


SURVEY PAPER

Open Access



# Big Data management in smart grid: concepts, requirements and implementation

Houda Daki<sup>\*</sup> , Asmaa El Hannani, Abdelhak Aqqal, Abdelfattah Haidine and Aziz Dahbi

<sup>\*</sup>Correspondence:  
daki.h@ucd.ac.ma  
Laboratory of Information  
Technologies, National  
School of Applied Sciences,  
University of Chouaib  
Doukkali, Route d'Azemmour,  
Nationale No 1, ElHaouzia,  
24002 El Jadida, Morocco

## Abstract

A smart grid is an intelligent electricity grid that optimizes the generation, distribution and consumption of electricity through the introduction of Information and Communication Technologies on the electricity grid. In essence, smart grids bring profound changes in the information systems that drive them: new information flows coming from the electricity grid, new players such as decentralized producers of renewable energies, new uses such as electric vehicles and connected houses and new communicating equipments such as smart meters, sensors and remote control points. All this will cause a deluge of data that the energy companies will have to face. Big Data technologies offers suitable solutions for utilities, but the decision about which Big Data technology to use is critical. In this paper, we provide an overview of data management for smart grids, summarise the added value of Big Data technologies for this kind of data, and discuss the technical requirements, the tools and the main steps to implement Big Data solutions in the smart grid context.

**Keywords:** Smart grid, SCADA, AMI, Demand response, Communication systems, Big Data, Real time processing, Batch processing, Hybrid processing, Customer analytics

## Background

Recently, the electricity consumption has changed in practice and in nature. The electricity uses are evolving: positive energy buildings, electric mobility, variable intensity urban lighting, storage batteries, etc. The electricity production modes are also evolving thanks to the development of renewable energies and the transformation of the energy mix. The electrical system must therefore evolve towards greater reliability, efficiency and flexibility in order to better take into account the development of new uses and to preserve the balance between consumption and production in a changing energy landscape. Smart grids become a real solution to these concerns, by introducing Information and Communication Technologies (ICT) into electricity grids and integrating efficiently the actions of all users (producers and consumers) in order to guarantee a sustainable, safe and cost-effective supply of electricity.

Smart grids ensure efficient connection and exploitation of all means of production, provide automatic and real-time management of the electrical networks, allow better measurement of consumption, optimize the level of reliability and improve the existing services which in turn lead to energy savings and lower costs [1–5]. The implementation of smart grids features leads to a very large increase in the volume of data to be

processed due to the installation of smart meters and various sensors on the network and the development of customer facilities, etc. For example a smart meter could send the consumer energy usage every 15 min, so every million meters can generate 96 million reads per day instead of one meter reading a month in a conventional grid. So, in addition to energy management, smart grids require great data management to be able to deal with high velocity, important storage capacity and advanced data analytics requirements.

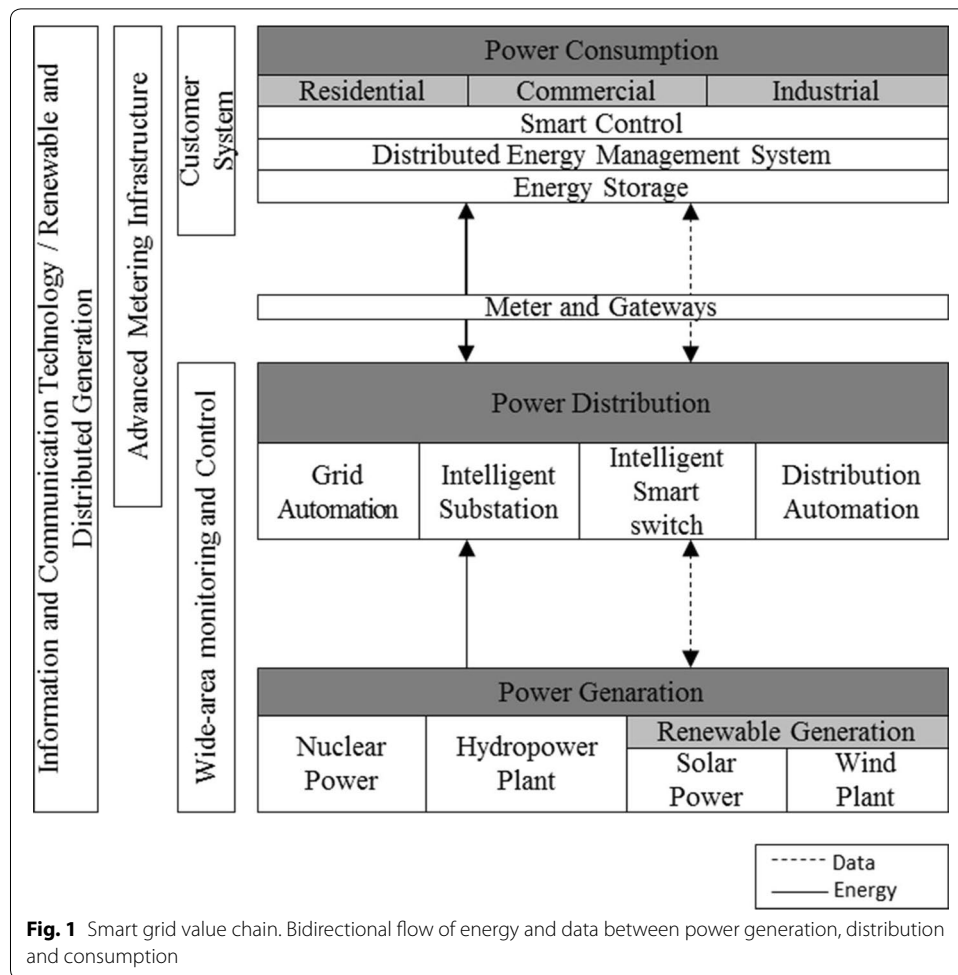
Indeed, smart grids data requires complex treatments, because of their nature, distribution and real-time constraints of certain needs. Big Data techniques are suitable for advanced and efficient data management for this kind of applications. The large volume of data will help utilities do things they never could do before such as better understanding the customer behaviour, conservation, consumption and demand, keeping track of downtime and power failures etc. At the same time, this will present challenges for utilities that lack the systems and data analysis skills to deal with these data. So, the main goal of utilities now is the ability to manage high volume data and to use advanced analytics to transform data collected to information, then to knowledge and finally to actionable plans.

In this context, this paper gives an overview of the opportunities, concepts and challenges of data management in smart grids with the emphasis of Big Data infrastructure. Furthermore, it describes the key criterias and resources requirements utilities should examine in order to select the right Big Data tools and implements given data analytics system. We aim at providing guidance to researchers and companies who have an interest in related issues. The rest of the paper is organized as follows: an overview of smart grid is given in "[Smart grid overview](#)". Description of smart grid systems is provided in "[Smart grid systems](#)". In "[Data management issues in smart grid](#)", we discuss data management issues for smart grid. "[Big Data for Smart Grid](#)" presents Big Data opportunities and infrastructure. Finally, "[Big Data implementation in smart grid: the case of customer data analytics](#)" describes the steps, tools and technical requirements for implementing and deploying big data technologies for smart grids.

