municipalities in the Information and

Communication division and the faster processing of the above requests. It uses an application developed at the PC level to control huge increases and keep statistics.

Interconnected System RES: It is the system for managing the requests of Renewable energy sources and monitoring them at all stages. It has been developed by the Information and Communication division and has been put into production since 14/1/2013.

- It covers the procedures for the management and implementation of RES requests from the initial application (in stages) until their activation)
- It exists in every HEDNO Area
- Illustration of Electrical Systems, High and Medium Voltage substations, transformers, Medium Voltage lines (tree structure)

System for clearing the Market of Unconnected Islands: System for recording measurements and calculations of production, pricing and liquidation of NII producers. The application includes the following:

- Monitors and manages RES stations (data of station, producer, connection to substation)
- Manages the monthly readings and telemetry readings
- Calculates the monthly invoices of the producers, issues the necessary letters and sends them by mail
- Prepares the accounting report for introduction in SAP ERP
- Calculation of loss rates
- Guarantees
- Exante (energy consumption and representation percentage by supplier and electrical system)
- Calculation of network usage charges
- Withholdings in favor of third parties
- Thermal generator pricing
- Cargo representation

Network Services Quality Monitoring (Netserv): Information system for the management of requests for network services (customer details, property details, request details) concerning the Attica region, about recording of possible delays that may be due to the customer, HEDNO or third parties, calculation of delay times and finally calculation of final completion times of requests . It also provides a set of reports and csv for in-house use. It has been developed by Information and Communication division and has been in production since 8/1/2016.

Electronic Protocol application: On-line application for sending all registered documents between units of the company, as well as internal flow (workflow) within the unit. The application is accessed by the staff of the secretariat of the unit and all executives.

Fixed Telephony Accounts application: The purpose of the application is to monitor the fixed telephony accounts used by HEDNO. For all fixed telecommunications connections (telephony, internet and leased circuits) the cost center of the unit that uses them is registered once and every month the account data sent by the telecommunications provider is entered. The application has the ability to display the following elements:

- Connection manager: Displays all the phone numbers that the user has the right to see. The screen does not contain invoices.
- Charges per Account: All accounts issued (one each month) are displayed for each telephone number
- Summary Phone Calls: Displays per bill issued for each number, number, total duration and total cost per call category.
- Call Analysis: A complete analysis of calls made from a number is displayed (such as a number dialed without displaying the last three digits, date, time and duration of the call).

HEDNO Electronic Portal / HEDNO Portal: Through the HEDNO electronic portal, users have the opportunity to be informed about the company's news / announcements, have access to the central applications, submit electronic support requests, have a complete picture of the company's organization chart, etc.

Application of Power Theft: Robbery Management enables the end user to manage the robbery report in the best way so as to monitor the progress of these cases, to search and export important statistics for in-house use and at the same time to print forms and letters to consumers or suppliers, when required.

Medium Voltage fault announcements: The purpose of the application is to record all medium voltage faults related to Medium Voltage line failures or M / S YT / MT or MT / MT or ASP or CIS faults in the Medium Voltage fault information file and the daily update of the HEDNO hierarchy for the failures of the previous day with the automatic sending of email. In addition, in cases of critical failures, i.e. in cases of anomalies in important HEDNO facilities or in important customers or in a long estimated duration of restoration of customers' electricity, immediate hierarchy is provided by sending an email for the critical failure. Finally, it is possible to search the application history file and extract fault statistics to draw useful conclusions.

Announcement of Consumer Failures: The Troubleshooting application is an online application, which transmits the information of the fault announcements recorded by the answering machines that work in the fault call center in the competent units of HEDNO throughout the territory. In addition, this application records all the intermediate actions performed by the competent units of HEDNO until the final repair of the faults, as well as the general power outages, planned or extraordinary, together with the affected geographical areas.

Distance education: Information and Communication division has developed a distance learning system to meet the e-learning needs of HEDNO. This system can support both modern (requires simultaneous training of trainer and trainee) and asynchronous training (does not require simultaneous participation of trainer and trainee and participants can choose their own training time frame). In both cases the trainees receive training services regardless of where they are located. The condition, of course, is that they have access to the HEDNO intranet.

This system allows SSO (single sign on) access. It has been implemented exclusively with free software and is hosted on two Ubuntu HEDNO servers. More specifically, the software used are wordpress, moodle, Big Blue Button and the Symphony programming framework. In addition, database server, web server, file transfer server, email server and cron scheduling services have been implemented. A backup server has been implemented to make the e-learning tests and experiments run smoothly and independently of the presence of trainees on the production platform. Asynchronous e-learning servers are equipped with security certificates (access with https protocol). There is classified access to the services according to the rights of each user (visitors, students, teachers, administrators).

Seal Management: The application allows the management, processing and monitoring of telemetry benefits stamps. Through the application, a stamp is charged from the area that will be received to the craftsman who will receive it as well as in which supply it will be installed. Therefore, there is the history of the seal and the immediate information about where the seal was at any given time. The application for office procedures is offered via the internet (web) and for the actions of the craftsman via a mobile terminal type tablet.

New Counting System: The application supports the download of indications based on the counting program created daily by ERMIS. It is possible to assign specific routes to a contractor and a meter for receiving indications. Also, through the application can be counted specific benefits out of program. There is the possibility of recording the coordinates of the supply and the management of the DAK (Counter Report Sheet). The application for office procedures is offered via internet (web) and for the actions of the meter via mobile terminal type tablet.

Maintenance of YT / MT Substation Equipment: The application monitors the maintenance of the substations M / T - M/T. It maintains a register of all components of the substation network (M/S, D/I, etc.), monitors their life cycle and plans the upcoming maintenance of the components. It also supports the creation of individual maintenance and maintains information on the history of maintenance and data. It is possible to schedule maintenance in calendar form. The application is offered via the internet (web).

Geographic Network Information System (GIS): The object of the GIS project is the digital geographical representation of the Electricity Distribution Network and more specifically, the digitization, mapping and registration of geographical and descriptive data for each of the five HEDNO departments in the geographic information system database. The System has capabilities for storing, analyzing, presenting and managing information concerning the networks under the jurisdiction of HEDNO with a geographical reference. It combines information levels that can be presented with maps, tables, objects, or any information contained in a geographic item.

The main axis of design and operation of the system is the following:

- Digital Geographic Database with three main parts: Background - Networks - Projects, in which all the information related to networks and topographic background is kept. Database data management.
- Support for the various phases of network projects such as design, construction, etc.
- Ability for each user to connect, be informed, acquire or update database elements that fall within his / her competence.
- Support for exploitation applications, network development.

Pricing of RES of Interconnected Islands: The application includes the following:

- Maintenance of the RES Register regarding the licensing process and the Technical Characteristics of the operating parks.
- Pricing on a monthly basis of the energy injected in the network.
- Management of RES metering data.

Requirements for RES of Unconnected Islands: The application includes the following:

- Managing the requests of the new parks and monitoring the licensing process.
- Management of the priority order of installation and activation of parks.
- Modification of existing requests.

Calculation of Monopoly Charges: The application calculates the charges (ETMEAR, XXD, XXS and YKO) that are charged to all consumers who use the National Electrical System(i.e. electricity transmission and distribution networks, utility services and special). This calculation is used by suppliers to invoice regulated charges.

Calculation of Representation Rates (ex ante consumption calculation): The application includes the following:
Next month consumption forecast per customer based on historical consumption and infusion data or ad hoc estimate in the absence of history.
"Building" the percentage of each supplier per month from the total of the individual forecasts per customer.
Sending representation percentages to suppliers for control and any objections and then sending the final data to suppliers and LAGIE.

Periodic Clearance of Network Suppliers: The application includes the following:

- Ex post calculation of the energy charged per hour for each supplier.
- Calculation of due payments or receipts per month for each supplier.
- Periodic settlement of payments and receipts.

Search for the expiration date of the Responsible Installer Declaration (RID): With the present application, the expiration date of a RID can be sought, as stated by the licensed electrical installer in the form of YDE (Responsible Installer Statement) kept by HEDNO in its file, so that the owner / user of the supply knows when he has to re-inspect his electrical installation.

Investment Aid Statement for park producers: The application was requested to meet the need for producers to register the investment aid they received as well as all the installments they received, with the aim of calculating for each installment the amount of operating aid impairment (PALE) to be given behind the producer , and in case an infringement has been committed, a sanction should be calculated.

Consumer Questionnaire: Customer Satisfaction Questionnaire. The research has been completed.

