Power Electronics and Drives
Volume 4(39), 2019 DOI: 10.2478/pead-2019-0011

sciendo

Power Electronics and Drives

A Review on Big Data Management and Decision-Making in Smart Grid

Research Article

Amira Mohamed[1,2], Shady S. Refaat[1], Haitham Abu-Rub[1]

[1]*Electrical and Computer Engineering Department, Texas A&M University at Qatar, Doha, Qatar*
[2]*Electrical and Computer Engineering Department, Texas A&M University, College Station, TX, USA*

Received May 08, 2019; Accepted June 26, 2019

Abstract: Smart grid (SG) is the solution to solve existing problems of energy security from generation to utilization. Examples of such problems are disruptions in the electric grid and disturbances in the transmission. SG is a premium source of Big Data. The data should be processed to reveal hidden patterns and secret correlations to extrapolate the needed values. Such useful information obtained by the so-called data analytics is an essential element for energy management and control decision towards improving energy security, efficiency, and decreasing costs of energy use. For that reason, different techniques have been developed to process Big Data. This paper presents an overview of these techniques and discusses their advantages and challenges. The contribution of this paper is building a recommender system using different techniques to overcome the most obstacles encountering the Big Data processes in SG. The proposed system achieves the goals of the future SG by (i) analyzing data and executing values as accurately as possible, (ii) helping in decision-making to improve the efficiency of the grid, (iii) reducing cost and time, (iv) managing operating parameters, (v) allowing predicting and preventing equipment failures, and (vi) increasing customer satisfaction. Big Data process enables benefits that were never achieved for the SG application.

Keywords: *Big Data • energy management • Big Data analytics • smart grid • decision-making*

1. Introduction

Smart grid (SG) has been emerged as the most ingenious idea worldwide towards solving the contemporary pressing power demand problems. SG provides a two-way power and communication flow between customers and utilities to improve grid controllability and reliability (Cagri Gungor et al., 2013). SG is supported by many data sources such as smart metres, sensors, detectors, and measurement units; therefore, a huge amount of data is being exchanged between the elements of SG. The biggest challenge facing SG is to collect and deal in real time with massive and important amount of data. Example is the data acquired by intelligent electronic devices, operating and maintenance data for electrical devices and equipment, and very large datasets used in decision-making, such as metrological service data and geospatial information system. These data are updated always through the fast technology growth. There is a consensus about the concept of the three V's characterizing Big Data: volume, variety, and velocity (Ward and Barker, 2013). SG data achieve the three properties of the Big Data: volume, variety, and velocity. Big Data in SG are classified into two broad categories: structured and unstructured data (Thakur and Mann, 2014). "Structured data" can be ordered in columns, rows, or binary. These data can be collected from various sources such as phasor measurement data, sensors, and smart metres. "Unstructured data", which are not organized, can be collected from photos, customer comfort level, social media, emails, etc. All generated massive amounts of data can provide better understanding of customer segmentation and can help improve the efficiency of electrical generation and scheduling. New forms of data processing are required to

* *E-mail: Amira.mohammed@qatar.tamu.edu, Shady.khalil@qatar.tamu.edu, haitham.abu-rub@qatar.tamu.edu*

