Big Data for Effective Management of Smart Grids

Alba Amato and Salvatore Venticinque

Abstract The Energy industry is facing a set of changes. The old grids need to be replaced, alternative energy market is increasing and consumers want more control of their consumption. On the other hand, the ever-increasing pervasiveness of technology together with the smart paradigm, are becoming the reference point of anyone involved in innovation, and energy management issues. In this context, the information that can potentially be made available by technological innovation is obvious. Nevertheless, in order to turn it into better and more efficient decisions, it is necessary to keep in mind three sets of issues: those related to the management of generated data streams, those related to the quality of the data and finally those related to their usability for human decision-maker. In smart grid, large amounts of and various types of data, such as device status data, electricity consumption data, and user interaction data are collected. Then, as described in several scientific papers, many data analysis techniques, including optimization, forecasting, classification and other, can be applied on the large amounts of smart grid big data. There are several techniques, based on Big Data analysis using computational intelligence techniques, to optimize power generation and operation in real time, to predict electricity demand and electricity consumption and to develop dynamic pricing mechanisms. The aim of the chapter is to critically analyze the way Big Data is utilized in the field of Energy Management in Smart Grid addressing problems and discussing the important trends.

Keywords Big Data · Smart Grid · Computational intelligence · Dynamic energy management · Predictive analytics · Peer to Peer · SCADA · Interoperability, NOSQL, learning theory

A. Amato (✉) · S. Venticinque
Department of Industrial and Information Engineering, Second University of Naples, Caserta, Italy
e-mail: alba.amato@unina2.it

S. Venticinque
e-mail: salvatore.venticinque@unina2.it

1 Introduction

Smart grids are an evolution of the existing power distribution networks. They respond to the growing demand for energy, and the availability of several solutions of renewable energy sources that have stimulated the formulation of plans aiming at expanding and upgrading existing power grids in several countries. A fundamental characteristic of the smart grid is also the ability to manage, via protocols and information flows, generators and loads active in the network, coordinating them to perform certain functions in real-time as, for example, to cope with a peak, balance the load of a power supply or make up for a sudden drop in voltage by drawing more districts where there is a surplus. By linking information technologies with the electric power grid to provide "electricity with a brain" the smart grid promises many benefits, including increased energy efficiency, reduced carbon emissions, and improved power reliability, but there is an urgent need to establish protocols and standards [1].

Nevertheless, in order to turn it into better and more efficient decisions, it is necessary to keep in mind three sets of issues: those related to the management of generated data streams, those related to the quality of the data and finally those related to their usability for human decision-maker.

The recent introduction of smart meters and the creation of the Smart Grid, the first of which constitutes one of the fundamental elements, has completely revolutionized the utility system. From a processing point of view of a smart grid is inexhaustible and valuable data source in order to analyse time series, crossing them with weather data to make predictions about the electricity consumption of long or very short time; Distribute more efficiently the supply on the territory; Evaluate the quality of service provided in real time; Analyse time series to prevent potential failures and to intervene promptly; Produce more accurate bills making more conscious consumer. Smart Grid therefore generates a stream of data that must be captured, processed and analyzed efficiently from all business areas. So the data obtained, together with the data coming from more traditional sources, can be used to perform several types of analysis such as churn analysis, or research and development of new tariff plans depending on the type of consumer or fraud detection.

In fact, in smart grid, large amounts of and various types of data, such as device status data, electricity consumption data, and user interaction data are collected. Then, as described in several scientific papers [2], many data analysis techniques, including optimization, forecasting, classification and other, can be applied on the large amounts of smart grid big data [3]. There are several techniques, based on Big Data analysis using computational intelligence techniques, to optimize power generation and operation in real time, to predict electricity demand and electricity consumption and to develop dynamic pricing mechanisms [4]. Computational Intelligence can provide effective and efficient decision support for all of the producers, operators, customers and regulators in smart grid enabling all those stakeholders to have more control over the energy utilization. The aim of the chapter is to critically analyze the way Big Data analysis using computational intelligence techniques is utilized in the field of Energy Management in Smart Grid addressing problems and

discussing the important trends. In particular computational intelligence techniques can be used to overcome the challenges posed by large and complex software systems. Computational intelligence aims to produce good solutions to problems in a reasonable amount of time and it is widely used for several real world applications and problems that are widespread in Smart Grid field, e.g., routing problems; assignment and scheduling problems; forecasting problems; etc.

The first part of the chapter presents overview, background and real life applications of Internet of Energy with particular emphasis on Smart Grid. Successively the state of the art of Big Data Analytics issues in IoE are discussed together with the current solutions and future trends and challenges.

The paper is organized as follows: in Sect. 2 smart energy concepts are presented, Sect. 3 introduces big data properties of smart grid and some important information in order to better understand the problem. Section 4 presents an overview of research lines and research project dealing with energy management in Smart Grid. Conclusions are drawn in Sect. 5.

2 Smart Grids and Smart Micro-Grids

Smart grids is now part of a wider smart energy concepts that includes not only the provisioning of intelligences to the power grid by an ICT solution, but also management of smart buildings, the monitoring and analysis of user's information, user's devices, environmental parameters and others [5].

In the context of smart energy, it is increasingly spreading the idea that the road leading to the reduction of global energy consumption depends of the capability to deliver usable information in the hands of energy managers, users and consumers for effective decision making. This would require lower investment than advancing the power grid or green restructuring buildings. In fact an immediate and effective solution to obtain improvements in fuel consumption and emissions appears to ensure that the existing infrastructure is at its maximum efficiency. To reach this goal it can certainly be useful the creation of a data analysis system, which is a convenient solution also from an investment perspective of their occupants. Such capability relies on innovative ICT solutions which can extract the information potentially contained in data collected by advanced metering technologies and deliver effectively it to applications and users.