## **Smart grid overview**

### **Smart grid architecture**

Smart grid is defined as an intelligent network based on new technologies, sensors and equipments to manage wide energy resources and to enhance the reliability, efficiency and security of the entire energy value chain [1]. The main advantage of smart grids is the ability to better integrate renewable energy sources into the system and supervise energy consumption and production thanks to a bidirectional flow of energy and data between power generation, distribution and consumption as shown in Fig. 1. Power generation is the first step in smart grid value chain, it includes power sources such as nuclear, hydropower and renewable and it relays on wide area monitoring and control technologies to communicate with the next step called power distribution. This later, is based on a proximity network that connects consumers with the electricity grid and transmits data using advanced metering infrastructure. Power consumption is the last step on smart grid value chain and it involves the users of electricity, both residential and industrial. It is increasingly common for the consumer to generate electrical energy

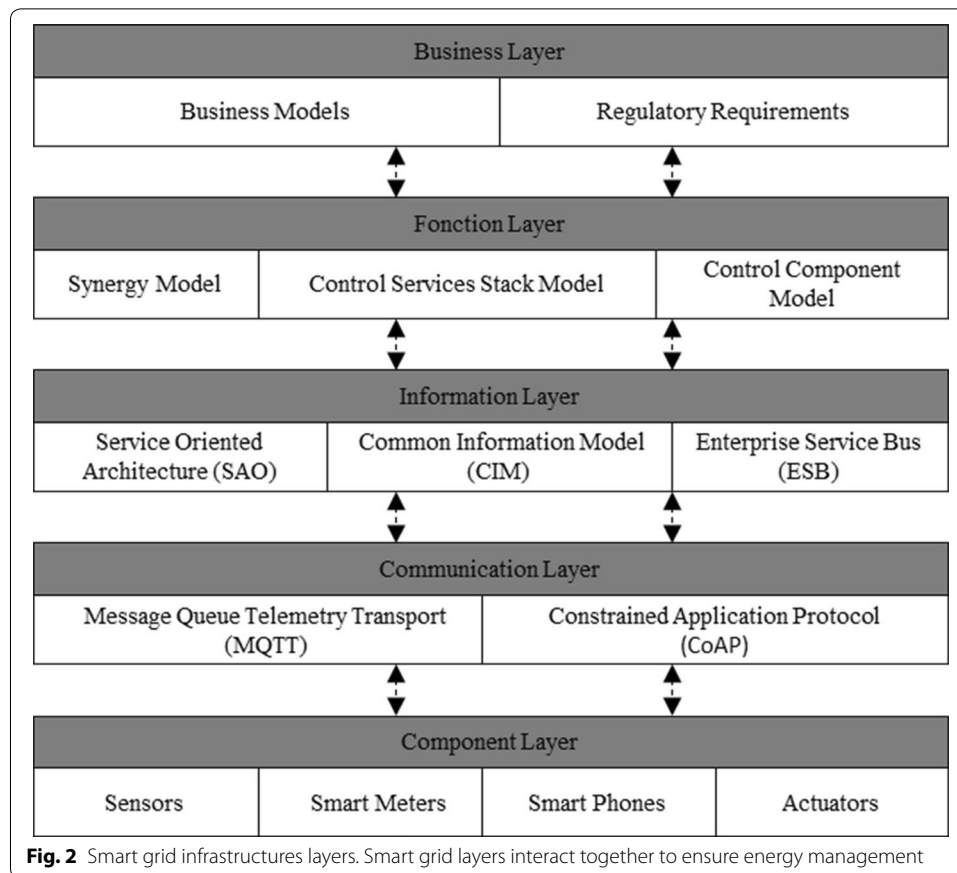


using alternative energy production methods (solar energy, biomass, wind, etc.). For this reason, it is very important to supervise their consumption and production in order to optimize the service.

The smart grid infrastructure is composed of several layers that interact with each other as illustrated in Fig. 2 [2]. The component layer is related to physical devices and is responsible of getting functions, information, and communication means from the other layers. The communication layer uses several techniques and protocols to transfer data between the components of the grid. The information layer describes the data model and the communication systems that will be used to exchange information. The function layer defines the logical functions or applications independently of the physical architecture. Finally, the business layer defines the business models and the regulatory requirements. To ensure energy management and data transfer, these layers communicate to each others and each layer relies on a great number of systems to accomplish its mission.

**Smart grid opportunities**

In the last years, all the world attached high interest in smart grid technologies, because in addition to reliable, secure and efficient electricity management, it enhances power quality which is a main factor in electrical grids. Smart grid concept guarantees an



efficient power quality management, based on intelligent transmission and distribution, policies and pricing mechanisms for real-time power markets [3, 4]. Power quality management relies on smart meters and intelligent energy distribution, which bring information about the power delivered for utilities and consumers.

High or low voltage causes undesirable impacts. It makes the operation of electronic equipments ineffective and may even damage them. The added value of smart grid is to improve the performance and efficiency of the power system by optimizing the voltage, using electronic devices at their highest efficiency and allowing fault tolerance in the electrical grid.

Smart grid brings a great number of added value for both utilities and customers:

- *Added value for utilities* To help utilities better manage the grid and thus make the right decisions at the right time, smart grids use several optimization, control and monitoring systems that allow utilities to have more details about the grid in real time. From the utilities viewpoint, the benefits of smart grids are numerous and can be summarized as follows: (i) Improving the overall management of the production, transport and distribution system, (ii) Enhancing energy independence through the integration of renewable energies, (iii) Optimising the management and modelling of the available capacities of energy production according to the real and/or spontaneous demand, (iv) Maintaining network balance by managing under-voltage and over-

voltage in real time, (v) Improving the security of electricity grids and reducing fraud, and (vi) Improving the quality of services and the customer service.

- *Added value for customers* Smart grids offer many options for customers by using interactive and scalable models of power grid and energy demand. The customers are the users (consumers) of electricity, both residential and industrial. It is more and more frequent that the customer himself produces electrical energy using alternative energy production methods (solar energy, biomass, wind ...). The use of real-time communications with smart grid control and monitoring systems enables the measurement and optimization of the energy value of the customers on the grid. In addition, with the help of smart meters and other equipments of the smart grid, consumers can control their consumption in real time and avoid peak loads through price benefits. They can run their washing machines, dryers and dishwashers at off-peak times, when energy price is very low. As a result, customers not only save money but also require less generation capacity.

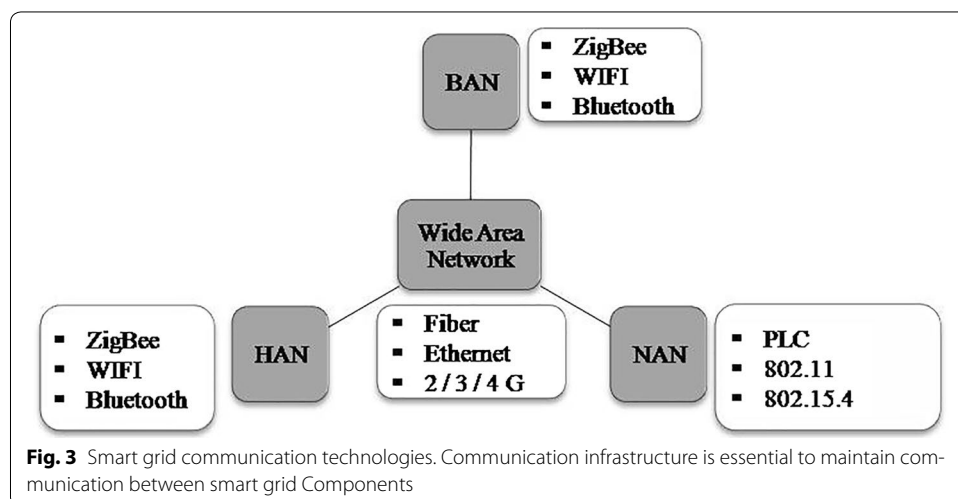
### Smart grid systems

Smart grid relays on advanced and modern communication and information infrastructure to improve energy production, distribution and storage which in turn help reduce the cost and efforts of management and planning.