Application of Private Insurance: The main purpose of the Private Insurance Application is the automation of the collection of data by DANP regarding the declaration of dependent members, giving employees the opportunity to electronically complete the required application.

Application of Medium Voltage Metering Data: The user (after registering) can choose from the list of supply numbers that belong to him (based on VAT), the one he wants. For each supply number, the application displays the indications history of the registers as well as their consumption curves. This application was requested to meet the need for Medium Voltage consumers to be able to see the indications of the registers as well as their curves.

Environmental Household Invoice: Implementation of environmental household electricity invoice for household consumers in the region of Western Macedonia and the municipality of Megalopolis in the regional unit of Arcadia.

Registration in HEDNO Services: The application was requested to have a single way to register in applications that require login. The applications that have been discussed to support registration are currently the Investment Support Statement, the Municipal Application and the metrics data portal.

Possibilities of power absorption of RES stations per geographical area in the Interconnected Network: The data displayed relates to the indicative current capacity of the Interconnected Network for power absorption by RES and SITHYA (High Efficiency Electricity-Heat Cogeneration) stations per geographical area.

1% return: Beneficiaries of RES return: Home benefits of the municipal / local communities where RES operate are entitled to a refund of 1%.

Loan Implementation: The Loan Management Information System was created in order to contribute to the management and liquidation of the loans of HEDNO employees and at the same time to provide online information on all loans and their development.

Fax Server application: The Fax Server application has been installed in the central infrastructure of HEDNO and has been interconnected with OTE (Hellenic Telecommunications Organization S.A) with a SIP line of 500 numbers and 30 channels. The logic of the application starts from registering users on the platform and assigning them at least one Fax number. After registration, the user has the ability to send Email from his corporate Email Account with recipient the number_receiver@fax.deddie.gr and attachment the PDF file he wants to send (mail2fax). Upon completion of the shipment, the Fax Server sends a notification to the sender with the status of its shipment. Similarly, when a Fax is sent to the number assigned to one most users; all users receive in their mail in PDF file the pages sent (fax2mail).

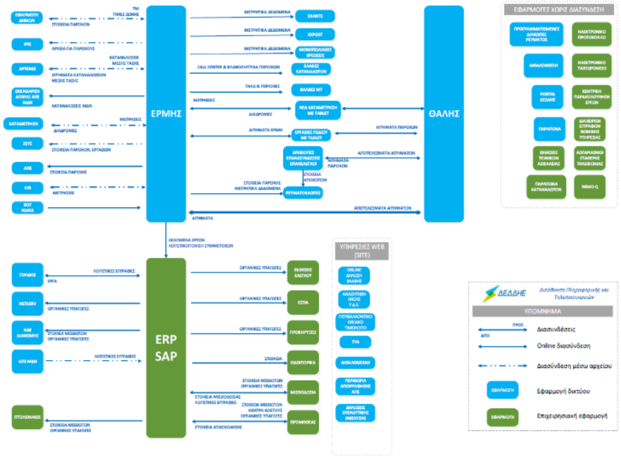

Figure 17: HEDNO Systems.

The structure of HEDNO's systems is as described above. However, HEDNO in order to respond more quickly and efficiently to all the requests of its customers (consumers and providers) through the digitization of documents, it is expected to be amended soon after the integration of the new system **Hercules**. Here are some of the most important benefits of installing the new system:

- Reduction of the number of failures in the Network and their duration as well (Enterprise Asset Management section). The workshops will work faster and more accurately in repairing the defects, thus reducing the inconvenience to consumers and / or businesses (Workforce Management module).
- All consumers will have the opportunity to know their energy profile through their smart meter and to manage this knowledge for the benefit of themselves and for the benefit of the community. Providers will be able to directly access energy data for their best business planning and commercial policy formulation (Metering and Energy Data Management module).
- The anti-social phenomenon of electricity theft will be significantly reduced with a positive impact on the budget of households, businesses and electricity providers.

Finally, it should be noted that with the implementation of the Hercules, HEDNO is moving in the direction of the Strategic Energy Technology (SET) Plan of the European Union. This encourages the consumer by giving him access to all the necessary data in a transparent, friendly and secure way to change and adapt his energy behavior, maximizing the benefits for himself and the environment in the context of the energy market.

7.4 Data Use Cases

In an energy and utility company, data originates from numerous sources and come in different formats, including grid equipment, sensors, smart appliances, smart meters, weather data, measurements from power systems, GIS data, distribution automation data, third-party data, consumer data, web data, population data, geographic information system data, DMS data, energy management system, SCADA and data related to asset management. HEDNO is using this data to bring in operational efficiencies, reduce costs, lower carbon emissions, and manage energy demand for end consumers.

Data need to be transformed into actionable insights by applying high volume data management and advanced analytics. Essentially, advanced analytics, such as predictive analytics, data mining, statistical analysis, machine learning and artificial intelligence (AI) techniques, which operate on large data sets having one or more features of big data are presented. Data need to be integrated and interoperability between different devices and control levels must be ensured. New regulation and standardized processes are required for data collection and governance. There is a lack of standards for data description and communication, essential for interoperability. Moreover, data integration and data sharing practices across institutions need to be defined for the benefit of all stakeholders.

Some data use cases are described below:

Power generation planning: HEDNO can optimize its power planning and generation using analytics. There are two key decision-making processes in power generation: power planning and dispatching of the economic load. Once we gather all the data from multiple sources, there are multiple models run on top of that data to arrive at power planning. By economic load dispatch, we mean matching energy demand with the optimal power supply from the grid over a specific time frame.

Efficient and accurate forecasting: Data analytics helps in accurately forecasting the energy consumption, which plays a pivotal role in the generation and thus, dynamic pricing. Similarly, it plays an important role in forecasting the power generation, especially for renewable energy sources, which include solar as well as wind, which gets impacted due to changing weather conditions. This all gets taken care by doing predictive analysis on all the data taken from weather systems.

Site selection: The integration of all the data, be in energy production, energy consumption, GIS, and weather data like wind direction, temperature, humidity, atmospheric pressure, cloud, and wind speed can support the sites selection where renewable power generation devices have to be installed. This improves energy efficiency as well as power output and brings in a lot of efficiencies. GIS data equally plays an important role. It includes geographical information data from satellite data or LiDAR (light detection and ranging) that helps in spatial (three dimensional) planning.

Asset management: The industry has asset-intensive units. HEDNO regularly faces a lot of asset management-related challenges. For instance, asset operations, asset monitoring, sharing of resources, asset maintenance, asset procurement and inventory management. HEDNO can achieve efficiency based on insights drawn from the analytics.

Energy efficiency: Data coming from smart meters, asset operations, business policies, and weather data can be integrated and analyzed over a period of time which helps in designing electrical devices with energy-efficiency parameters, thus reducing power requirements. Energy efficiency plays an important role to reduce carbon emissions. This also includes various other issues like equipment efficiency issues and problems in insulation, as well as improvements in operational areas. This way, HEDNO can forecast its energy consumption and predict energy savings.

8. Information Management System Project

HEDNO's mission is the operation, maintenance and development of the Distribution Network in economic terms to ensure its reliable, efficient & safe operation and its long-term ability to meet demand, while taking care of the environmental and energy efficiency. Moreover, its goals are to modernize the Distribution Network and transform it into a 'smart system' that will continually optimize the management of the connected consumers and producers, covering their emerging needs by an optimal techno-economical way.

8.1 The IMS Purpose

This Project is a key avenue and a catalyst for the HEDNO's strategic goals both in terms of its modernization and its effective adaptation to the new environment of electricity.

The IMS will include at least the following items:

1. Software for the IMS platform
2. Implementation effort for the alignment of the platform with HEDNO business and functional requirements
3. CIM compliant Adapters from third party IT/OT systems to be integrated with IMS
4. It is expected that the following IT/OT systems will be integrated :
    a. SCADA/DMS from EFACEC vendor
    b. ERP/SAP
    c. CRM/SAP IS-U
    d. Workforce Management as an implementation of SAP IS-U
    e. GIS Smallword
5. Adapters for IMS to enable functionalities of IMS and data validation
6. Integration effort
7. Upgrade and testing of new software versions of IMS should be quoted as part of non-binding maintenance agreement offer with possibility of annual renewal.

IMS as a new platform should add value to HEDNO's corporate business by providing an integrated point of data validation, network model and its visualization, network management, reporting and various what if analysis..