The logic architecture of a smart grid is shown in Fig. 1 that is also introduced in [6]. The fabric layer of smart grids is composed of a number of smart meters and controllers, which collect data, perform simple elaboration and feed directly SCADA and other types of systems. The more representative example is the network of downstream detection instruments, which connect the user's power grid to the distribution network, and are used to measure the consumption on the various branches of a network, in particular for billing purpose. They can be also used to check and control power peek, but do not allow for improving energy efficiency, in fact usually measures are not collected and processed in real time for management purpose. These
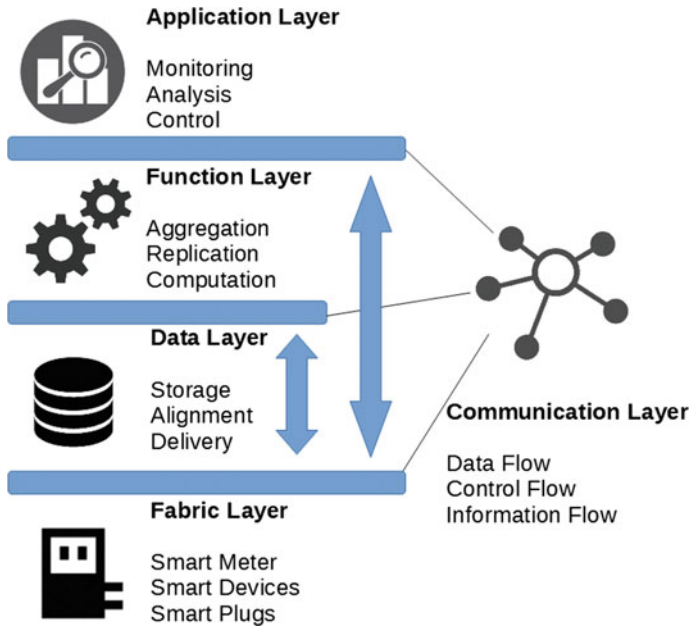
Fig. 1 Layered model of smart energy

usually represent the leaves of an energy provider and provide the finest grain of information.

On the other hand the spread of smart metering systems is growing as the information technology is assuming a central role in the technologies for energy management sector also at the building dimension. The instruments to detect energy consumption are now reaching affordable and their market penetration appears rapidly growing. For this reason the dimension of the fabric layer is growing, but it usually feeds, at building level, proprietary systems, which implement all the remaining stack of the layered architecture, caring of data communication, collection, monitoring, analysis and control. These solutions are usually provided to users which aim at increasing the degree of automation of their home, in this case to improve energy efficiency. Such kind of grid, limited to an apartment or to a building is called smart micro-grid.

The communication layer allows for transmission of data and control signals using heterogeneous technologies and across different kinds of area networks. For example data can be collected locally using embedded systems that are hosted at the user's home, in the Cloud, or directly to the user's smartphone, according to the complexity of applications, the amount of data and the provided functionalities. Hybrid solutions exists [7].

At the next layer data are collected in centralized or distributed repositories. Data flowing from the fabric layer are characterized here by complexity complexity of different type, that make challenging the extraction of relevant information they pro-

vide as a whole. First we must consider different solutions, because data are heterogeneous as they come from very different sources or they are representative of different magnitudes and phenomena (the sources can be rather than utility meter sensors that detect environmental quantities or human phenomena). Sometimes the same technology cannot adequately manage the characteristics of data from different sources, and for example, the integration of the electrical consumption data with those of other energy sources such as fuel or water is seldom supported. But also the data representation make it difficult the correlation between different kinds of energy related information or environmental variables. Furthermore, virtually all the data come from field measurements and need to be consolidated into a single data structure, there is the problem of managing the transport and ensure their integrity. Lies also the problem of data sets to be processed keeping in mind that interoperability with NoSQL database, it is not a very common feature among the energy data management systems, although this, like other technologies that allow you to work with big date, could be a very important resource to support processing in real time, ensuring accurate and timely analysis capabilities and deliver it to decision-makers with the latest information.

The collective layer services orchestrate services provided by the underlying layers to reduce the complexity due to distribution, heterogeneity, computational load, distribution etc. Effort spent to increase innovation could be spent at this layer to integrate available technologies such as in the Cloud field, to meet computational requirements of big data processing.

At highest level we found application which implement monitoring, analysis and sometimes business intelligence to decision makers for an effective management of the infrastructure. Building Management Systems (BMS) are implemented and Energy Information Systems (EIS) are implemented at this layer. BMS are conceived to perform operational tasks (maintenance of electrical and mechanical systems, the occupant complaints management ...), as well as the energy and the operation of the individual subsystems such as the heating or air conditioning. EIS allow for recognition and fault diagnosis, alarm management and energy management. In those three areas, if implemented properly, they are able to provide a very important support by speeding up the procedures of intervention in case of failures, focusing the attention of engineers on the most critical events that are happening in real time, integrating and consolidating data flows from environmental or energy meter and presenting the results of calculations with usable mode.

Available technological solutions implement all or part of the discussed architecture. For example in Fig. 2 we can see some example of implementation for smart micro-grids and smart grid:

- A commercial solution[1] that provides a wifi gateway from a sensor network of smart meters to an app that can be used for monitoring and management of the smart home. The apps implements all the layers of the architecture.
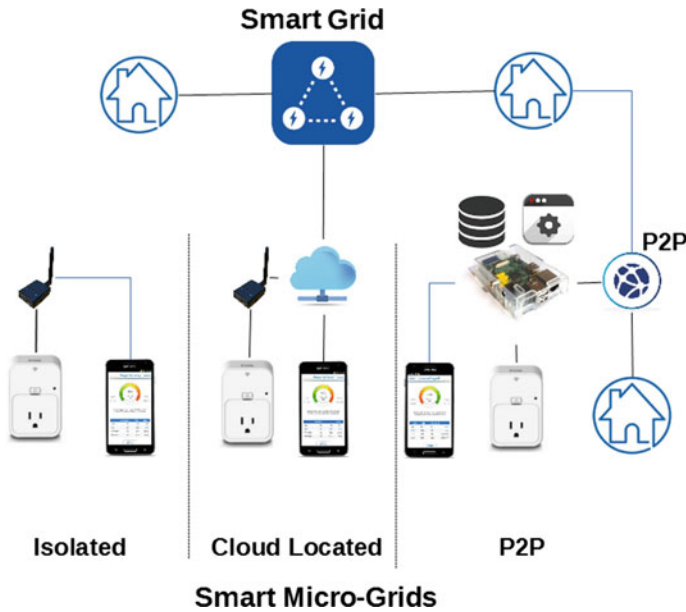
---

[1]http://www.4-noks.com.