manage these data and enable enhanced decision-making. Big Data process is divided into data management and analytics. Data management contains Big Data storage, mining, and integration to prepare and retrieve it for analysis. Data analytics include analyzing the managing data to be in a useful form for decision-making. This mass of information is essential to make SG more efficient, reliable, secure, independent, and supportive during normal conditions and contingencies (Fang et al., 2012). SGs allow greater possibility of energy delivery and power flows. The technology needed to collect this massive amount of data is obtainable today, but managing and extracting useful information, particularly in real time, are still a Big challenge (Weilki, 2013; Zhang, 2013). Recently, several strategies have been developed to manage and analyze Big Data in SG (Borkar et al., 2011, 2012; Cattell, 2011; Chandarana and Vijayalakslnni, 2014; Fadnavis and Tabhane, 2015; Hoffmann, 2013; Kersten et al., 2011, 2012; Mandai et al., 2015; Mohan, 2013; NIST, 2010; Oracle Corporation, 2011; Sadalage and Fowler, 2012; Stonebraker et al., 2010; Zikopoulos et al., 2013). Examples are Hewlett Packard Enterprise Extends Flagship (HPE IDOL), Hadoop Stream processing (Chandarana and Vijayalakslnni, 2014), Parallel Database Management System (DBMS), Google Cloud Dataflow, and Apache Spark (Hoffmann, 2013; Kleiner et al., 2012). Two systems have emerged to address Big Data process. The DBMS uses a shared-nothing architecture, where data are spread over a cluster based on a partitioning strategy, usually hash based (Borkar et al., 2011). This system is robust and stable and seems to achieve good performances under the condition of appropriate arrangement and organization of the datasets in a hierarchical manner. However, many obstacles must be overcome to achieve good results such as slow extraction if the datasets are not organized, as well as if the complexity of relational database increases. The skill set required by the DBMS also increases, and then the DBMS requires an installation of expensive storage systems and proprietary servers. Hadoop MapReduce has emerged as one of the major techniques used for specific data analytics tasks (Fadnavis and Tabhane, 2015; Stonebraker et al., 2010). Hadoop MapReduce is the most robust and widely used tool for very large-scale batch data processing. This system is highly applicable for offline Big Data analytics in SG due to high-throughput data; however, it becomes inefficient and requires algorithmic design effort to implement solutions for real-time analysis on Big Data such as weather forecasting and customer comfort level (Mandai et al., 2015). Recently, the NoSQL (non-relational or not only SQL) database was created for the development of efficient data management solutions (Cattell, 2011; Sadalage and Fowler, 2012). The NoSQL system seems to provide very simple indexing strategies in comparison to relational DBMS. However, there are two major drawbacks associated with this system: it does not provide general “one-fits-all” solution and it requires a lot of expertise and programming efforts for implementing solutions (Borkar et al., 2012; Mohan, 2013; Oracle Corporation, 2011). These systems were implemented as completely separate systems; however, the need is to develop innovative hybrid combinations of the three systems (NIST, 2010; Zikopoulos et al., 2013). In addition, using Big Data analytics as a key to deal with uncertainties and different sizes from terabytes to zettabytes of structured and unstructured data is a challenge. This paper provides a comprehensible analysis of theB Data in addition to its definition, classification, and utilization for SG advantages. The paper studies the impact and benefits of adopting Big Data mining, analysis, processing, and management to ensure better stability, safety, and reliability of SG. The objective of this review paper is to summarize and compare some of the well-known techniques used in various stages of Big Data process within the context of SG applications. This paper provides a methodical summary of the best techniques that can reduce the challenges in managing and analyzing Big Data in SG to increase the grid reliability and efficiency and to help in decision-making at various levels. The main contribution of the paper is providing recommended system to process the Big Data in future SG. Therefore, the objective of this paper is to offer the following:

1. A comprehensive review of Big Data processing techniques.
2. Discussing about challenges that these techniques face for the application in SG.
3. Developing a dynamic and effective Big Data management and processing system while overcoming Big Data processing challenges and improving the reliability, efficiency, and stability of SG.

2. Big Data sources

Reliable and real-time monitoring is one of the requirements to prevent the possible power disturbances and outages. Hence, extracting the relevant information from the available data needs deep understanding of the sources of these data. Each smart power grid is made of a large number of sensors, smart metres, detectors, and