8.2 Business needs

Increasing amounts of Distributed Energy Resources (DER) and the emergence of electrical commodity markets are changing the face of distribution management. Advancements in communications and computing power have enabled increased situational awareness as well as made possible protection that is more complex, reliability, and optimization algorithms.

The desired goal from IMS project is an integrated, upgradable, scalable and highly flexible System of Systems (SoS) with advanced capabilities to optimally manage HEDNO's existing and future grid. HEDNO Information Management System (IMS) tender does not intend to upgrade currently available SCADA/DMS applications. HEDNO Information Management System is expected to be a modular platform integrated with IT and OT existing systems and capable for integration with future IT/OT systems.

HEDNO seeks to implement a unified and centralized Information Management System (IMS) to:

1. Enable advanced, reliable, operational, integration platform, for integration of all COMPANY grid data into one and single operational network model,
2. Have capability for reporting on both data governance as well as operational KPIs
3. Have abilities to support data analytics and what if analysis performed on created network model.

IMS shall enable HEDNO to assure that IT and OT systems developed and delivered through the rest of 12 Strategic Projects will, from the data integration point, follow at least major requirements such us:

- One unified data model shall be implemented for all relevant IT and OT systems
- Single data entry must be built into the future HEDNO IT/OT systems,
- Validation of data shall be at the point of collection, with management of data corrections.
- Data must become a shared enterprise asset and can't be owned by a single department or system.
- Cyber security must be taken into consideration, since data as an asset as well as confidentiality of data must be protected throughout all IT/OT systems and integration processes.
- Common integration framework for all OT systems must be proved by HIMS, developed on relevant industry standards (i.e. CIM, ICCP).

The core components of IMS should include:

- Core modules of IT-OT integration management, used to collect, validate and integrate data between HEDNO systems
- Integration services, consisting of ESB that is leveraging existing HEDNO environment and ICCP module used to enabling integration between HEDNO applications. As HEDNO already uses SAP environment, SAP PO/PI is intended to be used as service bus. SAP PI enables to set up cross-system communication and integration that allows to connect SAP and non-SAP systems and provides an open source environment that is necessary for complex system landscape for the integration of systems and for communication.
- Visualization tool to visually present network of entire HEDNO and support reporting
- Reporting module, in support of management decision making and reporting with user customization capabilities
- What if analysis module, in support of management decision making, and improvement of network performance, simulation and testing of various use cases needed for optimal improvement of network operation

The following are indicative high-level processes that HEDNO utilizes in the organization:

Process 1: Procedure of applying GIS incremental updates to referent IMS model.

Process 2: Procedure of GIS incremental updates being rejected by IMS validation.

Process 3: Procedure of applying latest Customer account information from CRM to referent IMS model

Process 4: SCADA feed is available in referent operational IMS model.

Process 5: Procedure of executing DMS network operations and IMS is aligned.

Process 6: Procedure of advanced IMS functionality; outage received via IVR interface and confirmed by AMI.

No	Short Indicative Description of the Business Process implemented in IMS
1	Procedure of applying GIS incremental updates to referent IMS model
2	Procedure of GIS incremental updates being rejected by IMS validation
3	Procedure of applying latest Customer account information from CRM to referent IMS model
4	SCADA feed is available in referent operational IMS model
5	Procedure of executing DMS network operations and IMS is aligned
6	Procedure of advanced IMS functionality: outage received via IVR interface and confirmed by AMI interface
7	Procedures for planned and unplanned outages
8	Procedure for new customer connection based on CRM and GIS

Table 1: Standard distribution system operator processes supported by IMS

8.3 Assessment of HEDNO Integration Capabilities

In November 2017, HEDNO conducted an assessment using to evaluate landscape and existing smart grid capabilities, the extent of smart metering, existing communications equipment. Then compare these to what is required to support smart grid use cases and future innovation with IT/OT integration. Assessment Report has been produced in December 2017 and a high level concept for IMS developed, as below in figure 17:

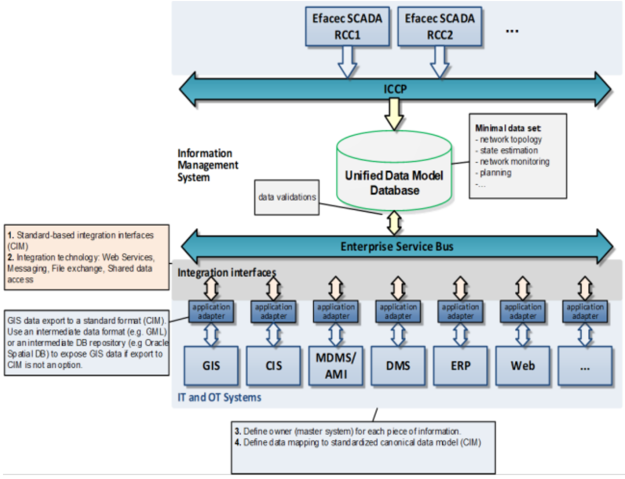
Efacec SCADA RCC1
Efacec SCADA RCC2
...
ICCP
Information Management System
Unified Data Model Database
Minimal data set
- network topology
- state estimation
- network monitoring
- planning
-...
data validations
1. Standard-based integration interfaces (CIM)
2. Integration technology: Web Services, Messaging, File exchange, Shared data access
Enterprise Service Bus
Integration interfaces
application adapter
GIS data export to a standard format (CIM). Use an intermediate data format (e.g. GML) or an intermediate DB repository (e.g Oracle Spatial DB) to expose GIS data if export to CIM is not an option.
GIS
CIS
MDMS/ AMI
DMS
ERP
Web
...
IT and OT Systems
3. Define owner (master system) for each piece of information.
4. Define data mapping to standardized canonical data model (CIM)

Figure 17: High-level concept for IMS resulting from Assessment

Below described the results of assessment addressed in the scope of IMS specification:

- The support for real-time operations and their data collection and integration with IT based on multiple sources of data, is restricted at the moment due to legacy systems
- Lack of definitions for end-to-end business processes
- HEDNO's current deployment of mainframe for integration platform is impacted by a number of

legacy applications; No ESB / integration bus use across the utility

- Real-time systems have not been prepared for integration with external ones
- Data management at HEDNO is organized across siloes due to the high number of legacy applications currently being planned for replacement
- No meta-data approach for consolidated metadata storage for:

1. Communications connectivity service
2. Customer data and Network models for distribution
3. Authorization in the integration layer based on system, not sufficiently on roles (security)

8.4 Functional specifications of the Project

Messages exchanged between the integrated 3rd party systems should be compliant and vendors are requested to provide certificate of CIM compliance. The CIM compliance certification must be provided for the IMS solution or details and certification/letter of acceptance from DSO that the deployment of their IMS solution in other DSO was based on IEC 61968 standard.

In principle, GIS should represent the master source for the network model data. However, there are other data sources that need to be merged with GIS data to complement important details, such as equipment catalog details and other asset data, delivery sites data, load data, etc. Some of this information should come from CRM, ERP or MDMS. On the other hand, IMS is also required to exchange on-line messages with corporate systems and to have ICCP link with SCADA for telemetered values.

IMS system should be designed for scalable growth as least taking into consideration existing and systems to be implemented in the next five years as shown in the figure 18 in order to facilitate integration of more business systems that HENDO S.A will implement in the future. The following figure gives an overview of systems and data flows anticipated for integration at present and in the future.

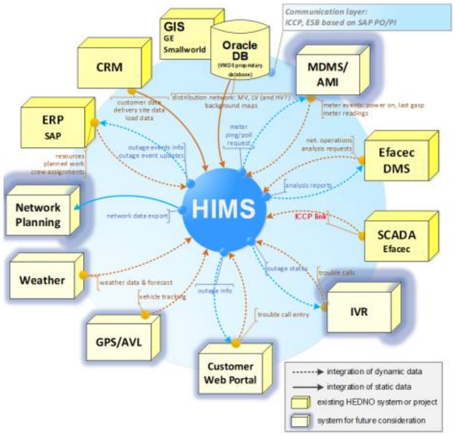

Figure 18: Data ownership and flow between HIMS and other systems (including existing and future)

It is necessary to define the master system, i.e. data owner, for each segment of data model and each integration message. Business procedures can then simply be defined on top of that operating environment and those conditions. Good practice and general recommendation for enterprise integration architecture is that single system should be the leading/master system for specific functionality and entity management, while shared responsibility should be used only if unambiguous preconditions for share of responsibilities are defined. Bi-directional integration on the level of same entity should be avoided whenever possible, because it makes any integrity update procedure (systems sync) too complex.