Fig. 2 Technological implementations of smart micro-grids

- A different technology[2] that use a resident gateway to forward collected data to the cloud, where they are stored and analyzed, monitoring and control dashboard are provided to customers by application as a service (AaaS). In this case the data, the application and the business logic are hosted in Cloud.
- The CoSSMic [8] research project prototype implements a peer to peer network of collaborating smart micro-grids, each of them collecting and storing data locally. All layers are hosted at home, in an PC or an embedded system with enough computational resources.
- Finally smart grid solutions are provided by some big commercial companies, which usually provide ad hoc solution which are not integrated with the micro-grids, a part of the monitoring and total energy flow between the household and the distribution network.

Figure 3 shows how the variation in energy consumption patterns between buildings in a neighbourhood can be optimized by coordinating load shifting and the use of storage capacities. The investigation of a scalable, at least at neighborhood level, is investigated in [6] by a Peer to Peer (P2P) collaboration of smart micro-grids [5, 7].

Unfortunately fragmentation of the smart grid is observed both across layers and within the same layers, because of different reasons which range from interoperability to security. In the current technological scenario more and more micro-grids are
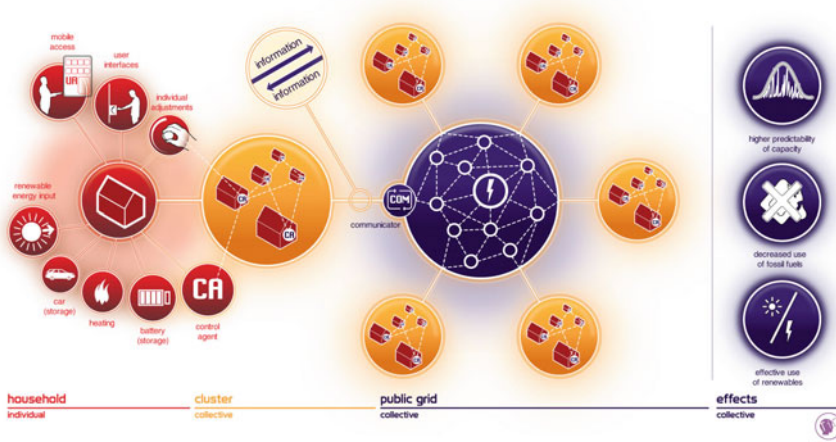

Fig. 3 Collaborating smart grids

growing, which are all connected to the power grid, but isolated from the smart grid. Open issues affect different layer of the presented architectures. Some of them can be addressed independently such as growing supply of services in mobility, flexibility and customization of user interfaces. Other ones rely on interoperability and security, that are drivers to deliver and integrate data to upper levels, for improving the application awareness, for aggregate processing and information fusion.
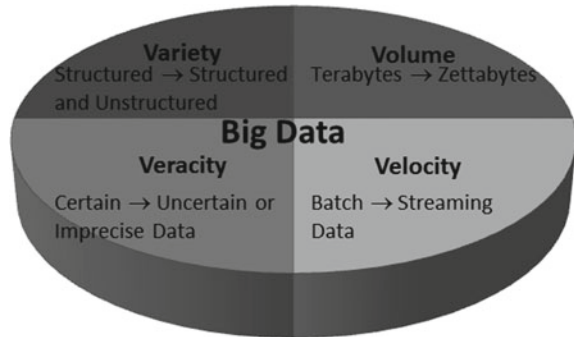
But this process has started and more and more data are collected and must be stored, aggregated and processed. For this reason the smart energy energy is already now a big data problem. In the following sections we will focus on the data and the application layer discussing issues, research efforts and technological solution related to big data properties of smart energy.

3 Big Data Properties of Smart Grid

"Not only will the integration of Big Data technologies help make the grid more efficient, it will fundamentally change who sells electric power; how it gets priced; and how regulator, utilities grid operators and end user interact" is claimed in [9].

An interesting view of what are the Big Data has been exposed to Gartner that defines Big Data as "high volume, velocity and/or variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" [10]. In fact the huge size is not the only property of Big Data. Only if the information has the characteristics of Volume,Velocity and/or Variety we can talk about Big Data [11] as shown in Fig. 4. Volume refers to the fact that we are dealing with ever-growing data expand-

ing beyond terabytes into petabytes, and even exabytes (1 million terabytes). Variety refers to the fact that Big Data is characterized by data that often come from heterogeneous sources such as machines, sensors and unrefined ones, making the management much more complex. Finally, the third characteristic, that is velocity that, according to Gartner [12], "means both how fast data is being produced and how fast the data must be processed to meet demand". In fact in a very short time the data can become obsolete. Dealing effectively with Big Data "requires to perform analytics against the volume and variety of data while it is still in motion, not just after" [11]. IBM [13] proposes the inclusion of veracity as the fourth Big Data attribute to emphasize the importance of addressing and managing the uncertainty of some types of data. Striving for high data quality is an important Big Data requirement and challenge, but even the best data cleansing methods cannot remove the inherent unpredictability of some data, like the weather, the economy, or a customer's actual future buying decisions. The need to acknowledge and plan for uncertainty is a dimension of Big Data that has been introduced as executives seek to better understand the uncertain world around them.

In the case of smart grids most of information are provided by *Machine-generated data* coming from smart meters to measure the consumption of electricity, but we have also human-generated data. Data frequency will varie a lot. In fact the plans will be update by the user at low frequency at not regular intervals. Monitoring must be available on demand, as with social media data. On the other hand prediction and measures could came periodically and at higher frequency. Even if data analysis can be performed in batch mode, on the other hand negotiation and scheduling should take place in near real time. In any case we handle structured data.