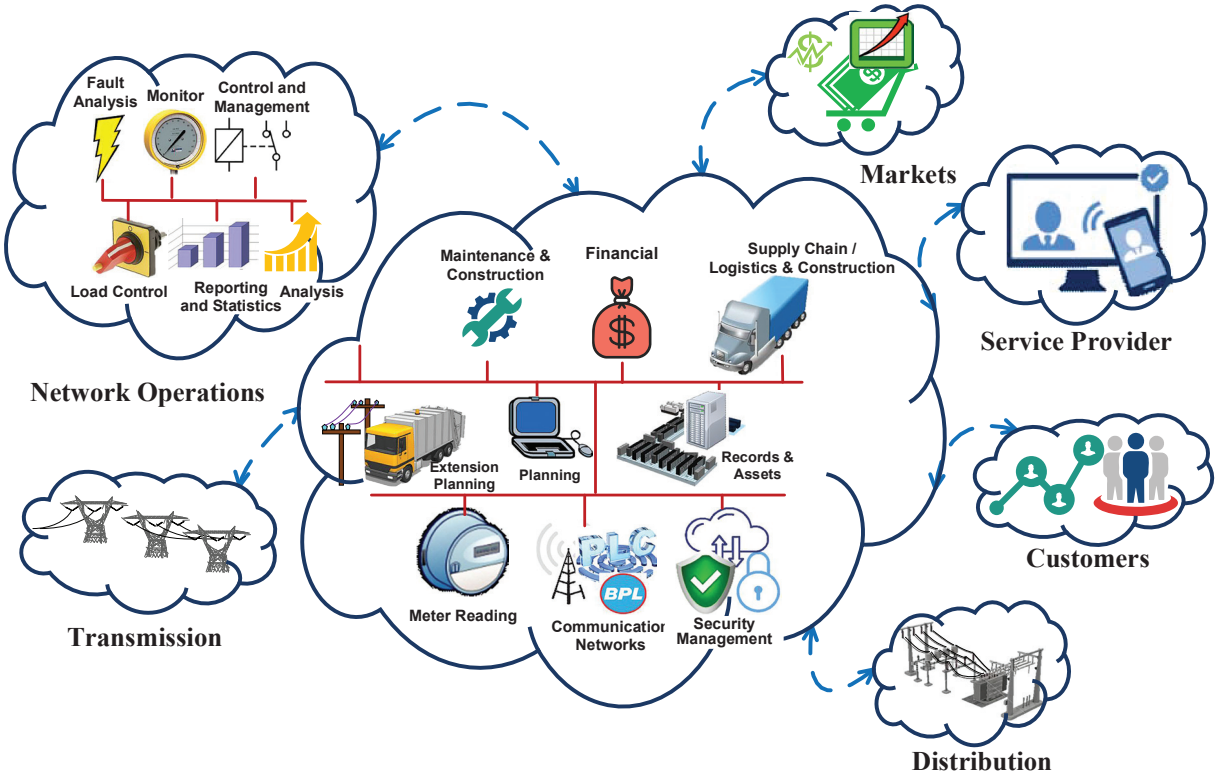

Fig. 1. SG components and architecture

measurement units. Fig. 1 shows the SG components and architecture. The data collected from these elements are huge and satisfy all the three V's characterizing Big Data: volume, variety, and velocity (Gungor et al., 2010; Sagiroglu and Sinang, 2013; Saha and Srivastava, 2014). Volume refers to the quantity of generated and stored data. The question that needs to be clarified is how the Big Data is. Fig. 2 answers this question by giving the graduation in the data volume and illustrating the criteria of velocity and variety as well. Velocity refers to the speed at which the data are generated and processed to meet the demands. Variety refers to the type and nature of the data. The use of Big Data in appropriate way requires studying and understanding different types of these data and managing the massive data in a simple manner that is easy to act upon it. Big Data in SG is divided into two types (Chandarana and Vijayalakshrni, 2014): structured and unstructured. "Structured data" refer to all data that have the advantage of being easily entered, stored, queried, and analyzed. Examples of these data are phasor measurement data, operational data for running utilities, energy consumption data measured by the widespread smart metres embedded in electronic devices, smartphones data, and data from global positioning system (GPS). Structured data also include energy market pricing and bidding data collected by advanced metering infrastructure, the management, control, and maintenance data for devices in the power system acquired by intelligent electronic device. "Unstructured data" include more complex information such as customer comfort level, photos, financial data, and very large datasets used in decision-making, such as weather and lightning data, geographic information system data, seismic reflection data, animal migration data, financial market data, social media data, Email, and regulatory reporting data. Unstructured data generally cannot be separated into categories or analyzed numerically easily. The impressive growth of Internet technology in recent years resulted in the appearance of a huge volume of different variety of datasets that vary in origin, size, speed, form, and function. Fig. 3 shows the Big Data sources used by SG. Smart monitoring and sensing capabilities are necessary to achieve real-time response from SG sensors, which are input devices that convert physical stimulus into an electrical signal. SG sensors are used to monitor and control different assets in SG, compared to traditional communication technologies. Nowadays, the direction is to use wireless sensor networks (WSNs) in SG because of their flexibility and low cost (Beyer, 2011).

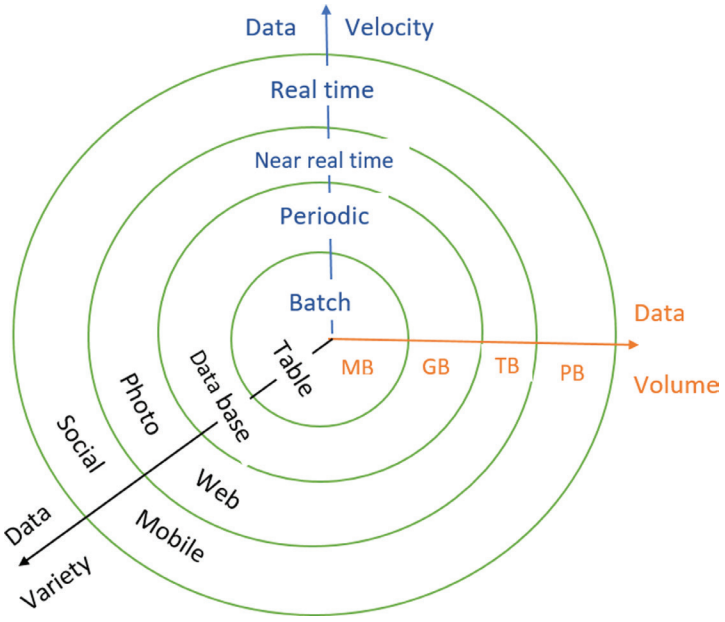

Fig. 2. Big Data: three V's characteristics (Demchenko et al., 2013)