8.5 Components of IMS

IMS should be integrated into the existing corporate environment and assimilated into already established corporate business processes.

One of the IMS integration aspects concerns the ability to provide a solution for network model data maintenance in a complex environment where data originates from multiple external source systems and needs to be merged in the destination database. Another IMS integration aspect concerns the ability for integrating near real-time data from field devices and on-line messages exchanged with a set of dedicated/specialized corporate systems.

IMS Services should be accessible through user applications/systems and results of running core services calculations should be stored and reproduced to user applications and administrators in report form. The integration of existing, currently available IT/OT systems with IMS is considered an important part of IMS project. The adapters acting as an intermediate between the IT/OT systems and the IMS are responsible for CIM compliant messages towards the IMS and the transformation of CIM compliant messages to proprietary DB or data model of the IT/OT systems.

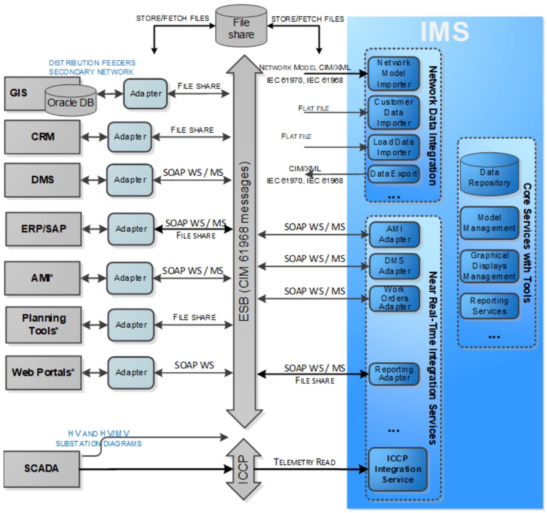

Figure 19: IMS integration interfaces and protocols (AMI, Planning and additional web portals are part of future implementation

9. Requirements of the System

9.1 The Project Purpose

The recent proliferation of utility operations systems involving advanced IT penetration and automation means utilities must consider a holistic approach towards data analytics. In other words, HEDNO S.A explores ways to derive maximum business advantages from aggregated and real-time data synthesis and trend visualization through advanced data analytics. The Smart Grid concept and its constituent technologies have added new dimensions to HEDNO by providing higher-resolution data for enhanced network operations. In HEDNO, analytics can be defined as the process of converting data from smart grid sensors and devices by integrating it with a variety of related data sets (including data from operational, non-operational and external systems) to develop models that predict and/or prescribe the next best action, thus creating deep situational awareness.

By exploiting the Big Data generated by smart meters and systems based on the Internet of Things (IoT), HEDNO will have a vast array of opportunities to improve operational and commercial efficiency using insights gained from data analytics that may not have been previously feasible. It is focused on addressing growing consumer demand and the related issues of access, availability, quality and affordability of power supply. They are just beginning to explore data analytics solutions that will eventually generate tremendous value in terms of efficiency improvements, enhance customer services and improve financial performance.

Specifically, HEDNO seeks to implement a new information system that integrates information residing in multiple IT/OT systems as well as data streams produced by smart meters in order to:

- Provide a holistic view of HEDNO’s data to better support decision making across all users (managers, technicians, customers, operators).
- Integrate with existing data sources (structured/unstructured/IoT) and operational processes.
- Support any data (relational, JSON, graph, spatial, text, OLAP, XML, multimedia) and any workload (transactions, analytics, ML, IoT, streaming, blockchain).
- Implement various analytics and visualization modules on top of HEDNO’s data.
- Consolidate, synchronize and validate data in existing HEDNO’s systems.
- Provide automated recommendations based on predictive analytics and business rules.
- Be most productive for developers and analysts (integrated microservices, events, REST, SaaS, ML, CI/CD, Low-Code).
- Connect on-premises, hybrid and cloud applications
- Enhance business productivity by automating business processes.
- Be open, scalable, flexible, B2B, and EDI.

9.2 The Architecture

The architecture should add value to HEDNO’s corporate business by providing an integrated point of data validation, visualization, BI reporting and analytics modules. The figure 20 gives an overview of the recommended architecture.

The new system will be used by different users (managers, analysts, technicians, customers, operators, public) and the recommended architecture will consist of five key layers:

1. Analysis Layer (BI Module, Machine Learning/AI, Spatial Analysis and Visualization, Stream Analytics, Data Governance/GDPR/Open Data)
2. Data Infrastructure Layer:
    - Federation/ Data Virtualization/ Interfaces
    - Persistent Data Layer (Data Lake) [(Data Warehouse, Analytics Database, Spatial Database, Unstructured Data)]
    - Transient Data Layer (Stream Engines)
3. Data Orchestration Layer (Data Validation & Data Transformations Tasks)
4. Connectivity Layer [ESB(CIM 61968 Messages), ICCP, Adapters]
5. Data sources Layer (GIS, CRM, DMS, ERP/SAP, MDMS/AMI, External, SCADA

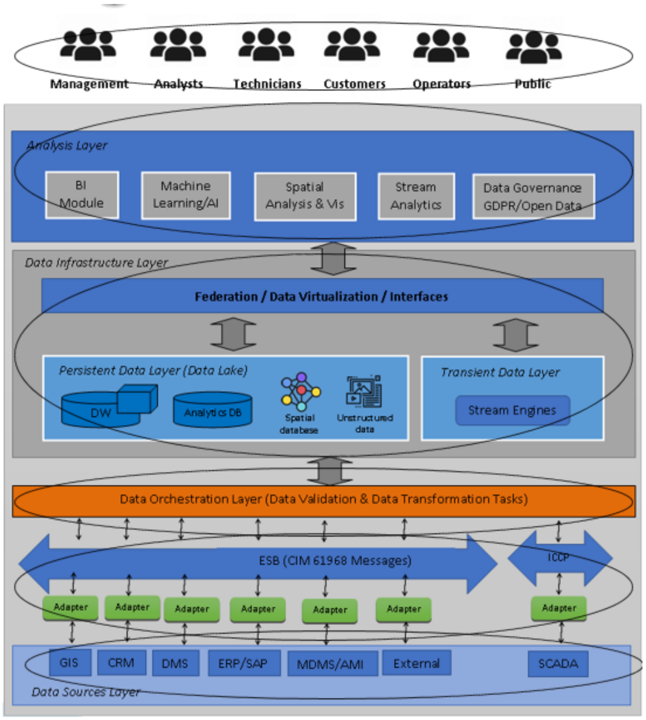

Figure 20: The Recommended Architecture

9.3 Analysis Layer

The analysis layer consists of:

- BI Module
- Machine Learning/AI
- Spatial Analysis and Visualizations
- Stream Analytics
- Data Governance/GDPR/Open Data

BI Module: This module is responsible for multi-dimensional analysis processes (roll up, drill down, slice and dice) and Key Performance Indicators (KPIs) reporting according to HEDNO needs. Dashboards shall present KPIs through widgets. Widgets shall be capable of visualizing KPIs, emphasizing when something is out of the ordinary using appropriate alert symbols. Some indicative KPIs are described below:

Transmission operating indices:

- Critical outages - up to 5 most critical outages
- Critical points - load of the most loaded section, transformer; min and max bus voltage
- Number of elements in violation – elements in violation per element type
- Outages severity – outages per severity type (good, unknown, medium, critical)
- Total active and reactive power per type – generators, consumers, network equivalents, exchanges and losses

Distribution operating indices:

- Consumption per type – active power per consumers, storages and electric vehicle stations
- Critical points – consumer voltage
- Critical points – section loading
- Critical points – supply transformer loading
- Load factor – average yearly demand / peak demand
- System load – actual power injection, DER production and total consumption
- System load look ahead – power injection, DER production and total consumption over a period
- System frequency
- System power factor

Distribution energy resources:

- CO2 reduction – achieved by DER resources
- CO2 reduction per year – achieved by DER resources
- Demand flexibility – actual demand flexibility
- Demand flexibility look ahead – system demand flexibility over a period
- Generation per type – actual DER production per technology type
- High DER production ration – feeders with the highest DER production ratio
- System DER penetration level – penetration level of renewable DERs per region and system
- System hosting capacity – hosting capacity per region and system
- Weather data – actual temperature, irradiation and wind speed per region

Quality of service:

- KPIs reported by HEDNO as per national regulation (reducing losses, improving reliability indices and customer satisfaction) that are calculated should be available to external actors (RAE) to retrieve the information without intervention.
- Customers affected / restored – number of affected and restored customers over a period
- Customers in outage / affected – number of affected and customers in outage over a period
- SAIFI - the average number of interruptions that a customer experiences daily / year-to-date
- SAIDI - measures the average duration of interruptions daily / year-to-date
- CAIDI - the average amount of time in a year that a customer's power service is interrupted

Finally, BI module should:

- Track key performance indicators in real time: machine utilization, process efficiency, costs, inventory levels, losses and cycle times.
- Help HEDNO team monitor real time data from multiple sources into one dashboard and drill down to identify issues when necessary.
- Monitor the supply chain end-to-end and identify problems and bottlenecks before they reach critical processes.
- Be able to incorporate visualization and reporting tools. Specifically, it must integrate vendor's tools, third party tools (PowerBI, Tableau, Qlik) as well as open source/PL libraries (DS3, Python, R).