### Communication systems

Communication infrastructure is very important to maintain communication between the components of smart grids. It allows to have, in addition to the electricity connection, data connection to vehicle information over the grid which is nowadays an essential part of the smart grids. Thus, a secure communication network with high bandwidth capacity and speed is needed. Communication infrastructure manages in general three types of networks: Home Area Networks (HANs), Business Area Networks (BANs) and Neighborhood Area Networks (NANs) [5]. Transmitting data over these networks categories is based on many technologies as shown in Fig. 3.

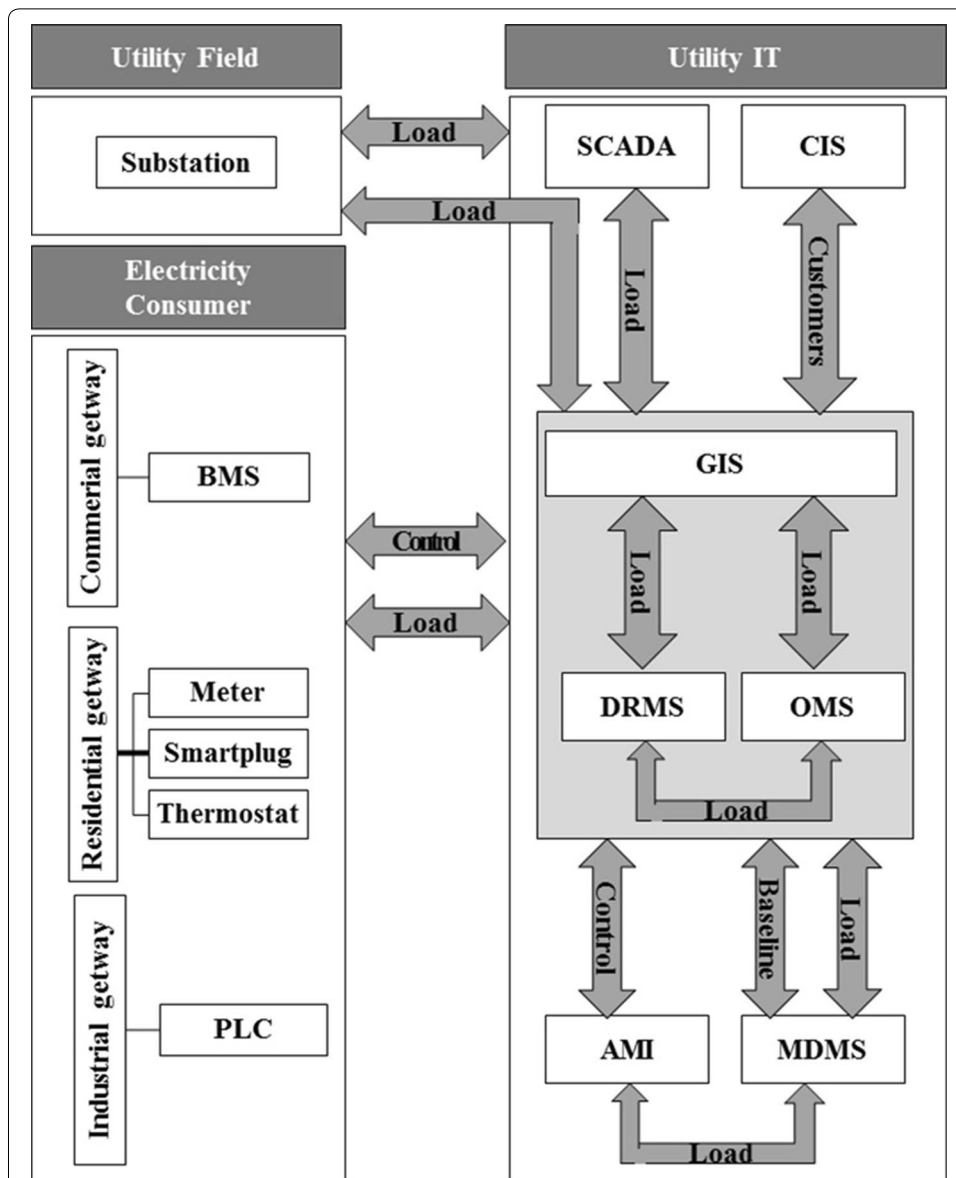
Network technologies in smart grid can be classified into two broad categories:



- *Wireless technologies* Smart grid is composed of a large number of devices of various types and most of them can only communicate using wireless channels [6]. Wireless technologies are facing challenging issues in term of bandwidth, scalability and distance requirements, especially for transmitting large data. As a result, wireless connection is used in multi-service layer to catch data. Technologies such as IEEE802.11, IEEE802.15.4, Bluetooth, Infrared, ZigBee and Radio frequency are applicable for smart grid applications. IEEE802.15.4 and ZigBee are the most successful, because the other air interfaces have some limits to face some nodes specific requirements such as energy consumption, bandwidth demand, throughput and latency. For example Radio frequency have a lack of protocols, a broadcast signal and there are no security system [7] and Bluetooth supports star topology only, operates in few nodes and requires low density. To optimize wireless network capacity, Multiple-Input Multiple-Output (MIMO), Orthogonal Frequency-Division Multiplexing (OFDM) technologies were proposed recently [8].
- *Wired technologies* Wireline networks are more efficient than wireless ones, because they offer higher capacity, optimal communication delay and wide coverage, but on the other hand they require an extra investment for cable deployment. Wireline technologies are based on lot of technologies including fiber optics, IP-based Wavelength Division Multiplexing (WDM) network and SONET/SDH etc [9]. Optical technologies make wireline networks support between 155 Mbps and 160 Gbps [10]. Power Line Communication (PLC) is another kind of wired technology used by electrical companies to transmit data over existing power cables. This technology helps utilities to reduce costs, because it is over traditional electric power grids. PLC presents some challenges such as limited data rates due to attenuation, delay and replication of the phase. Recently, broadband PLC and narrow-band PLC helped utilities to overcome the limited data rate that has reached more than 200 Mbps [11].

### Information systems

Information systems are crucial components in smart grids that communicate together for a flexible, scalable and efficient grid as illustrated in Fig. 4. Utility information systems control and load data coming from substations of the utility field or from electricity consumers including commercial, residential and industrial consumers, then use it to extract values about the state of the lines and equipments, the energy consumed, the consumption modes, etc. Utility information systems contain several components: Supervisory Control and Data Acquisition (SCADA) system collects data from utility field, then uses it to manage the electrical grid infrastructure. This system communicates with other information systems to report about the network. Customer Information System (CIS), Geographic Information System (GIS), Advanced Metering Infrastructure (AMI) and Meter Data Management System (MDMS) process the data coming from electricity consumers, and they also exchange data between each other. Demand Response Management System (DRMS) and Outage Management System (OMS) are the main systems in the grid because they interact with all other systems and even together to guarantee a global vision of the grid and consumers satisfaction. More details about each of these systems are provided below.