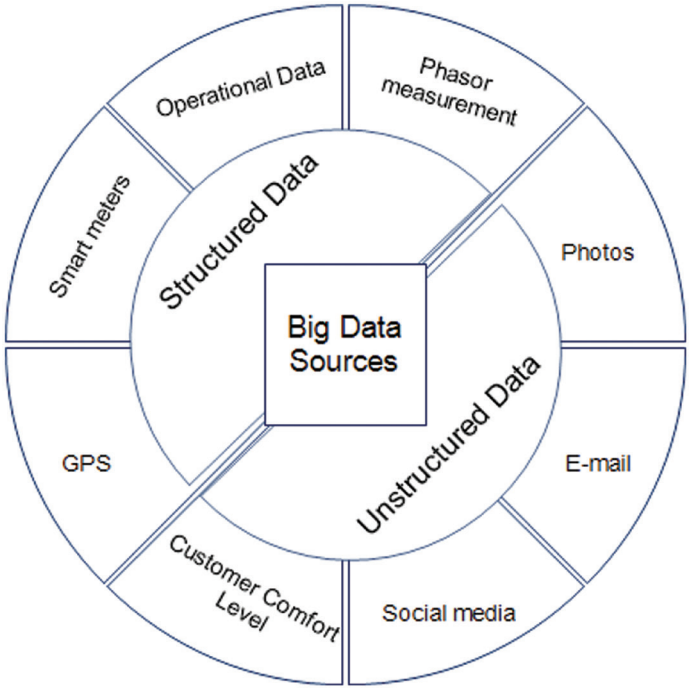

Fig. 3. Big Data sources in SGSG, smart grid

3. Big Data process

Big Data need more compelling techniques to process the huge volume of information within a limited time. The challenge is to manage SG through information management instead of focusing just on the volumes of data. In spite

of the importance of volume in managing the whole data information, variety and velocity are still very important to focus on. It is essential to process these data and obtain patterns from it to make better decisions on time. Creating such a strategy requires a robust understanding of Big Data process as illustrated in Fig. 4. Big Data process is divided into two main platforms: data management and data analytics. The data from the SG network equipment such as metres, sensors, devices, substations, and mobile need to be managed for extracting the useful data before entering into the analytics platform. “Data management” platform contains Big Data storage, mining, integration, aggregation, and representations to prepare and retrieve the data for analysis. “Data analytics” platform refers to modelling and analyzing the managed data and present it in readable form. The goal of Big Data analytics is to extract useful values and information to support decision-making. In the next sections, each step of Big Data process is discussed. Then, the discussion is moved to the crosscutting challenges and ending with a proposed system built using the best techniques that manage and analyze Big Data in SG. The goal is to decrease the challenges, improve the efficiency of the grid, increase grid reliability and customer satisfaction, and help in online decision-making.

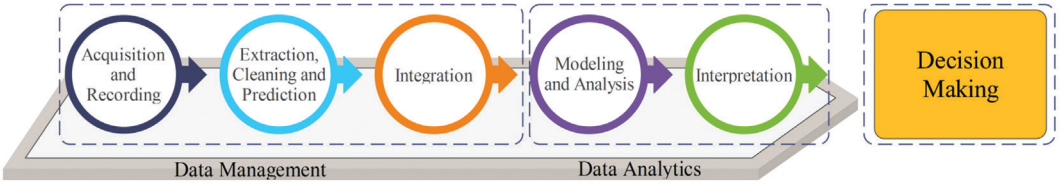

Fig. 4. Big Data process

3.1. Big Data management platform

The Nig Data management platform is a crucial step in the process to prepare massive amount of data for data analytics platform. The data management platform involves two major features: “identify” targeted plans according to the requirements, “extract and proportionate” useful information that achieve the targeted plans to minimize cost and time consumed for analyzing and processing all the obtained data from widely deployed measurement devices and other sources in SG.

3.1.1. Acquisition and recording

The first step in the Big Data management process is defining, storing, and mining massive amount of data from diverse sources. Supervisory Control and Data Acquisition (SCADA) systems collect data from sensors, perception devices, and substations. The main function of SCADA system is helping in managing the power flow by providing remote monitoring and control of electric network devices. With the Big Data size increased, the complexity of monitoring and controlling power network will significantly increase. Cloud computing services are emerging as an essential component to host SCADA systems. Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (Jararweh et al., 2014; Markovic et al., 2013). There are many advantages for using cloud services for SCADA systems such as achieving greater reliability, enhancing functionality, and reducing energy cost (Bonomi et al., 2012). However, it raises security and network concerns because the storage in the cloud is shared among multiple users.

Fog computing is used to overcome the drawbacks of currently available cloud computing. Fog computing is concentrated in devices and operates on network ends; this concentration means that data can be processed locally in smart devices rather than establishing channels for cloud storage and utilization (Srinivas and Togiti, 2015). By using this kind of distributed strategy, the method gains lower cost with improved security.