Machine Learning/AI Module: This module must provide a complete framework to implement models (prediction, segmentation and market basket analysis, association, time-series, detection of anomalies and outliers) using statistical/ML algorithms (chapter 1.6, 1.7) essential for HEDNO. Moreover, it is important to support multiple input datasets with configurable join relationships and automated and customized workflow via programming; non-experts must be able to exploit ML through detecting data that best predicts outcome and best prediction algorithm for each task as well as model parameters for best performance. Thus, users' time for data preparation will be saved.

This module should also:

- Make sure models continue to perform well by identifying and assessing the influence of inputs drifting from data used to build models.
- Allow model validators and risk managers to run tests, compare candidates, document results and determine when AI/ML models are ready for production.
- Monitor production models to accomplish accuracy and fairness.

What is more, HEDNO must suggest a number of important AI tasks using vendors tools and/or open source tools-programming languages such as TensorFlow, Python, R, and Spark, which must be assigned to the contractor for implementation. Indicative AI tasks (chapter 2.7) are described below:

Predictive maintenance: This technique helps determine the condition of in-service equipment in order to estimate when maintenance should be performed. Predictive maintenance leverages real-time condition monitoring to trigger workflows of centralized data from SCADA outages associated with maintenance registry. As maintenance managers can continuously monitor limit exceptions or fault detections in real time, HEDNO can use this maintenance to eliminate unnecessary maintenance costs, reduce downtime and extend asset lifecycle, reducing overall capital costs.

Customer segmentation: It is the process of dividing customers into groups based on common characteristics, so HEDNO can market to each group effectively and appropriately.

Load forecasting: It is a technique used to predict the power/energy needed to meet the demand and supply equilibrium and generate more accurate load forecasts with improved accuracy leveraging weather, CRM, forecasting and real time data.

Theft detection: HEDNO has to handle problems with the non-technical losses caused by frauds and thefts committed by some of their consumers. In order to minimize this, the contractor should apply methodologies to perform the detection of consumers that might be fraudsters.

Load profiling: It is a way to describe the typical behavior of electric consumption, which is usually represented in time domain for load forecasting, demand-side management and capital planning.

Renewable Energy Sources 360: HEDNO must have a complete sight of the demand for renewable energy sources such as wind in order to determine supply. This integration platform must provide starting models plus weather forecasts from a weather company.

Vehicle Routings Optimization: Finding the best route is of great importance in HEDNO. The benefits are the following:

- Process Automation
- Fast Adaptation to Capacity Constraints
- Enhanced Reputation and Safety Issues
- Government Regulation Compliance
- Reduced Inventory Cost
- Adaptive and Continuous Scheduling of fleet vehicles

***Customer Sensitive Information Anonymization*:** Identifying the concerns with access to and privacy protection for energy consumption data are essential to the development of HEDNO regulations. In addition, the anonymization of customer sensitive information is vital due to the potential of consumer and authorized third party access to energy usage data through using smart grid technologies.

Spatial Analysis and Visualizations Module: Spatial analysis technologies play a major role in planning, monitoring, and managing of the network by comprehensively considering environmental and economic issues. Geospatial data has the potential to improve the service capabilities of HEDNO. From immediate equipment repair to timely maintenance, spatial analysis provides better handling of services. The utility lines are prone to damages. GIS and Remote Sensing helps in tracking the exact location of the supply lines and prevent them from any destruction. For electricity mapping, these technologies can help in the preparation of baseline data and mapping the distribution of network over the areas.

Moreover, geospatial application helps in tracking and analyzing service stoppages or power outage. Since power failures can be due to many reasons like weather event or any fault in the equipment, a map indicating physical asserts is important for repair and maintenance. From navigation to the delivery flow of utilities, GIS and Remote Sensing data are essential everywhere. Even for planning where a new network will meet the demands, GIS-based data is used.

Spatial analysis can provide an integrated platform for proper management and planning of power systems. The Spatial Analysis and Visualizations module should help in developing visual representations of geographical and environmental conditions of the site. In particular, it must represent spatially HEDNO's entities, (network, assets, people, loads, etc.); based on preparative and measurement techniques of spatial analysis with built-in and custom map layers, entities should be studied according to their topological, geometric, or geographic properties. Moreover, they should be visualized in the network connectivity model. This model must include various data such as demographics, economics, and environmental. Spatial queries must be generated to provide answers to the application using GIS in developing a model (representation of reality). Finally, this module should provide access to software tools (ArcGIS, GeoDa, CrimeSTAT) in depicting and analyzing demographic patterns spatially.

Stream Analytics Module: The use of real-time data streams and predictive analytics will give utilities intelligent decision-making capabilities to improve operations, reduce downtime, and better serve their customers. Furthermore, the existence of IoT is of great importance since it lowers the cost of experimentation, provides HEDNO with new ways to analyze their networks and allows them to adapt a model to fit specific customer needs. IoT analytics solutions must be used in order to extend the analytics infrastructure, develop new business opportunities, improve asset performance while lowering maintenance costs, make predictions using advanced predictive modeling, exploit smart meter analytics, and get a view of asset performance. Specifically, this module should:

- Provide output modes and APIs for managing streaming queries.
- Support API for static bounded data and streaming unbounded data.
- Merge streaming, interactive and batch queries.
- Run gradually and update the outputs as data streams in.
- Support file sources such as json or csv.
- Develop apps in Java, Python and R.

Real time analytics solutions can help HEDNO make better decisions, support automation processes and help customers manage their utility lifestyles. This in turn can optimize business decisions, improve operations and increase customer satisfaction.

Optimizing asset performance: By identifying potential problems in field assets and grids, HEDNO can avoid unplanned service interruptions in advance. Also, by monitoring the performance and health of assets HEDNO can predict decay points or equipment failure. Finally, real-time insights into peak periods, supply and demand analysis, and abnormal conditions can improve asset performance.

Enhancing customer experience: Intelligence from IoT assets, smart grids and SCADA enriched with customer data can provide critical insights into a customer's utility usage. HEDNO can build systems of intelligence around the consumption pattern to empower consumers with usage insights and influence their usage behavior. Segmentation at a micro level to develop a 360-degree view of customer's usage behavior will bring the best possible ways to optimize their usage. It can thus provide cost-effective plans delivering personalized experience, thereby improving loyalty, customer lifetime value and reducing customer churn. It can now provide customers with self-service intelligence capabilities to view and manage their utility usage, consumption pattern, historical usage, billing, payment and abnormal usage conditions.

Improving operational efficiency: Real-time business intelligence systems bring forth an interaction process for HEDNO's operators to proactively view and analyze operational data. BI dashboards enable them to monitor everything from meter reading to uptime for every single minute. It helps gain high visibility of historical comparison of asset performance, access real-time data and promptly react to make better decisions in a timely fashion. It also displays the overall performance of the assets in real time on a periodic basis, i.e., year-to-year, monthly, weekly, daily or hourly. These capabilities of business intelligence automate and help optimize asset performance, reduce risk and operational costs, and enhance business responsiveness.

Prevention of utility loss and fraud management: As enormous volumes of data streams driven from smart meters, grids and sensors become readily available, HEDNO can perform real-time and predictive analytics to gain critical insights. This can help operators continuously monitor the vital signs of utility distribution systems, reliably and quickly assess system integrity, and gain hidden insights. This in turn helps them in outage and fault detection, and to determine anomalies on supply distribution lines, thereby optimizing utility distribution. Thus minimizing utility loss and providing early detection of supply, theft, leaks, consumption and over usage.