**Fig. 4** Information systems in smart grid. Smart grid information systems contain several components to control and load data

**Supervisory control and data acquisition**

Supervisory Control and Data Acquisition (SCADA) is a reliable and safe system for monitoring control in smart grid. This system manages and maintains electrical networks, because it has the ability to collect data from any system in order to control it, using several devices for example sensors, SCADA master and SCADA Remote Terminal Unit (RTU) [5]. The mission of a SCADA system is data acquisition, data communication, data presentation and control in order to increase automation, efficiency and reduce costs. It helps smart grid to deliver energy in optimal way by offering a great number of opportunities such as programmable controls, multi-protocol support and detecting network malfunction using alarms.

### ***Advanced metering infrastructure***

The Advanced Metering Infrastructure (AMI) allows having measurements about energy consumption and production, which help utilities control the energy and be efficient in term of cost and time. AMI relies on smart meters, which must meet several requirements such as the data storage capability, the duration of meter intervals etc. AMI uses a great number of technologies to manage smart meters, such as MDMS, operational gateways and systems for data integration into software application platforms. All these technologies ensure advanced management systems in smart grid [12].

### ***Outage management system***

Utilities always seek to ensure a very high rate of user satisfaction by discovering, locating and resolving power outages in a very efficient and short time. Outage Management Systems (OMS) are important to have a vision and more precision about outages that can happen, in order to take corrective actions, minimize the effect, diagnose the causes and improve the system's availability and reliability. There is a lot of collected data that can be used for operational improvements, which helps to manage outage and improve efficiency including customers data, mobile workforce, field crews, SCADA and smart meters etc.

### ***Geographic information system***

The Geographic Information System (GIS) is primordial for utilities because it helps them to have visualization of maps and points of interests, to manage spatial data and present it. This system can be considered as visualization technology of the grid to have a global vision of consumers, generators and power lines position etc.

### ***Customer information system***

The Customer Information System (CIS) comes in order to develop the relationship between utilities and customers using every customer interaction. CIS helps utilities to deliver their services efficiently, to automate periodic tasks and to understand customers requirements and how each customer is connected to the grid.

### ***Demand response management system***

The Demand Response Management System (DRMS) gives the utilities the ability to create automated, integrated, and flexible platforms to manage demand response solutions in an efficient and smart manner. This system brings a great number of benefits such as reduce energy costs, improve stability and security and ensure satisfaction for customers and regulatory requirements for demand-side.

### **Data management issues in smart grid**

Smart grids systems generate a large quantity of data, for example SCADA system collects data every 2–5 s, AMI system collects data every 1–15 min etc. So, utilities face great number of challenges from strategy to performance in data management.



### **Standards and interoperability**

A smart grid is an heterogeneous and complex environment that contains different kind of devices, networks, systems and data. As examples, there are networks with fast or low processing, devices with or without energy constraints, interactive or non interactive systems, continuous or non continuous data, etc. So, smart grids face different requirements and challenges to manage data integration in term of bandwidth constraints, errors, limited resources and high scalability. If standards are not provided, utilities are dealing with different protocols with different definitions and different communication techniques which would make interoperability impossible. To standardize smart grids, many information models have been developed, starting by IEC 61850 used to communicate with MDMS and related enterprise applications. IEC 61970/61968 Common Information Models (CIM) is another information model which uses IEC 61850 as the basis of information exchanges and messaging. Recently, the integration of smart inverters required advanced protocols, so a more developed information model has been published as IEC 61850-90-7. There is also other advanced protocols like IEEE 1815 (dnp3) and IEEE 2030.5 (Sep2) which allow the use of existing communications infrastructure [2].

### **Management of massive data volume**

Smart grids bring to surface the cost of storing and processing the huge quantity of data used to manage the grid. Unfortunately, utilities still don't make full use of the new data collected because they lack the infrastructure and/or data analysis skills to deal with it. In addition to utilities massive data management issues, there are some challenges for customers data management. The unused data will reduce smart grid opportunities for consumers as active participants, in term of controlling their energy consumption and avoiding peak loads through price benefits. Customers should be aware of the huge quantity of data that can be extracted using smart grid platforms. They should be educated about how to communicate with their meters and with the different platforms in the grid so they can make efficient choices to conserve energy and save money.

### **Security and data privacy**

In the smart grid ecosystem, millions of devices are inter-connected via communication networks which surrenders the grid to potential vulnerabilities. Moreover, virtualization technologies, the key to use cloud computing technologies, makes electrical companies able to run their applications in virtual machines, which reduce the investment cost in terms of hardware and energy [13]. But on the other hand, it has some limitations in security due to the shared platform between several users. The network bandwidth is also another challenge, which causes low latency problems in real time applications, that require highly scalable, available, and fault-tolerant connection [14]. All these ICT dimensions increase the risk of compromising smart grid security objectives, namely: availability, integrity, confidentiality and accountability. The data confidentiality cannot be guaranteed if there is no secure connectivity between devices [15]. To secure the communication, there is a need to adopt an authentication mechanism [16]. The most common is the 'authentication, authorization and accounting' (AAA) mechanism. The

authentication defines users using credentials, the authorization describes for each user his own permissions, and the accounting is responsible of supervising users [17].

**Big Data for smart grid**

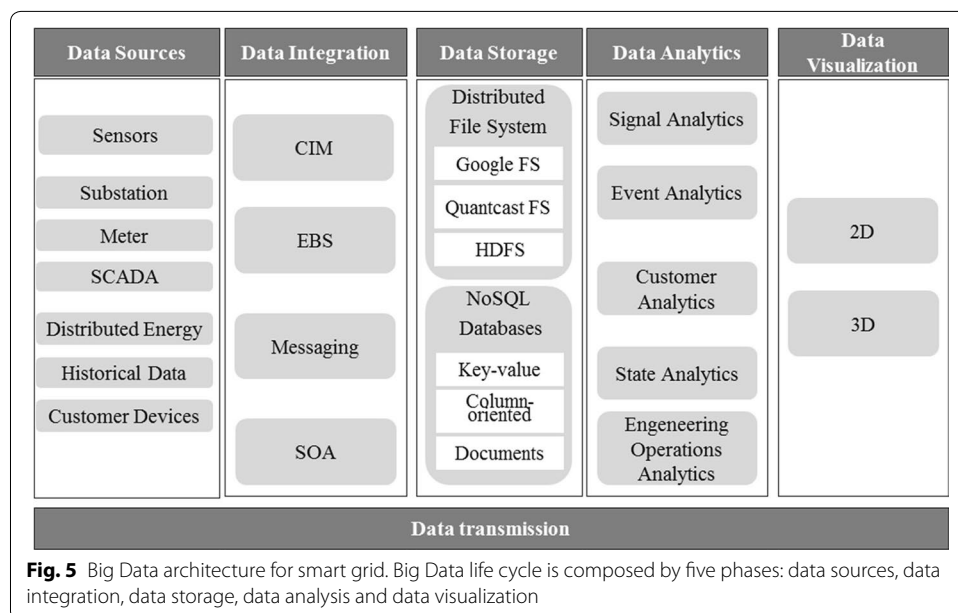
Big Data technologies are a good opportunity for utilities to bring new methodologies, evaluation models and applications and improve data management in smart grids.