3.1.2. Extraction, cleaning, and prediction

Big Data mining is the capability of extracting useful information from large dataset. Technically, data mining is the process of finding correlations or patterns among large-volume, heterogeneous, autonomous data in large relational database (Thakur and Mann, 2014). Data mining process includes extraction, cleaning, annotation, and analyzing a vast array of information through six steps (Kusum and Rupali, 2013). Those steps are as follows: classification, estimation, prediction, association rules, clustering, and description: “classification step: the data classified into groups” (Wu et al., 2014). There are several algorithms used in data classification such as decision tree, Naive Bayes, AdaBoost, and Apriori (Fahad et al., 2014). Classification is based on checking the properties of a newly presented

data and assigning it to a predefined class. Classified data include some unknown continuous variables that must be estimated via “estimation step”: get particular values that can be processed with the current state variables, along with their uncertainties. Once the outcome of the classification step includes uncertain data, the estimation using the presented input measurements and the previously calculated state, along with the uncertain data, helps prepare the data for the prediction and to update the outcome. “Prediction step” refers to the expected outcome in the future based on the knowledge and experience. Prediction step attempts to form patterns that permit to predict the next measurement, given the available input data. Output data are composed of various rules. “Association rules step” is an attempt to build relationships between a set of data in the database. Complex rules are always generated by the same grouped data attributes identified through variable clustering. “Clustering step” is an attempt to organize a particular set of measured data into a partition or a set based on their characteristics, aggregating according to their similarities. The most popular Big Data’s clustering techniques that can be applied in SG are data mining clustering algorithms, dimension reduction, parallel classification, and MapReduce framework (Nagpal and Mann, 2011; Ben Ayed et al., 2014; Electric Power Research Institute (EPRI), 2009; Ferreira Cordeiro et al., 2011; Sherin et al., 2014; Yadav et al., 2013). The choice of most suitable clustering techniques for SG applications depends on the dataset characteristics. Fig. 5 illustrates the different techniques of clustering in terms of Big Data in addition to the advantages and limitations of each technique. However, the implementation of the theory of comparative advantage between those clustering techniques shows clearly that the data mining clustering algorithms for Big Data in SG have a major weak point: it does not have the capacity to deal with a huge amount of data (Shirkhorshidi et al., 2014). Even if supported by using the dimension reduction technique to reduce the size of the dataset, the result would be undesirable (Zerhari et al., 2015). The parallelization technique is applicable to implement in many SG applications, but the main drawback in using it is the complexity of the parallel computing systems such as clusters, grids, and distributed systems (Shirkhorshidi et al., 2014). On the other hand, the MapReduce framework is another technique that presents very satisfactory results, offers scalability, and generates instant response. The last step in the data mining process is the “description step”, which refers to how to describe the chosen data in a structured form, patterns, or in a row data suitable to match with the analyzing techniques that will be implemented to help in decision-making.