Data Governance/GDPR/Open Data Module: HEDNO complies with the legislation on personal data protection (L.2472/97) and/or Regulation 2016/679 (GDPR) of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Data Governance must be taken into consideration, since data as an asset as well as confidentiality of data must be protected throughout all IT/OT systems and integration processes.

The new system should support data portability, so consumers be capable of transferring their energy usage data, account information and billing information to any third party service provider, such as a smartphone app, a demand response provider or a commercial building energy management system. Moreover, it is crucial some of the energy data resulted from power systems to be open to the general-public. Before publication, the contractor must use suitable methods to anonymize this data so that there is no direct correlation with HEDNO. There are four reasons why these energy data should be published:

- Improved quality of science
- More effective collaboration across the science-policy boundary
- Increased productivity through collaborative burden sharing
- Profound relevance to societal debates

9.4 Data Infrastructure Layer

Federation/ Virtualization/ Interfaces

Nowadays, the ability to view, access, manipulate and analyze data without the need to know or understand its physical format or location, and without having to move or copy it, is vital. HEDNO needs a data virtualization platform that will allow it to build and manage virtualized views of different entities, which access, transform and deliver the data required to accelerate revenue, reduce costs, lessen risk and improve compliance.

Furthermore, a data virtualization layer is of great importance as ad hoc schemas for data scientists should be created and entity-specific extraction of data should be supported. This layer must communicate with the data sources layer. Data virtualization can help HEDNO turn their data into analytic insights. Since, data virtualization can combine data in real time, even across structured, semi-structured, or unstructured sources, it can support real-time analytics based on streaming data, social media data, or sensor based data. The inherent agility and flexibility of data virtualization enables IT teams to alter the data architecture on an as-needed basis, as newer technologies and infrastructures emerge.

Federation/ Virtualization/ Interfaces layer should also provide API management in order to work with any host, API, and scale; improving and publishing the APIs securely as well as connecting to systems hosted anywhere is essential.

Persistence Layer

The Reporting and Analysis functions of Business Intelligence will be deployed in HEDNO by first combining data from different systems. The extracted and filtered data will be transformed into operational data warehouses by ETL (Extract, Transform and Load) processes. Then, techniques such as OLAP cubes will be used in order to enhance the data storage to deliver better analytical performance. The multidimensional analysis provided by OLAP tools helps analysts "slice and dice" relationships between different variables within different levels of their own hierarchies.

Since OLAP queries are generally organized around partial aggregations along the different dimensions, the data can be organized along the different dimensions. Data warehouse must manage the dimensions, measures and fact tables of the star schema (a model that uses the star technique reflects the way a user sees the data) some of which are the following: network, assets, people, loads, workshop, amount of bill, total sales, supply, requests, metering, municipality, and power outage. What is more, various third party tools should be integrated such as Tableau and PowerBI. These tools will provide interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards (chapter 1.9). Finally, advanced analytics engines should be contained like Hadoop, Spark and/or NOSQL databases (e.g. graph databases).

Transient Data Layer

This layer must involve stream engines. It is crucial for HEDNO to handle data streams attributed to SCADA, smart meters and events (IVR, web). For this reason, this layer must involve stream engines such as:

- Azure Event Hubs
- AWS Kinesis
- Kafka (Confluent)
- SAP Streaming Analytics
- Others, open source (e.g. Spark streaming)

9.5 Data Orchestration Layer

In this layer, low impact capture, routing, transformation, and delivery of database transactions across homogeneous and heterogeneous environments in real time with no distance limitations must be provided. Moreover, the data originated from data sources layer will be validated and transformed through various tasks.

Specifically, data validation should occur to confirm whether the data pulled from the sources has the correct/expected values in a given domain (such as a pattern/default or list of values). If the data fails the validation rules, it should be rejected entirely or in part. The rejected data is ideally reported back to the source system for further analysis to identify and to rectify the incorrect records.

During transformation, a set of rules or functions will be applied on the extracted data to convert it into a single standard format. The data transformation tasks are important so as to:

- Populate data stores at Data Infrastructure Layer.
- Transfer, exchange and synchronize data between existing systems.

9.6 Connectivity Layer

The connectivity layer should contain the following:

Enterprise Service Bus (ESB) is based on a great number of approaches to manage communication between different kinds of systems such as GIS, etc. Decoupling and providing durability are primary drivers for using ESB. Moreover, ESB must provide:

- a place to ensure safety and compliance.
- high availability through built-in disaster recovery and geo-replication.
- handling of poison messages.
- detection and delete of duplicate messages.
- load balancing and scalability to improve performance and failover support.

In electric power transmission and distribution, the **Common Information Model (CIM)** is a standard developed by the electric power industry that has been officially adopted by the IEC, which aims to allow application software to exchange information about an electrical network. The CIM is currently maintained as a UML model. It defines a common vocabulary and basic ontology for aspects of the electric power industry.

Messaging represents communication systems based on exchanging messages. These messages include data and other information from different applications managed by messaging server.

For the purpose of integrating the infrastructure layer with EFACEC SCADA systems across HEDNO and exchanging the SCADA data with other IT/OT systems, the **Inter-Control Center Communications Protocol (ICCP)** or Telecontrol Application Service Element 2 (TASE.2) will be used; it is defined with the IEC 60870-6 standard series and is based on client server architecture. ICCP protocol implementation must:

- Support procedures, extensions and algorithms.
- Support message level authentication and encryption of the messages.
- Secure functionality must include the capability for simultaneous secure and non-secure associations.

The implementation of near real time integration component must start with the installation and configuration of ESB and subsequently ICCP, implementation of data interfaces. SAP PO/PI is intended to be used as a service bus. SAP PI enables to set up cross-system communication and integration that allows connecting SAP and non-SAP systems and provides an open source environment that is necessary for complex system landscape for the integration of systems and for communication. ESB will use appropriate adapters that will translate messages and application data models into a common schema and IEC 61968-100 compliant messages; it needs to be scalable to support integration APIs.

HEDNO does not have accurate information about which customers are connected to which electrical phases at their substations. Hence, a **network connectivity model** must be deployed to avoid imbalance in loading on their network, limit outages and make fault restoration cheaper. With this model HEDNO will have the ability to identify how the customers and assets are interconnected together downstream of a substation. The benefits of this model are:

- Accurate Outage Reporting (compliance)
- Improved Fault Detection, Isolation and Restoration
- Balanced Feeders (efficient operations)
- Accurate Power Flow Analysis
- Accurate Load Balancing
- Energy Loss/Theft Detection
- Finding mismatches between records and reality, with very good accuracy
- Customer satisfaction

9.7 Data Sources Layer

The HEDNO's data sources are its operational and information systems. In data management terms, the overall information storage system is the “system of record” and is a trusted data source. Due to the nature of these purpose-built systems, they tend to be siloed, that is, the data tend to be stored in different ways without a common organizational principle. The main HEDNO's data sources are in the table 2.

	Information Technology (IT) Systems	**Operational Technology (OT) Systems**	**Third Party Systems**
Input Sources	SAP/ERP software CRM Call center	▪ SCADA (Supervisory Control and Data Acquisition) system ▪ GIS (Geographic Information System) ▪ AMI/MDMS (Advanced Metering Infrastructure/Meter Data Management System) ▪ DMS (Distribution Management System) ▪ OMS (Outage Management System)	▪ Weather forecast systems ▪ Social media ▪ Websites
Volume	Medium	High	Low
Data Type	Structured	Structured/Semi-Structured	Unstructured

Table 2: HEDNO's Data Sources

Geographic Information System (GIS)

HEDNO utilizes GE's Smallworld GIS, including Oracle database allowing access for different external users to network data. This GIS contains a geographic-based connectivity model of the complete Distribution Network.

Customer Relationship Management system (CRM)

HEDNO is tendering for new CRM with ambition to replace a number of legacy applications and mainframe in this IT initiative. It is expected that the suite of SAP IS-U with SAP PO will be the new CRM, under the assumption of successful technical and financial evaluation. The CRM will send customer information to the SCADA-DMS spontaneously, so customer data can be maintained inside DMS and available for the operators.