**Big Data life cycle**

Big Data can be defined as a huge quantity of datasets, but in fact it includes other features. In addition to (1) the volume, Big Data is based on (2) the variety to present various data formats (structured, semi-structured or unstructured), (3) the velocity to provide timeliness requirements, (4) the value to give the ability to extract the meaning from the collected datasets, (5) the variability to provide inconsistency concept of the data, and (6) veracity to work on the trustworthiness of the data [18]. Figure 5 presents Big Data technologies for smart grid, in it different levels from data sources to visualization.

**Data sources**

Actually, there are distinct data classes according to the type of extracted values: (i) Operational data which is the electrical data of the grid that represent real and reactive power flows, demand response capacity, voltage etc. (ii) Non-operational data is not related to grid power but it refers to master data, data on power quality and reliability etc. (iii) Meter usage data is another kind of data associated to power usage and demand values such as average, peak and time of the day etc. (iv) Event message data comes from smart grid devices events like voltage loss/restoration, fault detection event etc. Finally, (v) Metadata, which is used to organize and interpret all the other kind of data. All these data are collected from several sources such as meters, sensors, devices, substations,



mobile data terminals, control devices, intelligent electronic devices, distributed energy resources, customer devices and historical data.

### **Data integration**

Modern information and communication technologies and advanced operation are used actually to improve smart grid reliability, persistence, efficiency and performance. That's the reason, to have several technologies and approaches to ensure data integration:

- *Service Oriented Architecture (SOA)* all enterprise systems combine a great number of software, each one has its own way to provide services to users. So the problem is how to manage and maintain all these systems. As a solution, SOA makes software communicate together using a single approach which makes data integration easier and more flexible [19]. In smart grids, SOA is used essentially on demand systems.
- *Enterprise Service Bus (ESB)* is based on a great number of approaches to manage communication between different kinds of systems such as GIS, OMS, CIS etc. ESB brings a lot of benefits to reduce cost and time in term of management, monitoring and divergence of integration [20]. In smart grid, ESB technologies are strongly related to SOA, since it makes it more robust and flexible.
- *Common Information Models (CIM)* are used for smart grid persistence and for the integrated data architecture and are critical, especially in the success or failure of data management. CIM refers to UML models for the electric power industry. It plays a very important role in energy management systems in term of data integration, time and cost. In general, CIM help to exchange data with technical grid infrastructure. The CIM become primordial in power systems in order to guarantee the data interoperability, in the case of implementing different applications. CIM operate in data transformation level, it is used with ESB for the normalization and standardization of the data between smart grid systems.
- *Messaging* represents communication systems based on exchanging messages. These messages include data and other information from different applications managed by messaging server [21].

### **Data storage**

Data storage in smart grid has a critical role, because it is based on collecting data from dispatched sources and delivering data to analytics tools in fast input/output operations per second (IOPS). So there is a need for a developed and scalable data storage mechanism to meet Big Data requirements.

- *Distributed File System (DFS)* is a file system that allows multiple users on multiple machines to share files and storage resources. It is based on client/server as storage mechanism, and it permits every user to get a local copy of the stored data. There is a great number of solutions that use DFS for example: Googles GFS, Quantcast File System, HDFS, Ceph, Lustre GlusterFS, PVFS etc.
- *NoSQL databases* is a new database approach to overcome the limitations of traditional relational SQL databases in the case of massive data. This kind of databases present three architectures: key-value solutions such as Dynamo and Voldemort,

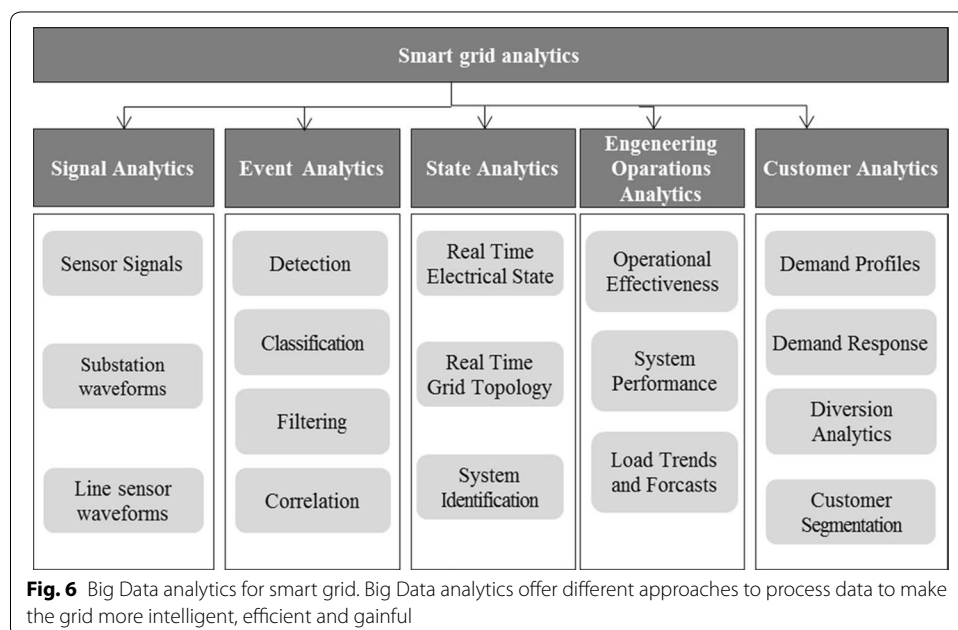
column-oriented solutions such as Cassandra and HBase and documents databases solutions such as MongoDB and CouchDB.

**Data analytics**

The grid collects data from different sources and stores it as a huge quantity of dataset that should be easily consumable for analytics. Analytics has a critical role to make the grid more intelligent, efficient and gainful. Figure 6 presents various kind of analytics in smart grids: (i) signal analytics which is based on signal processing, (ii) event analytics which focus on events, (iii) state analytics which help to have a vision about the state of the grid, (iv) engineering operations analytics which is responsible of the grid operating side, and (v) customer analytics which process customer data.

There are actually several models that can combine the various kind of the previous analytics classes such as descriptive, diagnostic, predictive, and prescriptive models. Each model describes an operation side of the grid. Descriptive models are used to describe customers behaviours in demand response programs and provide a basic understanding of their practices. After customers description, diagnostic models come to understand particular customers behaviours and analyse their decisions. All these previous models are useful to make predictive models to predict customers decisions in the future. Finally, there is prescriptive models which are the high level of analytics in smart grid, because they affect marketing, engagement strategies and the decisions to make [22].

Big Data processing can be done in two manners: The first is batch processing, which process data in a period of time and is used for data processing without high requirements on response time. The second, is stream processing and is used for real-time applications. This kind of processing requires a very low latency of response.



### **Data visualization**

Data visualization has a great role, because it improves the assessment of smart grid. Actually, there is a great number of visualization techniques based on multivariate high dimensional visualization which gives the ability to use 2D and even the 3D visualisation. But smart grids face enormous variables that complicate data presentation such as 3D Power-map etc. Scatter diagram, parallel coordinate, and Andrew curve for example resolve the problem of high dimensional data [23].