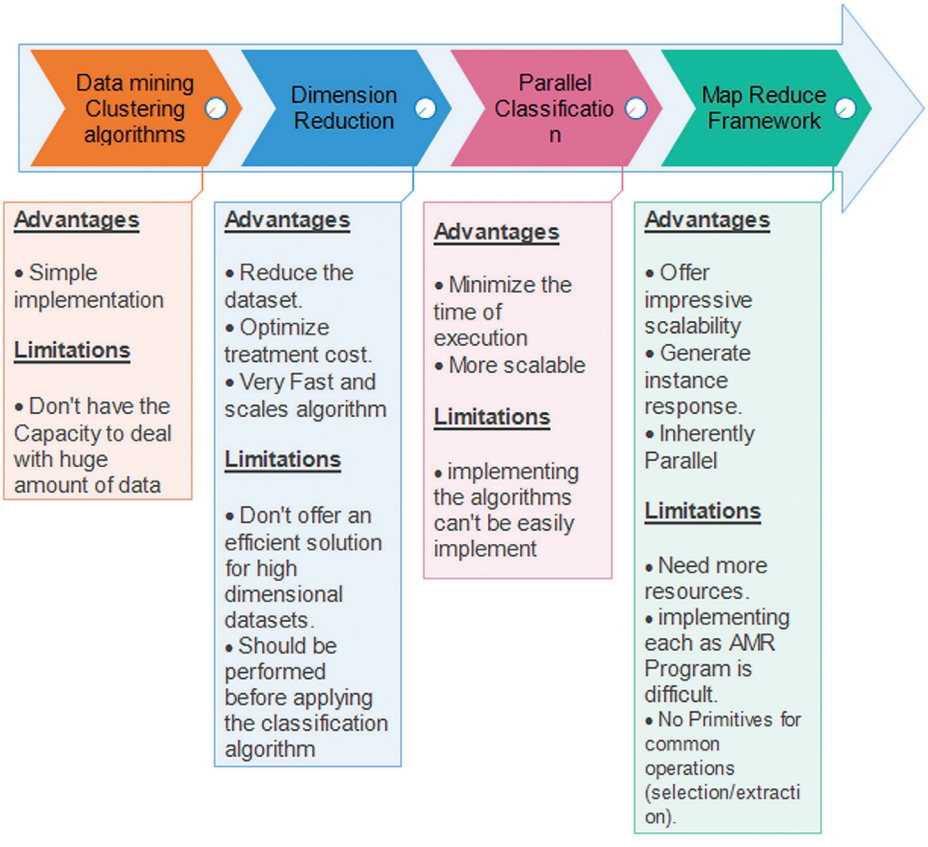

Fig. 5. Big Data clustering techniques

3.1.3. Big Data integration

SG data require integration stage to extract details of asset information and maintenance work. The main purpose to integrate the Big Data patterns is supporting standards governing sources of Big Data such as real-time pricing, distributed energy resources including demand response, distributed generation, energy storage, and consumer access to energy usage information. Architectural frameworks such as Zachman enterprise architecture framework, ISO Open-Electronic Data Interchange (EDI) reference model, and the Common Information Model (CIM) are used to integrate the Big Data to be ready for the analytics process (Reinprecht et al., 2011). CIM architecture is the most common technique fitting into SG applications due to flexible integration and layered systems (Rohr et al., 2011).

3.2. Big Data analytics platform

The current and future deployment of SG devices is producing a huge amount of data: these data characterized in specific patterns can be turned into comprehensible information for operational decision-making using analytics technique, which is divided into advanced analytics and predictive analytics (Shyam et al., 2015). “Advanced analytics” are used to analyze data that help in improving customer service. The time needed in analytics process can warn customers and utility during normal and contingency operating conditions. However, as fast the data can be analyzed and create the action, the fast recovery will be gained. “Predictive analytics” is analyzing the current data to make predictions about future and anticipating outcomes and behaviours based upon the data and not on assumptions. Implementing analytics techniques on the Big Data in SG can improve power distribution reliability as well as achieve continuity to the power grid by reducing power outages. The main benefit of Big Data analytics in SG are divided into four basic categories: first, “improving economic dispatch” based on optimizing asset life cycle and reducing the cost of condition-based maintenance; second, “improving reliability” through better management of operating parameters of SG and allowing predicting and preventing equipment failures; third, improving “efficiency” by reducing maintenance visits through monitoring the grid performance and giving the opportunity to predict any failure and take the suitable timely decision; fourth, increasing “customer satisfaction” through providing field insights into customer experience and taking into consideration the customer complaints and comfort level. The Big Data analytics process is divided into two main steps: modelling and analyzing, and the interpretation is discussed in the next section with a brief description of the most used techniques in each step and the challenges that each technique faces. To gain all the previous benefits from data analytics to serve SG, the selected techniques must be carefully chosen.

3.2.1. Modelling and analysis

After the integration stage, data are becoming ready to go through analytic techniques. In this paper, a brief description is given for the famous processing techniques used to model and analyze diverse data sources in SG. In addition, the most suitable technique will be selected which can fit Big Data analysis to achieve better performance and scalability in SG. Three types of processing techniques can be implemented for Big Data analysis (Afrati et al., 2011): batch, stream, and iterative processing. For “batch processing”, data are divided into small sets to process through Hadoop MapReduce technique, which is the most considerable technique used for analyzing large datasets (Shahrivari, 2014). MapReduce technique is used only for static applications in SG but for all applications that require real-time response such as demand response, real-time usage and pricing analysis, online grid control and monitoring, etc. Hadoop MapReduce is not efficient for most applications of SG. “Stream processing or real-time processing” deals with the new data independently instead of waiting for the next stream of data to process. Processing data through Apache Spark technique needs more analytics: this technique is combining batch, real-time, and iterative data processing requirements (Alexandros and Jagadish, 2012; Elluri and Salim, 2016; Xhafa et al., 2015). This makes stream processing very useful in SG applications. “Iterative processing” is a combination between the batch and stream processing: the data are analyzed through Apache Spark technique. However, iterative approach is processing all the variety of data but it is considered as time-consuming technique (Shyam et al., 2015). SG can be made more intelligent by choosing the optimal technique that can analyze all the diversity of data in SG.

3.2.2. Interpretation

The output from analytics process is gathered into different forms such as binary codes, tables, or functions. The output data are substantially containing information. Such information is the data that are accurate and timely:

specific and organized for a purpose; these data are presented in a form that gives it meaning and relevance and can lead to an increase in understanding and decrease uncertainty in SG (Park et al., 2015). Transferring output data into clear and readable information is the last step before decision-making. That can be achieved by creating an effective infrastructure to provide supplementary information that is known by interpretation. There are many ways to display data such as infographics, dials, gauges, geographic maps, spark lines, heat maps, detailed bar, pie, and fever charts. The information that has been interpreted is presented in a more meaningful context and moves to last platform which is decision-making.

3.3. Decision-making

Ultimately, a decision is made based on the result of data analysis after being interpreting to clear information. Decision-making is the appropriate plan that can be selected to protect the grid from any sudden outage or any failure through transforming the combined function into probability distributions.