About data complexity we have to consider that volume of data increases continuously, either because of the continuous production of metering information and because the foreseen connection of new users and devices. Such issue pose a challenge about using traditional methods, such as relational database engines to store, search, share, analyze, and visualize using. Data processing can make use of massive parallel processing power on available hardware because of an application model based on collaborating autonomous agents, however location of data and privacy is an open issue in current big data solutions.

Data velocity can obviously increase and response adaptation may be required.

The quality of the data made available by an energy information system is certainly another area of research on which it is important to reflect. Quality can be understood from the point of view of fairness and integrity of the data source to user, but also from the perspective of adherence data to the needs of decision-makers and their proper context. The first part of the problem may seem simple solution, but taking into account the heterogeneity of the sources and the absence of specific reference models, combined with the specific nature of the data collected in each real application context, you realize that it can be a longer process complicated than expected. It must be added that not always the transport of energy data takes place in an ideal way, since the communication networks built for this purpose are, in many cases the result of compromises due to the need to make them coexist with the existing classical structures of the buildings which are the subject Business management Energy. Standardize the data in a consistent and functional structure to the needs of the organization that will exploit them, it is also a prerequisite for people and structures who base their decisions on these data. The second part of the reasoning on data quality has instead to do with both the design that with the implementation of these systems. The fact that the data provided by a Smart Grid are more or less appropriate in decision-making that are intended to support, depends both on the instrument's ability to offer the depth and breadth of analysis, that the skills of those who are planning to deploy to a specific application context. By focusing on the power of the analysis instruments, the question regards the minimum level of granularity provided, and it is evident that this can only be a trade-off between the complexity of the system and the large number of listening points (the meters and sensors) it has. It is inevitably an issue that involves aspects concerning the scalability of the systems, the complexity of the necessary data structures and performance processing; identifying the right balance between these three dimensions is surely one of the most significant issues to be addressed, and will be both a mobile equilibrium influenced by the evolution of the technologies that make up the energy management systems.

Big data is changing the way of energy production and the pattern of energy consumption. Energy big data have brought opportunities and challenges at the same time for us. Some of the primary and urgent challenges include: (a) how to effectively collect, store and manage the energy big data; (b) how to efficiently analyze and mine the energy big data; (c) how to use the energy big data to support more effective and efficient decision makings; (d) how to get insights and obtain values from the energy big data; and (e) how to effectively prevent risks and protect privacy while utilizing the energy big data [14].

So, the challenge is to find a way to transform raw data into valuable information. To capture value from Big Data, it is necessary an innovation in technologies and techniques that will help individuals and organizations to integrate, analyze, visualize different types of data at different spatial and temporal scales. Based on Big Data Characteristics, [14] describes the characteristics of energy big data and proposes a mapping of the two. So the 4 V characteristics became the 3E (energy, exchange and empathy) characteristics of energy big data. Energy (data-as-an-energy) means that energy savings can be achieved by big data analytics. Energy big data with its easy to

transport properties, and in the course of constantly refining and value-added. Under the premise of can protect the interests of users, in each link of the low energy. By saving energy, the process of energy big data is the process of water and electricity energy release in a sense, through the energy big data analysis to achieve the purpose of energy saving, is the largest investment in energy infrastructure [15].

Exchange (data-as-an-exchange) refers to that the big data in energy system need to exchange and integrate with the big data from other sources to better realize its value. Energy big data has a value that reflect the entire national economy, social progress and development of all walk of life and other aspects of innovation and to play a greater value of its premise and the key is to interact with external data. In fact big data energy fusion and a full range of mining, analysis and presentation on this basis, can effectively improve the current industry [15].

Empathy (data-as-an-empathy) means that better energy services can be provided, users needs can be better satisfied, and consumer satisfaction can be improved based on energy big data analytics. Enterprise's fundamental purpose is to create customers, create demand. Energy big data natural contact households, factories and enterprises to promote power industry by the electric power production as the center to take the customer as the center which is the essence of ultimate concern for power users. Through the power user needs to fully tap and met to establish an emotional connection, for the majority of electricity users to provide more high-quality, safe and reliable electric service. In the process of energy industry contribution to maximize the value of energy industry also found a new source of power often changed frequently, empathy can benefit [15].

In conclusion, big data have a large impact on the management of power utilities as smart grid operation and future energy management will be hugely data-intensive. There are many challenges which affect the success of big data applications in smart grid. It is necessary to gain practical experience in integrating big data with smart grid together with more effort to develop more advanced and efficient algorithms for data analysis. With this aim, in the next section, we introduce several research contribution in the field of big data analytics for smart grid.

4 Research Lines and Contribution

In this section we present an overview of research contribution for big data analytics applications to the smart grid. In [16] authors identify several future research directions about smart grid, we focus here on those which are related to big data.

An effort to improve interoperability needs to aggregate all available information from different sources, such as individual smart meters, energy consumption schedulers, solar radiation sensors, wind-speed meters and relays.

As it is strengthened in [17], it is relevant to improve data collection and management, starting from optimizing data sampling, choosing effective storage solutions.

The next open issue deals with real-time monitoring and forecasting, which requires effort in the application and development of data mining techniques and

big data analytic. Techniques for categorization of the information and successful recognition of the different patterns are necessary here and at the business level.

Big data analytic is also relevant to provide at the right communication point necessary decision making support.

Research Focus	Requirements	Technology
Interoperability and standardization	Metrics, taxonomies, protocols	Ontologies, semantic
Big Data storages and Cloud availability	Performance, reliability, aggregation, processing, NOSQL	
Big data analytic	Categorization, pattern recognition, decision making, monitoring and forecasting	Data mining, business intelligence

4.1 Interoperability and Standardization

Smart Grid data are collected from different sources, such as individual smart meters, energy consumption schedulers, aggregators, solar radiation sensors, wind-speed meters. Some effort have been spent by widespread deployment of wireless communication power meters, availability of customer energy usage data, development of remote sensing for determining real-time transmission and distribution status, and protocols for electric vehicle charging.