### **Data transmission**

Data transmission in Big Data plays a critical role, because it affects all the previous phases. So it should maintain high bandwidth capacity and speed, data security and privacy etc. Data transmission in smart grids is based on communication technologies as described in "[Communication systems](#)", starting by access network technologies including PLC, ZigBee, WIFI etc., followed by area network technologies, using M2M, Cellular networks, Ethernet etc. Then core network technologies with IP, IMPLS etc. Finally, backbone network technologies, which rely on fiber technologies, microwave link, IP-based Wavelength, Division Multiplexing (WDM) network and other optical technologies.

### **Criteria for choosing Big Data technologies**

Big Data technologies propose several tools, so utilities should determine which platforms and tools to deploy to meet their goals. Previous subsections have shown that Big Data life cycle is composed of five phases: data sources, data integration, data storage, data analytics and data visualization. Big Data analytics is the most important step in the life cycle. So, depending on the analytics process, utilities can identify data to acquire and how to store it and even the visualization techniques to use.

Electrical companies should consider certain amount of precautions to choose the right analytics solutions. There are a lot of criterias to take into account in term of speed of computation, compatibility, graphic capabilities, possibility to work on the cloud etc. As a result, utilities need a Multiple Criteria Decision Making (MCDM) tools. For decision making applications, the Analytic Hierarchy Process (AHP) is considered one of the most popular MCDM methods, because it takes in consideration the quantitative and qualitative performances. The AHP model can be used for the Big Data analytics platform selection based on criteria definition including technical, social, cost and policy perspectives [24]. Table 1 describes Big Data technical perspectives, including hardware and resources configuration requirements [24].

### **Big Data resources requirements**

Big Data solutions have large amount of challenges in term of storing and processing. Thus, utilities should be aware of all Big Data requirements before implementing it.

### **Big Data hardware requirements**

Big Data solutions require high volume of data storage and high velocity in processing. So, before installing these technologies, utilities should ensure all hardware requirements. The most essential components of a Big Data system are the processing

**Table 1 AHP model technical perspective**

Technical perspective	Criteria
Availability and fault tolerance	Redundancy and resilience in networks, servers, physical storage, etc.
Scalability and flexibility	Tools must be evolutionary and scalable
Performance (latency)	Data processing time (single transaction, query request)
Computational complexity	Computation tools extension (data mining, business intelligence)
Distributed storage capacity and configurations	Storage systems parameters, such as storage nodes needed in terms of availability, periodic basis, etc.
Data processing modes	Batch, real and hybrid processing
Data security	Security compliance according to the platform requirements

frameworks and processing engines which are responsible for computing over data. Table 2 provides hardware requirements for the most popular Big Data processing frameworks: Hadoop, Storm, Spark and Flink.

To run these Big Data engines, especially for real-time processing, utilities should dedicate additional funds and resources. Cloud computing helps electrical companies to overcome power and cost requirements and bring a great number of benefits.

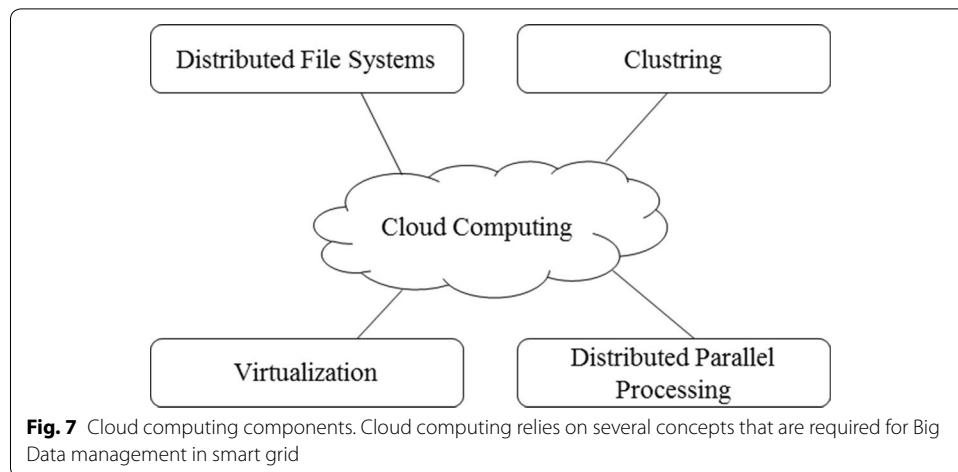
**Cloud computing frameworks**

Cloud computing solves many problems related to Big Data management for smart grids. It helps utilities to ensure the flexibility, agility and efficiency in terms of saving cost, energy and resources [25]. Cloud computing relies on several concepts that make it suitable for Big Data management in smart grids as illustrated in Fig. 7. The use of cloud computing in smart grid brings enormous benefits, due to redundancy, rollback recovery and multi-location data backup which increase data fault tolerance and security [13].

Cloud computing is based on service models. These models can be offered in public, private, or hybrid manner: (1) software as a service (SaaS) which provides applications and make them available to customers over the Internet, (2) platform as a service (PaaS) delivers hardware and software tools and gives customers the ability to create their own applications, (3) infrastructure as a service (IaaS) offers hardware, software, servers, and

**Table 2 Hardware requirements of some Big Data processing frameworks**

Framework	Hadoop	Storm	Spark	Flink
<b>Operating systems</b>	Red Hat Enterprise Linux (RHEL) v5.x or 6.x (64-bit) CentOS v5.x or 6.x (64-bit) SUSE Linux Enterprise Server 11, SP1 (64-bit)	CentOS Linux Windows	Windows XP/7/8 Mac OS X 10.7-9 Linux	Linux Mac OS X Windows (Cygwin)
<b>RAM</b>	64 GB at least	8 GB at least	8 GB at least	8 GB at least
<b>CPU</b>	2 cores at least	8 cores at least	8 cores at least	8 cores at least
<b>Network</b>	10 Gigabit at least	10 Gigabit at least	10 Gigabit at least	10 Gigabit at least
<b>Hard disk</b>	12–24 disks per node for each 1TB at least	6 disks per node for each 1TB at least	4–8 disks per node for each 1TB at least	12–24 disks per node for each 1TB at least



other IT infrastructure components over the Internet, (4) data as a service (DaaS) allows customers in addition to run applications to store data on-line, (5) communication-as-a-service (CaaS) is useful for messaging tools including voice over IP (VoIP), instant messaging (IM), and video conferencing, and (6) monitoring as a service (MaaS) is used for security services to ensure a third party security [26].

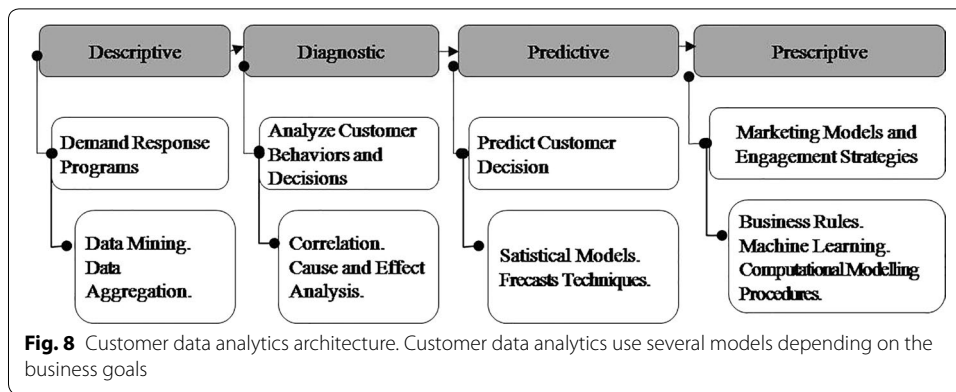
Cloud computing frameworks satisfy challenges of using Big Data technologies, so there is a great number of cloud solutions that can handle Big Data such as Amazon Elastic Compute Cloud (Amazon EC2), Google Compute engine, Microsoft Azure Cloud, IBM Docker Cloud, etc.