Decision-making step should make those decisions quickly and provide clear information. There are some concomitant challenges to decision-making process (Acquisto et al., 2015): they are as follows:

- Decision-makers must be trained with complex problems and aware of Big Data analytic results.
- Cyber security is a key challenge of the SG decision-making: it must be built-in as part of SG design and must not come afterthought.
- he massive amounts of data available are often incomplete or unreliable which affect the validity of the decision.
- Decisions must be made in dynamic environments based on partial information.

4. SG data process: key challenges

The Big Data challenges are becoming one of the most exciting research and investment opportunities for the next years. The most famous Big Data process challenges in SG are divided into three major groups of challenges as shown in Fig. 6, which are the data challenges, the data management challenges, and the data analytics challenges.

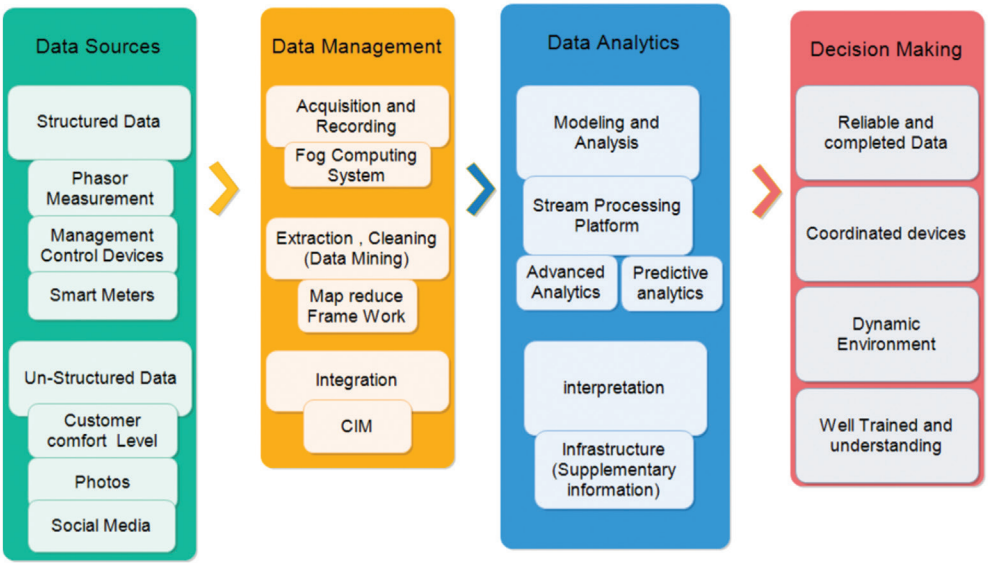

Fig. 6. Big Data challenges

4.1. Data challenges

Decisions that previously were based on some information and calculations now can be done based on the flood of data. Such Big Data have different properties and lead to some challenges. The three V's characterizing Big Data are volume, variety, and velocity (Gungor et al., 2010). "The volume" of data is already enormous and increasing

every day. Managing these large amounts of data is a challenge as the data volume is scaling faster than the speed of computer resources. Cloud computing now works on dealing with this challenge. “Velocity” refers to the speed of data generation and the time for its processing and analyzing. Most of the analytical approaches are unable to cope with this huge and fast flood of data. “The variety” of data being generated is also expanding, and the possibility to process this data is difficult. There is a need to change the techniques that are used for planning and managing the process to achieve the potential of Big Data.

4.2. Data management challenges

The fast growth in the SG data is faced with the challenge of managing these huge data while ensuring the security and reliability of the power grid. Big Data management challenges are divided into the following: “accessible information and security”: the significance of some requirements such as keeping the data accessible and usable is important for managing Big Data in SG to achieve preventive decisions; however, this needs to be aligned with all relevant legislative and regulatory instruments with data security; “precise data”: if the data used to support complex analysis and decision-making, these data need to be accurate and complete; “distributed mining”: many data mining techniques are not accurate; therefore, practical and theoretical analyses have to be provided to choose the useful data for Big Data decision if tagged and analyzed.

4.3. Data analytics challenges

Frequently, the Big Data collected is not in a ready form for analysis through some techniques, unless some challenges are considered such as heterogeneity and incompleteness. Although the data are filtered, some errors in data are remained. These errors must be managed during data analysis. Doing this correctly is a challenge.

- “Scale”: in the past managing the Big and rapidly increasing volumes of data has been done through fast processors , but now due to the faster development of the Internet, the data volume is scaling faster than computing resources. Recent work on managing probabilistic data suggests move towards cloud computing (Weilki, 2013; Zhang, 2013).
- “Timeliness”: the larger the dataset to be processed, the longer it will take to analyze. To decrease the time taken through the analytic process, new index structures are required for the used technique.
- “Privacy”: protecting the privacy rights of the information is considered a Big challenge. The potential value of Big Data is a function of the number of relevant, disparate datasets that can be linked and analyzed to reveal new patterns, trends, and insights and to be protected which will need to be carefully managed (Bonomi et al., 2012).