National Institute of Standards and Technology (NIST) ongoing efforts aim at facilitating and coordinating smart grid interoperability standards development and smart grid-related measurement science and technology, including the evolving and continuing NIST relationship with the Smart Grid Interoperability Panel (SGIP[3]) public-private partnership. NIST developed an initial (now completed) three-phase plan whose first step is the identification and consensus on smart grid standards. In particular interoperability on data is addressed at technical level, focusing on syntactic interoperability, and at information layer, where it deals with semantic understanding business context. Syntactic interoperability means understanding of data structures, while semantic means understanding of concept contained in data structures. At the state of the art improving interoperability requires development and application of techniques and technologies to achieve such alignment on big data which already now are available in smart grids.

Table Definition Language (TDL), described in terms of the XML-based Exchange Data Language (EDL) that can be used to constrain oft-utilized information into a well-known form, are proposed for data representation.

---

[3]http://www.sgip.org/.

Many standard and proposal are already available.[4]

They are related to information exchange among many components of a smart-grids such as energy usage in kilowatt hours from a meter, load profiles and control information, communications protocols for field devices and systems in SCADA Distribution Grid Management, Markets Energy Price and Transactions, Schedules and Calendars, Standard for Common Format for Event Data Exchange (COMFEDE) for Power Systems.

4.2 Big Data Storages

Widespread integration of grid-tied renewable along with attendant large-scale storage has been recognized as a key scientific and technological areas related to smart-grid [18].

Research effort at this layer deals with the effective utilization and integration of available technologies to address different kind of data of smart grid.

Big Data are so complex and large that it is really difficult and sometime impossible, to process and analyze them using traditional approaches. In fact traditional relational database management systems (RDBMS) can not handle Big Data sets in a cost effective and timely manner. These technologies may not be enabled to extract,from large data set, rich information that can be exploited across of a broad range of topics such as market segmentation, user behavior profiling, trend prediction, events detection, etc. and in many fields like public health, economic development and economic forecasting. Besides Big Data have a low information per byte, and, therefore, given the vast amount of data, the potential for great insight is quite high only if it is possible analyze the whole dataset [11].

The term NoSQL (meaning 'not only SQL') is used to describe a large class of databases which do not have properties of traditional relational databases and which are generally not queried with SQL (structured query language). NoSQL data stores are designed to scale well horizontally and run on commodity hardware. Also, the 'one size fit's it all' [19] notion does not work for all scenarios and it is a better to build systems based on the nature of the application and its work/data load [20]. They can provide solution both at data and collective layer of a smart grid architecture, in fact some functionalities can be:

- Sharding, also referred to as horizontal scaling or horizontal partitioning. It is a partitioning mechanism in which records are stored on different servers according to some keys. Data is partitioned in such a way that records, that are typically accessed/updated together, reside on the same node. Data shards may also be replicated for reasons of reliability and load-balancing.
- Consistent hashing [21] The idea behind consistent hashing is to use the same hash function, used to generate fixed-length output data that acts as a shortened refer-

---

[4]http://collaborate.nist.gov/twiki-sggrid/bin/view/SmartGrid/SGIPCoSStandardsInformation Library.

ence to the original data, for both the object hashing and the node hashing. This is advantageous to both objects and machines. The machines will get an interval of the hash function range and the neighboring machines can take over portions of the interval of their adjacent nodes if they leave and can assign parts of their interval if some new member node joins and gets mapped to a nearby interval. Another advantage of consistent hashing is that clients can easily determine the nodes to be contacted to perform read or write operations.

- Advanced computation capabilities such as MapReduce [22] for processing large data sets. A popular open source implementation is Apache Hadoop [23], a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage and to execute queries and other batch read operations against massive datasets that can be tens or hundreds of terabytes and even petabytes in size. It supports both high performant stream and batch processing, which are respectively used for smart grids to detect in real time alerts, error or for business analytics.

On the other hand, to chose the right technological solution we have to take into account the type of data of the smart energy domain and the kind of requirements for processing them. In particular format of the content for the energy measure is of course composed of real or integer values. Hence we have to store, communicate and process time series which should be available for off-line analysis and reporting but also for on-line monitoring and learning. Moreover we need to handle transaction for negotiation and historical data and prediction.

Key-value data stores (KVS) typically provide replication, versioning, locking, transactions, sorting, and/or other features. The client API offers simple operations including puts, gets, deletes, and key lockups. Notable examples include: Amazon DynamoDB [24], Project Voldemort [25], Memcached [26], Redis [27] and RIAK [28]. They can be exploited to organize information. Many of them offer in-memory solution and workload distribution [26], which can be used to improve performance for stream processing, when it needs to elaborate on line high frequency measures which are received from smart meters.

Document data stores (DDS). DDS typically store more complex data than KVS, allowing for nested values and dynamic attribute definitions at runtime. Unlike KVS, DDS generally support secondary indexes and multiple types of documents (objects) per database, as well as nested documents or lists. Notable examples include Amazon SimpleDB [29], CouchDB [30], Membase/Couchbase [31], MongoDB [32] and RavenDB [33]. They could be exploited at business level to store report, result of analysis, billing documents.

Extensible record data stores (ERDS). ERDS store extensible records, where default attributes (and their families) can be defined in a schema, but new attributes can be added per record. ERDS can partition extensible records both horizontally (per-row) or vertically (per-column) across a datastore, as well as simultaneously using both partitioning approaches. Notable examples include Google BigTable [34],

HBase [35], Hypertable [36] and Cassandra [37]. In smart grid they are the solution for building datawarehouse and supporting business analytics.

Another important category is constituted by Graph data stores. They [38] are based on graph theory and use graph structures with nodes, edges, and properties to represent and store data. Different Graph Store products exist in the market today. Some provide custom API's and Query Languages and many support the W3C's RDF standard. Notable examples include neo4j [39], AllegroGraph [40] and Infinite-Graph [41]. The graph data model fits better to model domain problems that can be represented by graph as ontologies, relationship, maps etc. Particular query languages allow querying the data bases by using classical graph operators as neighbor, path, distance etc. Unlike the NoSQL systems we presented, these systems generally provide ACID transactions. In smart grid they can be exploited to record relationships among events, users, devices and in general to represent the knowledge of the smart energy domain. They allow to apply at application level effective inference models to recognize situation, deduct reactions and predict future events.