### **Big Data implementation in smart grid: the case of customer data analytics**

In this section, the focus will be on customers data analytics, because it involves the smart consumers concept, which makes consumers as potential producers of clean energy, players in their consumption and also main actors in production and consumption balancing. Customer data analytics is a great opportunity for utilities to understand customer behaviour better, and be able to make strategic decisions.

#### **Added value of customer data analytics**

Big Data analytics of customers data become a necessity and not a choice for electrical companies. Consumers are participating in smart grids as end customers through smart meters that offer them better control of their own consumption. Demand Response (DR) programs are used by utilities to obtain real-time information of the demand curves in the various points of consumption in order to calibrate and prognosticate more precisely. Thus, the production curve can be regulated according to demand more efficiently and reduce the losses of "overproduction". This will also make it possible to make a real-time diagnosis of meters and equipment close to the consumer, sending alarms, executing "self-healing" systems, etc.. Improving customer engagement is among the motivations of DR, because it helps utilities interact with the customers energy needs even during power outage. Dynamic pricing is also involved by DR; consumption monitoring avoid usage in peak time, so customers can check prices in real time and adapt their usage

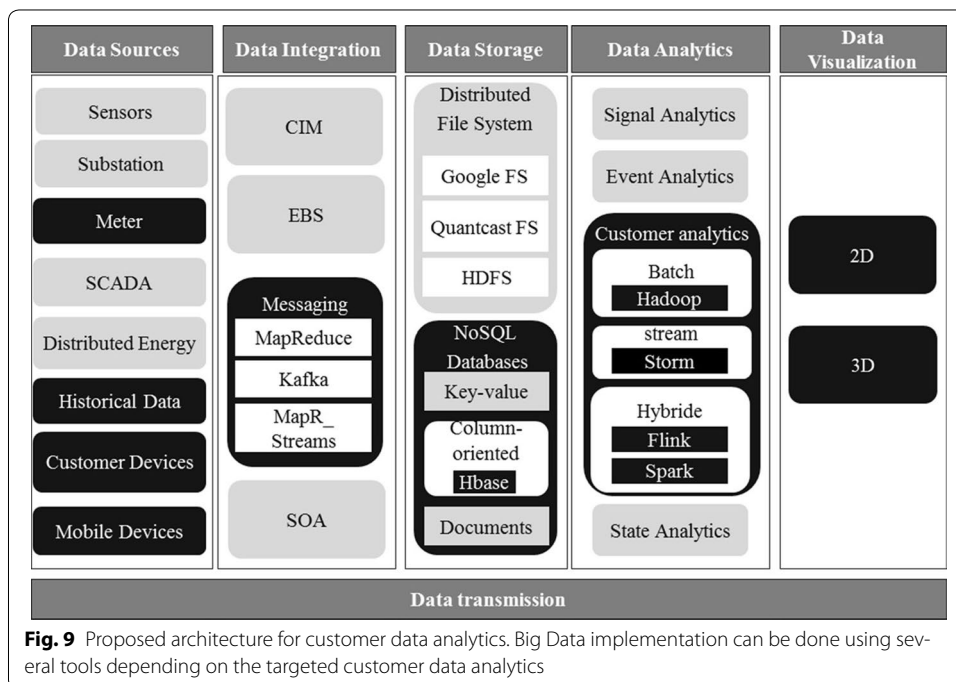


according to the electrical bills [12]. All of which is only possible using customer data analytics techniques as shown in Fig. 8.

**Big Data tools for customer data analytics**

Customers data is in the order of Terabytes and in a variety of formats. So, it requires high velocity, scalability and fault tolerance in data processing, storage and visualization. Big Data implementation can be done using several tools, but the analytics tools are the most critical in business choice. Figure 9 provides several Big Data technologies that can be used to manage smart grid data. The variety of customer data sources (smart meters, devices, historical data, etc.) requires the use of integration tools to make data uniform. Messaging tools are the most efficient for raw data integration and hence can be used for customer data integration.

Big Data analytics can be done using several processing mode:





- *Batch processing tools* Big data analytics offers a great number of methods to process data starting from batch processing. Hadoop [27] is a suitable choice for batch analytics for smart grid. Since smart grid systems are distributed geographically, distributed file systems are very useful for it. Hadoop has Hbase as a database system, Hadoop Distributed File System (HDFS) as a storage system, and MapReduce as a processing engine. Although, Hadoop can't handle modern Information Technology (IT) systems in data velocity, scalability and machine learning algorithms [28].
- *Real time processing tools* Real time processing is fast in term of execution than batch processing, because it handles data with high velocity requirements using stream processing or complex event processing systems. Real time processing can be implemented using several solutions such as S4, Splunk, Storm etc. Storm [29] is the most appropriate real time processing solution for smart grids, because it is open source, distributed and fault-tolerance and offers great number of opportunities as real time processing system, including message handling reliability, parallel computations and simple programming model etc. Storm can be used with Kafka for data integration and and Hbase for data storage.
- *Hybrid processing tools* Hybrid processing can handle both batch and real time processing. Spark [30] is a framework used for batch processing, but it has also real time processing solution with Spark streaming. Spark handles large-scale data processing, and also it includes useful tools such as Spark SQL, Spark Streaming, machine learning library and GraphX. All that make Spark meet Big Data requirements in smart grid. Spark streaming uses real time complex event processing engine to handle velocity issues. When using Spark, data storage can be done using HDFS or even Hbase [31]. Apache Flink [32] is another framework able to process data in both batch and stream modes. Flink is based on enormous APIs like transformations functions (map/reduce, group etc.), that make it scalable, easy to deploy, fault tolerance and fast in execution. Flink is efficient in machine learning, because it adopts its own machine learning library called FlinkML. Flink already has libraries to access HDFS, so it can be easily used with HDFS to store data.

## Conclusion

Smart grid systems collect huge quantity of datasets to bring smartness to the grid. In the same time this present challenges for utilities to deal with the nature, the distribution and the real-time constraints of the collected data. In this paper we have presented an overview of the opportunities, concepts and challenges of data management in smart grids and summarized the Big Data technologies and mechanisms that can be used to handle smart grid requirements including processing, storage and even visualization. We also provided the steps, tools and technical requirements for implementing and deploying Big Data technologies for smart grids in order to have an efficient and scalable data management.

### Authors' contributions

All authors read and approved the final manuscript.

### Authors' information

Houda DAKI is a Ph.D. student at the National School of Applied Sciences, University of Chouaib Doukkali, EL Jadida (Morocco). She received network and telecommunications engineering Diploma in 2015 from the National School of Applied Sciences, University of Chouaib Doukkali, EL Jadida (Morocco). Then she joined the the Laboratory of

Information Technologies at the National School of Applied Sciences as a PhD student. Her research interests include issues related to Big Data, Data analytics, Data mining and Smart Grids technologies.