ERP/SAP

The ERP - SAP information system supports the basic business processes of the company and is structured in "functional subsystems" (functional modules). The basic procedures supported are summarized below:

- Resources and crew model
- Planned work orders
- Crew assignments
- Planning and Business processes

Meter Data Management System (MDMS)/ Advanced Metering Infrastructure (AMI)

HEDNO uses two MDMs, but a new system is expected to be implemented. This interface shall provide the capabilities to:

- Allow operators to ping meters to verify "Power On" at the customer location
- Receive last gasp notifications as well as restoration messages

The source of the data is always the AMI system and the interface on the Information Management System towards DMS/OMS.

Supervisory Control and Data Acquisition (SCADA)/ Distribution Management System (DMS)

SCADA and DMS systems installed in Remote Control Centers (RCC) are based on EFACEC solution. HEDNO operates EFACEC SCADA in different regions. This system includes the communication Front Ends that perform the data acquisition and control output with all the field RTUs and devices. The SCADA should be interface to the real-time data from the substations shared by Distribution Operations. The interface shall use a bidirectional secure ICCP link between both systems.

External Systems

- Weather Data
- Social Media
- Web Services

9.8 Other Issues

Scalability: Scalability is the measure of a system's ability to increase or decrease in performance and cost in response to changes in application and system processing demands. As the HEDNO's system grows (in data volume, users, data sources), there should be reasonable ways of dealing with that growth. So, scalability is essential in that it contributes to competitiveness, efficiency, reputation and quality.

High Availability/Fault Tolerance: HEDNO's systems should be designed to be fault tolerant and of no less than 99.9% available. Hence, they can continue to operate even when an error, exception, or invalid input encounters, as long as they have been designed to handle such errors rather than defaulting to reporting an error and halting. They can also use replication to provide fault tolerance; a critically important database can be continuously replicated to another server, so that if the server hosting the primary database goes down, then operations can instantly be redirected to the replica database.
The architecture shall be considered unavailable (the time to be calculated as unavailable during testing) when:

- Database is not accessible (i.e. lost connection or no data).
- Any integration interface is not functioning (i.e. data from SCADA systems are not updated).
- Incremental or total import of data is not possible.
- User cannot log in.
- User with administrative rights cannot create new user, or change access rights for existing user.
- Users from remote locations cannot log in.
- Any function does not perform as it is specified and proved during the testing periods.

Users/Roles/Access Rights: The new system of HEDNO must restrict a network access based on the roles of individual users within it. It will let employees have access rights only to the information they need and prevent them from accessing information that doesn't pertain to them. Limiting network access is significant for HEDNO as it has many workers, employs contractors and permits access to third parties, like customers and vendors, making it difficult to monitor network access effectively. Since it is a role based control access company, it is of great importance to secure their sensitive data and critical applications.
Security System: Enhanced safety and security of supply should be provided through more complete, accurate and timely information at key decision points. Solution to be proposed should meet all the required features at least comply with the following IT security requirements:

- Capability to operate over encrypted communication protocols between the various architectural components and functional subsystems.
- Block unauthorized access by checking inputs for each operation that it has been deemed necessary.
- Enforce (dynamic) authorization rules for viewing/updating data.
- Support authentication protocols based on open standards to enable future interoperability with a single access system for HEDNO's applications.
- Meet requirements of secure system development (security by design and by default), as well as personal data and privacy requirements (privacy by design and by default).

System Performance: The contractor should suggest some performance indicators so as to evaluate how well the system constituents perform based on standardized benchmarks like latency, loading and response time.

Cloud/ On Premises: The new platform must integrate both cloud and on-premises applications. Moreover, it is important to:

- Publish and manage APIs whether those run in the cloud or on-premises.
- Support communication via events.
- Ensure communication of APIs and integration technologies asynchronously, even across diverse technology platforms.

Specifically, a combination of cloud and on premises solution can:

- Allow data scientists to work the way they want to, and provide access to automated workflows, the best of open source, and a streamlined approach to building models.
- Enable data science teams to work together with ways to share and reproduce models in a structured, secure way for enterprise-grade results.
- Provide a fully managed platform built to meet the needs of the modern enterprise.

Simplicity:

- Single repository for all types of data.
- Proven, robust and easy to use integration & analytics tools.

Automation:

- Minimal administration cost.
- Minimal risk regarding security, availability, performance.

Single Data Entry: Currently, in some cases, data entry of the same entity has to be done in multiple systems - e.g. an asset in GIS/SCADA. That requires to rely on multiple third party software platforms that need to be accurately communicating with each other at all times. Single point of entry means that there is one access point for the central station automation software. Hence, it is essential for HEDNO so as to avoid:

- Duplicate Data Entry: When managing multiple software components that do not communicate with each other, there is no way to get around duplicate data entry. With one single point of entry, the data are entered once.
- Database Costly Errors: When working with different platforms, the risk of errors is greatly increased.

References

[1] Pullum, Laura L., et al. “Big Data Analytics in the Smart Grid.” *IEEE*, 2 Nov. 2017, https://smartgrid.ieee.org/images/files/pdf/big_data_analytics_white_paper.pdf

[2] Arass, M. El, et al. “Data lifecycles analysis: Towards intelligent cycle.” *IEEE*, 2 Apr. 2017, https://ieeexplore.ieee.org/document/8054938

[3] Chauhan, Rajeev Kumar, et al. “Intelligent SCADA System.” *ResearchGate*, Jan. 2010, https://www.researchgate.net/publication/313477795_Intelligent_SCADA_System

[4] AllumiaX Staff Engineers. “SCADA and Its Application in Electrical Power Systems.” *Allumiax*, Aug. 2020, https://www.allumiax.com/blog/scada-and-its-application-in-electrical-power-systems

[5] Naeem, Tehreem. “Database Management Software: Features, Types, Benefits, and Uses.” *Astera*, 11 Feb. 2021, https://www.astera.com/type/blog/database-management-software

[6] Vassiliadis, Panos. “A survey of Extract–transform–Load technology.” *ResearchGate*, Sep. 2009, https://www.researchgate.net/publication/220613761_A_Surveyof_Extract-Transform-Load_Technology

[7] Zhao, Bo. “Web Scraping.” ResearchGate, May 2017, https://www.researchgate.net/publication/317177787_Web_Scraping

[8] Saurkar, Anand V., et al. “An Overview On Web Scraping Techniques And Tools.” *ISSN*, Apr. 2018, http://www.ijfrcsce.org/download/browse/Volume_4/April_18_Volume_4_Issue_4/1524638955_25-04-2018.pdf

[9] “Statistical Analysis.” *SaS*, https://www.sas.com/en_us/insights/analytics/statistical-analysis.html

[10] “What is Artificial Intelligence in the Energy Industry?” *next*, https://www.next-kraftwerke.com/knowledge/artificial-intelligence

[11] Marr, Bernard. “The 9 Best Analytics Tools For Data Visualization Available Today.” ©*Bernard Marr & Co*, 2020, https://www.bernardmarr.com/default.asp?contentID=2051

[12] Caldarola, Enrico Giacinto, and Antonio Maria Rinaldi. “Big Data Visualization Tools: A SurveyThe new paradigms, methodologies and tools for large data sets visualization.” ResearchGate, Jul. 2017, https://www.researchgate.net/publication/318679685_Big_Data_Visualization_Tools_A_Survey_-_The_New_Paradigms_Methodologies_and_Tools_for_Large_Data_Sets_Visualization

[13] Hoofnagle, Chris Jay, et al. “The European Union general data protection regulation: what it is and what it means.” *Taylor and Francis Online*, 10 Feb. 2019, https://www.tandfonline.com/doi/full/10.1080/13600834.2019.1573501

[14] European Distribution System Operators. “E.DSO Policy Brief on Open Data.” *E.DSO*, Dec. 2018, https://www.edsoforsmartgrids.eu/wp-content/uploads/EDSO-Open-Data-Policy-Brief_1812_final-1.pdf

[15] Sumic, Dr. Zarko. “Business Process Management in Energy.” *UtilitiesProject*, https://mthink.com/legacy/www.utilitiesproject.com/content/pdf/utp6_wp_sumic.pdf

[16] Alfa Consulting. “7 upcoming technological innovations on Energy Distribution.” *Alfa Beyond Consulting*, 7 Feb. 2018, https://alfaconsulting.com/en/7-upcoming-technological-innovations-on-energy-distribution/

[17] Gökalp, Mert Onuralp, et al. “Big-Data Analytics Architecture for Businesses: a comprehensive review on new open-source big-data tools.” Oct. 2017 https://www.researchgate.net/publication/320719910_Big-Data_Analytics_Architecture_for_Businesses_a_comprehensive_review_on_new_open-source_big-data_tools