We cannot forget a Time Series Databases which are optimized for handling time series data. Meters output are mainly represented as an array of numbers indexed by time. They are used also to represent profiles, curves, or traces in the smart energy domain. Some available technologies are Graphite,[5] InfluxDB[6] and OpenTSDB.[7] Some of them are not really conceived to store big data, but to effectively manage time series in window time-frames.

Table 1 provides a comparison of all the examples given in terms of Classification, Licence and Storage System. Comparison based on several issues are available at [42].

Exploitation of big data storages in Cloud is another scientific and technological challenge.

In [43] authors discuss about about how Cloud computing model can be used for developing Smart Grid solutions. They propose to exploit to use advantages of Cloud computing to achieve the most important future goals of a large-scale Smart Grid, such as energy savings, two-way communication, and demand resource management. In [44] smart grid data management is based on specific characteristics of cloud computing, such as distributed data management for real-time data gathering, parallel processing for real-time information retrieval, and ubiquitous access.

4.3   Big Data Analytic

In order to build an accurate real-time monitoring and forecasting system, it is necessary to integrate all available information from different sources, such as individual smart meters, energy consumption schedulers, aggregators, solar radiation

---

[5]http://graphite.net.

[6]https://influxdata.com/.

[7]opentsdb.net/.

Table 1 Data store comparison

Name	Classification	License	Data storage
Dynamo	KVS	Proprietary	Plug-in
Voldemort	KVS	Open Source	RAM
Memcached	KVS	Open Source	RAM
Redis	KVS	Open Source	RAM
RIAK	KVS	Open Source	Plug-in
SimpleDB	DDS	Proprietary	S3 (Simple Storage Service)
CouchDB	DDS	Open Source	Disk
Couchbase	DDS	Open Source	Disk
MongoDB	DDS	Open Source	Disk
RavenDB	DDS	Open Source	Disk
Google BigTable	ERDS	Proprietary	GFS
HBase	ERDS	Open Source	Hadoop
Hypertable	ERDS	Open Source	Disk
Cassandra	ERDS	Open Source	Disk
Neo4J	Graph	Open Source	Disk
AllegroGraph	Graph	Proprietary	Disk
InfiniteGraph	Graph	Proprietary	Disk
RRDtool	TSDB	Open Source	Disk (Circular Buffer)
Graphite	TSDB	Open Source	Disk
Druid	TSDB	Open Source	NFS
Riak TS	TSDB	Open Source	Hadoop
OpenTSDB	TSDB	Open Source	Hadoop, Hbase
kdb+	TSDB	Commercial	Disk, column oriented
splunk	TSDB	Commercial	Oracle DB

sensors [45]. Moreover [45] also individuate two important issues. the appropriate forecasting system should rely on effective data sampling, improved categorization of the information and successful recognition of the different patterns. Second, suitable adaptive algorithms and profiles for effective dynamic, autonomous, distributed, self-organized and fast multi-node decision-making have to be designed. This requires to invest effort in big data analytics, as it will give utilities and grid operators insights on smart grid data for grid management, reliability and efficiency. Analytics application can utilize various big data options for administering smart meter gateways and meter data processing for market participants. Energy theft and overloaded distribution equipment detections through grid load analysis, grid incident analysis and end-customer consumption load analysis are examples of information that can be extracted. The big data option will also allow the creation of load forecasts for different levels in the distribution grid as well as an analysis of distributed energy resources.

Descriptive smart meter analytics have already proved to be quite valuable for utilities that are searching for ways to use data to understand root causes and provide awareness for better decision making.

Diagnostic analytic models is closely coupled with descriptive ones a diagnostic model will further analyze the data to look for trends and patterns in the monitoring stage. Thus, the successful model will use drill-downs, factor analytics, and advanced statistical exploration.

Predictive big data analytics is a powerful tool to expose risks, uncover opportunities, and reveal relationships among myriad variables.

Relevant use case are the need for load balancing according to the day-to-day and hour-to-hour costs of power. The goal is saving both money and energy by predicting the costs of power and demand based on a constant flow of signals, allowing the distributors to buy and sell accordingly while shaving load during peak hours. Predictive analytics, applied to big-data, allows to create an interaction layer between the bulk power system and the distribution systems.

Decision making support can be implemented, exploiting Big Data Analytics Strategies for producing recommendations to find the best decision in a particular context making a more informed guess about the most high-value action.

Research contribution have proposed how to use computing resources for applying such techniques. A hierarchical structure of cloud computing centers to provide different types of computing services for information management and big data analysis is described in [46]. Security issues are addressed by a solution based on identity-based encryption, signature and proxy re-encryption.

In [47] it has been shown that the performance of multi-node load forecasting is clearly better than that of single-node forecasting. [47] proposes a load data hierarchical and partitioned processing method, establishes a formula to reflect their mutual restraint and relation, creates a model to describe transmission system multi-node load dynamic characteristic on the basis of top layer forecasting using recursive least square support vector machines algorithm, and constructs an ultra-short term load forecasting overall frame of adaptive dynamic model. As shown in [48], the designed algorithms should be based on realistic consensus functions or voting by incorporating probability terms models, where the large computations can be parallelized. The algorithmic results are the state estimation, the estimated production and consumption, and the STLF in SGs.

For the most efficient pattern-recognition and state estimation in the SGs environment, the following methodologies and technologies can be used:

- *Feature Selection and Extraction*. The traditional factors include the weather conditions, time of the day, season of the year, and random events and disturbances. On the other hand, the smart grid factors include the electricity prices, demand response, distributed energy sources, storage cells and electric vehicles.
- *Online Learning*. In contrast to statistical machine learning, online learning algorithms do not make stochastic assumptions about the observed data. Some application use a streaming processing of incoming data o recognize already known shape-lets and classify events.