Dr. Asmaa EL HANNANI is an Assistant Professor in Computer Science at the National School of Applied Sciences, University of Chouaib Doukkali, EL Jadida (Morocco). She received a Diploma degree (M.Sc) in computer science from the University of Fribourg (Switzerland), in 2003. In 2007, she obtained a joint Ph.D. degree in computer science from the University of Fribourg (Switzerland) and Institut National des Télécommunication, Evry (France). Then, she joined the Department of Computer Science, University of Sheffield (UK) as Research Associate within the Speech and Hearing Research. In 2010 she joined the Department of Telecommunications, Networks and Computer Science at the National School of Applied Sciences, teaching engineering students in the area of software engineering with a focus on web/mobile app design and development. Her research interests include biometrics technologies, speech processing and issues related to Big Data analytics.

Dr. Abdelhak AQQAL received a joint Ph.D. degree in Sciences and Technologies of Information and Communication (STIC) from the Chouaib Doukkali University (Morocco) and Darmstadt University of Technology (Germany) under the DAAD Sandwich Programme. He joined the National School of Applied Sciences of ElJadida (ENSAJ) in 2010 as Assistant Research Professor in the Department of Telecoms, Networking and Informatics. His research interests are mainly in the area of STIC and centered around developing and integrating innovative approaches to advance human education and life, with a focus on using Green and Smart Technologies in Morocco.

Dr. Abdelfatteh HAIDINE received his Ph.D. from Dresden University of Technology in Germany. He worked as consultant and manager with big companies (KEMA, ACCNETRUE) for deployment of smart metering and smart grid applications. Currently he is a Lecturer at the department of Telecoms, Networking and Informatics at the National School of Applied Sciences El Jadida, Morocco. His research interests include issues related to Machine-to-Machine (M2M) communications, networking technologies for smart city and smart grid applications, as well as application of combinatorial optimization in network planning and migration.

Dr. Aziz DAHBI received his Ph.D. in Sciences and Technologies of Information and Communication (STIC) at the University Chouaib Doukkali in El Jadida, Morocco. He is an Assistant Professor at the Department of Telecommunications, Networks and Computer Science, National School of Applied Sciences, University Chouaib Doukkali. His research interests include issues related to Elearning, Green Networking, and Emerging technologies. He is author of research studies published at national and international journals, conference proceedings and book chapters.

#### Acknowledgements

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 28 February 2017 Accepted: 17 March 2017

Published online: 28 April 2017

#### References

1. Wang W, Lu Z. Cyber security in the smart grid: survey and challenges. *Comput Netw.* 2013;57:1344–71.
2. McGranaghan M, Schmitt DHL, Cleveland F, Lambert E. Enabling the integrated grid: leveraging data to integrate distributed resources and customers. *IEEE Power Energy Mag.* 2016;14:83–93.
3. Agarwal V, Tsoukalas LH. Smart grids: importance of power quality. In: Proceedings of first international conference on energy-efficient computing and networking. 13–15 April 2010. Berlin; 2010. p 136–143
4. Amin SM. Smart grid: overview, issues and opportunities. *Advances and challenges in sensing, modeling, simulation, optimization and control.* *Eur J Control.* 2011;17:547–67.
5. Yan Y, Qian Y, Sharif H, Tipper D. A survey on smart grid communication infrastructures: motivations, requirements and challenges. *IEEE Commun Surv Tutor.* 2013;15:5–20.
6. Verdone R, Dardari D, Mazzini G, Conti A. *Wireless sensor and actuator networks: technologies, analysis and design.* Cambridge: Academic Press; 2008
7. Xu J, Wang J, Xie S, Chen W, Kim J-U. Study on intrusion detection policy for wireless sensor networks. *Int J Secur Appl.* 2013;7:1–6.
8. Xiaomeng Y, Fangming L, Jiangchuan L, Hai J. Building a network highway for big data: architecture and challenges. *IEEE Netw.* 2014;28:5–13.
9. Chen M, Mao S, Liu Y. Big data: a survey. *Mob Netw Appl.* 2014;19:171–209.
10. Wang W, Xu Y, Khanna M. A survey on the communication architectures in smart grid. *Comput Netw.* 2011;55:3604–29.
11. Yigit M, Gungor VC, Tuna G, Rangoussi M, Fadel E. Power line communication technologies for smart grid applications: a review of advances and challenges. *IT Comput Netw.* 2014;70:366–83.
12. Siano P. Demand response and smart grids—a survey. *Renew Sustain Energy Rev.* 2014;30:461–78.
13. Fang B, Yin X, Tan Y, Li C, Gao Y, Cao Y, Li J. The contributions of cloud technologies to smart grid. *Renew Sustain Energy Rev.* 2016;59:1326–31.
14. Jaradat M, Jarrah M, Bousselham A, Jararweh Y, Al-Ayyoub M. The internet of energy: smart sensor networks and big data management for smart grid. *Proc Comput Sci.* 2015;56:592–7.

15. Singla A, Sachdeva R. Review on security issues and attacks in wireless sensor networks. *Int J Adv Res Comput Sci Softw Eng*. 2013;3:529–34.
16. Khushboo G, Vaishali S. Design issues and challenges in wireless sensor networks. *Int J Comput Appl*. 2015;112:26
17. Malhotra J. Review on security issues and attacks in wireless sensor networks. *Int J Future Gener Commun Netw*. 2015;8:81–8.
18. Rehman MHU, Batool A. Pattern-based datasharing in big data environments. *Digit Technol*. 2015;1:39–42.
19. Minguez J, Jakob M, Heinkel U. A soa-based approach for the integration of a data propagation system. In: *Proceedings of the 9th international conference on information reuse and integration*. 10–12 August 2009. Melbourne; 2009. p 47–52
20. Vera-Baquero A, Colomo-Palacios R, Molloy O. Business process analytics using a big data approach. *IT Prof*. 2013;15:29–35.
21. Messaging Integration. <http://www.informationbuilders.com/messaging>
22. Stimmel CL. *Big Data analytics strategies for the smart grid*. Boca Raton: CRC Press; 2014.
23. Nga DV, See OH, Quang DN, Xuen CY, Chee LL. Visualization techniques in smart grid. *Smart Grid Renew Energy*. 2012;3:175.
24. Lněnička M. Ahp model for the big data analytics platform selection. *Acta Inform Pragensia*. 2015;4:108–21.
25. Mastelic T, Oleksiak A., Claussen H, Brandic I, Pierson JM, Vasilakos AV. Cloud computing: survey on energy efficiency. *ACM Comput Surv (CSUR)* . 2015;47:33
26. Rittinghouse JW, Ransome JF. *Cloud computing: implementation, management, and security*. Boca Raton: CRC Press; 2016.
27. Apache Hadoop. <http://hadoop.apache.org>. Accessed 29 Jan 2017
28. Shyam R, Ganesh HBB, Kumar SS, Prabakaran P, Soman KP. Apache spark a big data analytics platform for smart grid. *Proc Technol*. 2015;21:171–8.
29. Apache Storm. <http://storm.apache.org>. Accessed 5 Feb 2017
30. Apache Spark. <http://spark.apache.org>. Accessed 20 Feb 2017
31. Liu G, Zhu W, Saunders C, Gao F, Yu Y. Real-time complex event processing and analytics for smart grid. *Proc Comput Sci*. 2015;61:113–9.
32. Apache Flink. <https://flink.apache.org>. Accessed 15 Jan 2017

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)

---