[18] Hassan, Najmul, et al. “The Role of Edge Computing in Internet of Things.” *ResearchGate*, May 2018, https://www.researchgate.net/publication/323268025_The_Role_of_Edge_Computing_in_Internet_of_Things

[19] Fernández, Cristian Martín, et al. “An Edge Computing Architecture in the Internet of Things.” *IEEE*, 31 Jul. 2018, https://ieeexplore.ieee.org/document/8421152

[20] Mangat, Mona. “Edge Computing vs Cloud Computing: Key Differences.” *phoenixNap*, 2 Dec. 2019, https://phoenixnap.com/blog/edge-computing-vs-cloud-computing

[21]Xue, Guoliang, et al. “Smart Grid — The New and Improved Power Grid: A Survey.” *ResearchGate*, Jan. 2012, https://www.researchgate.net/publication/260670952_Smart_Grid_-_The_New_and_Improved_Power_Grid_A_Survey

[22] Zheng, Jixuan, et al. “Smart Meters in Smart Grid: An Overview.” *IEEE*, 4-5 Apr. 2013, https://ieeexplore.ieee.org/document/6520030

[23] Mohassel, Ramyar Rashed, et al. “A Survey on Advanced Metering Infrastructure and its Application in Smart Grids.” *ResearchGate*, Jan. 2014 https://www.researchgate.net/publication/265905362_A_Survay_on_Advanced_Metering_Infrastructure_and_its_Application_in_Smart_Grids

[24] Trong, Nghia Le, et al. “Advanced Metering Infrastructure Based on Smart Meters in Smart Grid.” *IntechOpen*, 29 Jun. 2016, https://www.intechopen.com/books/smart-metering-technology-and-services-inspirations-for-energy-utilities/advanced-metering-infrastructure-based-on-smart-meters-in-smart-grid

[25] International telecommunication union. “Simplified Domain Model in ICT perspective.” *Atelim*, 27 Jun. 2016, https://atelim.com/international-telecommunication-union-v2.html?part=2

[26] Bhattarai, Bishnu P., et al. “Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions.” *IET*, 1 Mar. 2019 https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/iet-stg.2018.0261

[27] Circutor, “Advantages of Smart Grids.” *Circutor*, http://circutor.com/en/documentation/articles/4162-advantages-of-smart-grids

[28] Zhang, Yang, and Tao Huang. “Big data analytics in smart grids: a review.” ResearchGate, Aug. 2018 https://www.researchgate.net/publication/326996236_Big_data_analytics_in_smart_grids_a_review

[29] Mohassel, Ramyar Rashed, et al. “A survey on Advanced Metering Infrastructure.” ResearchGate, Dec. 2014 https://www.researchgate.net/publication/263699831_A_survey_on_Advanced_Metering_Infrastructure

[30] Krogstie, John, et al. Conceptual Modelling in Information Systems Engineering.Springer,2007. https://link.springer.com/chapter/10.1007/978-3-540-72677-7_3

[31] Chaudhuri, Surajit, et al. “An Overview of Business Intelligence Technology.” COMMUNICATIONS OF THE ACM, Aug. 2011, https://cacm.acm.org/magazines/2011/8/114953-an-overview-of-business-intelligence-technology/fulltext

[32] Gallo, Crescenzio, et al. “Data Warehouse Design and Management: Theory and Practice.” IEEE, Jul. 2010, https://core.ac.uk/download/pdf/6342303.pdf

[33] Maguire, James. “Top 10 Benefits of a Data Warehouse.” Datamation, 15 Jun. 2020, https://www.datamation.com/big-data/top-10-benefits-of-a-data-warehouse/

[34] “25 BEST Data Warehouse Tools in 2021 (Open Source & Paid).” GURU99, 2021, https://www.guru99.com/top-20-etl-database-warehousing-tools.html

[35] Davis, Judith R., and Robert Eve. *Going Beyond Traditional Data Integration to Achieve Business Agility.* Nine Five One Press, 2018, https://business-iq.net/assets/4318-going-beyond-traditional-data-integration-to-achieve-business-agility

[36] “Data Virtualization.” *ScienceDirect* https://www.sciencedirect.com/topics/computer-science/data-virtualization

[37] Chatziantoniou, Damianos. “Data Virtual Machines: A Novel Approach to DataVirtualization in Big Data Environments.” *Course Hero*, https://www.coursehero.com/file/74238172/DVM-A-Novel-Approach-to-Data-Virtualizationpdf/

[38] Bologa, Ana-Ramona, and Razvan Bologa. “A Perspective on the Benefits of Data Virtualization Technology.” *ResearchGate*, Jan. 2011, https://www.researchgate.net/publication/227363869_A_Perspective_on_the_Benefits_of_Data_Virtualization_Technology

[39] Timothy, King, “The 13 Best Data Virtualization Tools and Software for 2020.” *Solutions Review*, 10 Nov. 2020, https://solutionsreview.com/data-integration/the-best-data-virtualization-tools-and-software-for-2020/

[40] “Gotland's Smart Grid.” *VATTENFALL*, 10 Nov. 2019, https://network-solutions.vattenfall.co.uk/case-study/gotland-smart-grid

[41] “ACON SMART GRIDS.” *ACON*, https://www.acon-smartgrids.cz/upload/Brozura_Acon_EU_web.pdf

[42] Establishing the Smart Grid in Austria.” *nes,* https://www.networkedenergy.com/en/success/establishing-the-smart-grid-in-austria

[43] Jowitt, Tom. “Microsoft Goes Green With Agder Energi Smart Grid Project.”, *Silicon*, 7 Oct. 2016, https://www.silicon.co.uk/e-innovation/green-it/microsoft-smart-grid-project-198806?print=print&cmpredirect

[44] “Optimized microgrids deliver a smarter approach to energy delivery.” *IBM,* Feb. 2020, https://www.ibm.com/case-studies/cleanspark-analytics

[45] “Streamlining optimization and data science workflows to help system operation for the Canary Islands.” IBM, Oct. 2018, https://www.ibm.com/case-studies/red-electrica-de-espana-hybrid-cloud-data-science

[46] “Turning up cost efficiency and output for windfarms with predictive maintenance solutions in the IBM Cloud.” *IBM*, Mar. 2018, https://www.ibm.com/case-studies/performance-for-assets-ibm-cloud-energy

[47] “Keeping the lights on and power grids stable in Belgium.” *sas*, https://www.sas.com/hu_hu/customers/eandis-be.html

[48] “Combating energy theft with analytics.” *sas,* https://www.sas.com/en_us/customers/cemig-br.html

[49] Hill, Stephen. “How Uttar Pradesh is meeting the challenges of a smart meter rollout.” *Utilities Blog*, 22 Dec. 2020,

https://blogs.oracle.com/utilities/post/how-uttar-pradesh-is-meeting-the-challenges-of-a-smart-meter-rollout

[50] “Generating power – and more value for customers.” *sas*, https://www.sas.com/th_th/customers/odec.html

[51] “Southwest Power Pool.”, *informatica*, https://www.informatica.com/it/about-us/customers/customer-success-stories/southwest-power-pool.html

[52] “Enverus Pumps Data-driven Applications Faster Using Denodo’s Data Virtualization Platform.”, denodo, https://www.denodo.com/en/customer/enverus

[53] Ritterbusch, Einar. “AI analysis of big data prepares Denmark for a greener future.” *IBM*, 26 May 2020, https://www.ibm.com/blogs/client-voices/denmark-advances-renewable-energy-big-data-ai/

[54] “Energy Utility's Oracle and SQL Server Migrations Revitalize - Data Infrastructure with EDB Postgres and Virtualization.” *EDB*, https://www.enterprisedb.com/resources/case-studies/energy-utilitys-oracle-and-sql-server-migrations-revitalize-data

[55] Cisco Public Information. “An IT Transformation for Power Outage Management.” *Cisco*, https://www.cisco.com/c/dam/en/us/products/collateral/analytics-automation-software/data-virtualization/li-power-casestudy.pdf

[56] “TransAlta Case Study: A Cloud Modernization Story.” *Denodo* https://www.denodo.com/en/video/case-study/transalta-case-study-cloud-modernization-story

[57] HEDNO, *deddie* https://www.deddie.gr /

[58] “INFORMATION MANAGEMENT SYSTEM.” *IMS-Technical Specification Final*

[59] “Μητρώο Εφαρμογών ΔΠΛΤ.” *ΔΕΔΔΗΕ*, 22 Sep. 2020