- *Randomized Model Averaging*. It is concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data. A major focus of research is to automatically learn to recognize complex patterns such as the features of smart grids, and make intelligent decisions based on data.

Even if many contributions focused on short term load forecasting, using regression models, linear time-series-based, state-space models, and nonlinear time-series modeling, on the other hand they have been seldom used at meter aggregate levels, such as distribution feeders and substation. Moreover, very little progress has been made in the field of the very-short-term load forecasting, which could be very useful to compensate aggregate power fluctuation of co-located photovoltaic panels, when storages are full or do not exist.

4.4   Research Projects Networked with Companies

Power utilities are cooperating with IT companies to develop big data analytics for smart grid applications. There are several research project, networked with big companies, aimed at investigate big data analytics applications to the smart grid. The aim of those research project networked with companies is to achieve also competence and results that can be used as a springboard for the companies' own research and development projects.

Global technology company Siemens announced on February 2016 the integration of a big data option in its smart grid application. They are providing more and more standard Business Intelligence (BI) reports to their products and supporting applications into the smart grid solution. Also European Commission is funding several European Project based on Big Data driven Smart Grid solution. Some important project aiming at solve the issue related to big data analytics, data transformation and management, demand response, settlement end forecasting are the following.

The EU-funded SPARKS[8] (Smart Grid Protection Against Cyber Attacks) project aims to provide innovative solutions in a number of ways, including approaches to risk assessment and reference architectures for secure smart grids. The project will make recommendations regarding the future direction of smart grid security standards. Furthermore, key smart grid technologies will be investigated, such as the use of big data for security analytics in smart grids, and novel hardware-supported approaches for smart meter (gateway) authentication. All of these contributions and technologies will be assessed from a societal and economic impact perspective, and evaluated in real-world demonstrators.

RealValue[9] is an European energy storage project funded by Horizon 2020, the largest Research and Innovation Programme in Europe. RealValue will use a combination of physical demonstrations in Ireland, Germany and Latvia along with innovative modelling techniques, in order to demonstrate how local small-scale energy

---

[8]https://project-sparks.eu/.

[9]http://www.realvalueproject.com/.

storage, optimised across the whole EU energy system, with advanced ICT, could bring benefits to all market participants. Using Big Data technologies, the smart system will also interpret data to forecast heat demand and consumer comfort preferences, and it will interact with the grid to manage supply and congestion.

IES (Integrating the Energy System)[10] is an European Project that combines Big Data technologies and cyber security in order design a modular process chain to achieve interoperability of standards, a specification of a normalised use of these standards in interoperability profiles and a demonstration of the processes for testing interoperability. The description of the processes and the practical implementation is to be vendor-neutral, in order to ensure long-term interoperability and acceptance in energy domain. The transparency of the method and the open database for technical specifications and profiles should be accessible for technology providers for interoperable products and services. In this way interoperability will create increasing competition with decreasing prices for better products on binding security level.

Secure Operation of Sustainable Power Systems (SOSPO)[11] is a project whose main goal is to carry out research and development for the purpose of methods for a real-time assessment of system stability and security, as well as methods for intelligent wide-area prosumption control that can ensure stable and secure operation of the future power system. The research in the SOSPO project focuses on methods that enable system stability and security assessment in real-time and on methods for automatically determining control actions that regain system security when an insecure operation has been detected.

5   Conclusion

In this chapter we have presented a survey of the various research issues, challenges, technical and technological solutions and analyzed critically the utilization of Big Data to manage smart grid introducing several research problems and describing several research project trying to bring out the weaknesses and strengths of the different solutions. The main goal of the research was to identify the critical issues and highlight strengths and potential on the basis of direct experience accumulated. The main challenges are [49]:

- the difficulty in collecting the data by itself due the existence of multiple sources with different formats and types and different usage and access policies;
- the difficulty in categorize and organize and an easily accessible way for applications to use the data itself due to the unstructured nature of the data;
- the difficulty in create a unified understanding of data semantics and a knowledge base in order to extract new knowledge based on specific real-time data;
- the difficulty in retrieve and transform the data automatically and universally into a unified data source for useful analysis;

---

[10]http://www.offis.de/en/offis_in_portrait/structure/structure/projekte/ies-austria.html.

[11]http://www.sospo.dk/.

- the data uncertainty and trustworthiness;
- security and privacy issues, as databases may include confidential information, so it is necessary to protect this data against unauthorized use and malicious attacks;
- the size of generated data that rapidly grows according to the population grows. So Smart Grid applications need to evolve quickly and extend efficiently to handle the growing volume and variety of big data.

Nevertheless, the success of the new energy transition relies on the ability to adopt Big Data analysis using computational intelligence techniques. Computational Intelligence provides solutions for such complex real-world problems to which traditional modeling can be useless for a few reasons: the processes might be too complex and it contains some uncertainties during the process. Computational Intelligence represents a powerful and versatile methodology for a wide range of data analysis problems, so to effectively and efficiently overcome the challenges it is necessary to pay more attention to these techniques.

References

1. NIST: Nist smart grid. http://www.nist.gov/smartgrid/ (2012)
2. Jin, X., Wah, B., Cheng, X., Wang, Y.: Significance and challenges of big data research. Big Data Research **2** (2015) 59–64
3. Diamantoulakis, P., Kapinas, V., Karagiannidis, G.: Big data analytics for dynamic energy management in smart grids. Big Data Research **2** (2015) 94–101
4. Pakkanen, P., Pakkala, D.: Reference architecture and classification of technologies, products and services for big data systems. Big Data Research **2** (2015) 166–186
5. Amato, A., Aversa, R., Di Martino, B., Scialdone, M., Venticinque, S., Hallsteinsen, S., Horn, G.: Software agents for collaborating smart solar-powered micro-grids. Volume 7. (2015) 125–133
6. Amato, A., Di Martino, B., Scialdone, M., Venticinque, S.: Design and evaluation of p2p overlays for energy negotiation in smart micro-grid. Computer Standards and Interfaces **44** (2016) 159–168
7. Amato, A., Di Martino, B., Scialdone, M., Venticinque, S., Hallsteinsen, S., Jiang, S.: A distributed system for smart energy negotiation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **8729** (2014) 422–434
8. Horn, G., Venticinque, S., Amato, A.: Inferring appliance load profiles from measurements. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9258** (2015) 118–130
9. Stimmel, C.L.: Big Data Analytics Strategies for the Smart Grid. CRC Press (2014)
10. Gartner: Hype cycle for big data, 2012. Technical report (2012)
11. IBM, Zikopoulos, P., Eaton, C.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. 1st edn. McGraw-Hill Osborne Media (2011)
12. Gartner: Pattern-based strategy: Getting value from big data. Technical report (2011)
13. Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., Tufano, P.: Analytics: The real-world use of big data. Ibm institute for business value - executive report, IBM Institute for Business Value (2012)
14. Zhou, K., Fu, C., Yang, S.: Big data driven smart energy management: From big data to big insights. Renewable and Sustainable Energy Reviews **56** (2016) 215–225

15. Wan, X., Wang, B.: Key technology research based on big data era hydroelectric energy. In: Advances in Energy Equipment Science and Engineering. CRC Press (2015) 81–85
16. Diamantoulakis, P.D., Kapinas, V.M., Karagiannidis, G.K.: Big data analytics for dynamic energy management in smart grids. Big Data Research **2** (2015) 94–101
17. Yang, P., Yoo, P., Fernando, J., Zhou, B., Zhang, Z., Zomaya, A.: Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications. IEEE Transactions on Cybernetics **44** (2014) 445–455
18. Amato, A., Venticinque, S.: Big data management systems for the exploitation of pervasive environments. Studies in Computational Intelligence **546** (2014) 67–89
19. Stonebraker, M., Cetintemel, U.: "one size fits all": An idea whose time has come and gone. In: Proceedings of the 21st International Conference on Data Engineering. ICDE '05, Washington, DC, USA, IEEE Computer Society (2005) 2–11
20. Gajendran, S.K.: A survey on nosql databases. Technical report (2012)
21. Karger, D., Lehman, E., Leighton, T., Panigrahy, R., Levine, M., Lewin, D.: Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the world wide web. In: Proceedings of the twenty-ninth annual ACM symposium on Theory of computing. STOC '97, New York, NY, USA, ACM (1997) 654–663
22. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51** (2008) 107–113
23. Apache: Hadoop (2012) http://hadoop.apache.org/, [Online; 10-July-2016].
24. DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., Vogels, W.: Dynamo: amazon's highly available key-value store. SIGOPS Oper. Syst. Rev. **41** (2007) 205–220
25. Sumbaly, R., Kreps, J., Gao, L., Feinberg, A., Soman, C., Shah, S.: Serving large-scale batch computed data with project voldemort. (2009)
26. Memcached: Memcached (2012) http://memcached.org/, [Online; 7-July-2016].
27. Redis: (2012) http://redis.io/documentation, [Online; 10-July-2016].
28. Riak: (2012) http://basho.com/riak/ [Online; 6-July-2016].
29. Amazon: Simpledb (2012) http://aws.amazon.com/simpledb/, [Online; 6-July-2016].
30. Apache: Couchdb (2012) http://couchdb.apache.org/, [Online; 6-July-2016].
31. Couchbase: (2012) http://www.couchbase.com/, [Online; 6-July-2016].
32. MongoDB: (2012) http://www.mongodb.org/, [Online; 6-July-2016].
33. RavenDB: (2012) http://ravendb.net/, [Online; 6-July-2016].
34. Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, T., Fikes, A., Gruber, R.E.: Bigtable: A distributed storage system for structured data. ACM Trans. Comput. Syst. **26** (2008) 4:1–4:26
35. HBase: Hbase (2012) [Online; 6-July-2016].
36. Hypertable: (2012) http://hypertable.com/documentation/, [Online; 6-July-2016].
37. : Cassandra (2012) http://cassandra.apache.org/, [Online; 6-July-2016].
38. Robinson, I., Webber, J., Eifrem, E.: Graph Databases. O'Reilly Media, Incorporated (2013)
39. Neo Technology, I.: Neo4j, the world's leading graph database. (2012) http://www.neo4j.org/, [Online; 7-July-2016].
40. AllegroGraph: (2012) http://www.franz.com/agraph/allegrograph/, [Online; 6-July-2016].
41. InfiniteGraph: (2012) http://www.objectivity.com/infinitegraph, [Online; 6-July-2016].
42. findthebest.com: Compare nosql databases (2012) [Online; 6-July-2016].
43. Markovic, D.S., Zivkovic, D., Branovic, I., Popovic, R., Cvetkovic, D.: Smart power grid and cloud computing. Renewable and Sustainable Energy Reviews **24** (2013) 566–577
44. Rusitschka, S., Eger, K., Gerdes, C.: Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain. In: Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on. (2010) 483–488
45. Diamantoulakis, P.D., Kapinas, V.M., Karagiannidis, G.K.: Big data analytics for dynamic energy management in smart grids. CoRR **abs/1504.02424** (2015)
46. Baek, J., Vu, Q.H., Liu, J.K., Huang, X., Xiang, Y.: A secure cloud computing based framework for big data information management of smart grid. IEEE Transactions on Cloud Computing **3** (2015) 233–244

47. Han, L., Han, X., Hua, J., Geng, Y.: A hybrid approach of ultra-short term multinode load forecasting. (2007) 1321–1326
48. Hajdu, A., Hajdu, L., Jns, ., Kovacs, L., Tomn, H.: Generalizing the majority voting scheme to spatially constrained voting. IEEE Transactions on Image Processing **22** (2013) 4182–4194
49. Al Nuaimi, E., Al Neyadi, H., Mohamed, N., Al-Jaroodi, J.: Applications of big data to smart cities. Journal of Internet Services and Applications **6** (2015) 1–15