4.4. Proposed system

This paper provides a proposed system that can be used to minimize the challenging factors and improve the grid efficiency. Fig. 7 shows the proposed system architecture starting from collecting the Big Data from the SG components. These flows of collected data pass through Big Data management followed by data analytics and reach the last stage which is decision-making. It is clear that there is a huge amount of data collected from various sources in SG. These collected data are divided into structured data and unstructured data. The first category represents the data that have a defined length and format such as the smart metre reading, operational data for utilities, energy consumption data, and measured data. The unstructured data represent the data that do not fit neatly into the traditional structure of relational databases such as videos, audio files, web pages, and user interface through emails, social media, customer comfort level, and demand response programs. A centralized server called SCADA system analyzes the information, while a decentralized model with microgrids can complement SCADA. In the proposed system, fog computing operates on network ends, which allows to process the data in smart devices instead of establishing special channels to utilize these data. This can improve system scalability, reduce costs, ensure rapid response of energy, improve security, and integrate distributed power generators with the main power grid (Béjar Alonso, 2013; Wei et al., 2014). The interplay between the fog and SCADA can benefit SG greatly. These data must be filtered to extract useful information from large datasets that can be achieved by classifying the data, then organizing each particular set of objects into a partition using different algorithms. The choice of algorithm will always depend on the characteristics of the dataset. In the proposed system, Hadoop MapReduce framework is used, as it is more cost-effective for handling large, complex, or unstructured datasets than conventional approaches. Furthermore, it offers massive scalability and speed. MapReduce simplifies processing on large datasets (Bredillet et al., 2010). To build a flexible system architecture that can adapt in line with SG infrastructure development and

Fig. 7. Proposed Big Data management and decision-making system architecture

can deal with the fast growth of these numerous datasets, Big Data integration is used as a process which is highly combined and iterative to add new data sources. The CIM, which is an abstract information model, can provide data understanding through the identification of the relationships and associations of the data within a utility. In addition, implementing a communication infrastructure is required. This will protect the security and privacy of high volumes of data transfer and can help in managing the interfaces between new analytics platforms and legacy systems (Yan et al., 2013). The challenge is not only to store and manage the vast volume of data but also to analyze and extract meaningful value from it. The resulting data are then ready to move to the next platform: analytical and modelling. In the proposed system, streaming analytics is the platform responsible for management, monitoring, and real-time analytics of SG live streaming data with a high scalability and availability. Analytics platform provides more than packaged application. A platform can be considered the foundation which integrates certain data into the application flow and updates database with processed information which helps in achieving a high level of dynamic feasibility with the proposed system. Fig.7 shows how the analytics platform is approached, starting from receiving the managed data which have different effects on the system based on its type. Modelling and analyzing are performed using real-time analytics through advanced analytics and predictive analytics. Advanced analytics are used to visualize, optimize, or automate the grid operations. In addition, they create new values from the data to be useful in decision-making. Predictive analytics platform analyzes the current data to make predictions about the future and anticipates the outcomes and behaviours based upon the data and not on assumptions. Predictive analytics have a great impact on SG through improving power distribution reliability by achieving continuity of the power and reducing power outages. The final output of analytics platform should be stored in a format/model after being categorized based on its intelligence and importance. Interpretation aggregates the results in time series intervals, displaying data in more sophisticated ways such as infographics, dials and gauges, geographic maps, spark lines, heat maps, and detailed bar, pie, and fever charts. The obtained data characterized in specific patterns can be turning into comprehensible information for decision-making. Decision-making is quite adaptive with dynamical environment where it depends on variable data all the time.

From the previous description, it is evident that the proposed system presents a suitable approach that can be implemented to process Big Data in SG to make better decisions based on calculated prediction, not assumptions. The major attributes and benefits for implementing the proposed system are generating viable values from Big Data through both predictive and advanced analytics in SG. This in turn improves economic dispatch based on optimizing asset life cycle, improves reliability through better management of operating parameters of SG, improves efficiency by reducing maintenance, gives the opportunity to predict any failure, and takes the suitable timely decision. In addition, this helps in increasing customer satisfaction through taking into consideration the customers' complaints and comfort level.

5. Conclusion

Big Data emerges as a promising solution to better decision-making in SG. The need for Big Data management in SG has become a major trend to resolve robustness challenges to extract useful data from this massive amount of data. This paper characterizes different techniques and algorithms used to manage Big Data in SG. It shows that many technical challenges described in this paper must be addressed before choosing the techniques that can be used in each step starting from collecting data, mining, storing, integration, and analysis to achieve the promised benefits of Big Data. A new suggested system is presented in this paper trying to tackle the major data process challenges to develop SG performance. The proposed system is effective to recognize SG data and their characteristics to develop comprehensive data management and reach the optimum decision-making, allow predicting and preventing equipment failures, which can be helpful to reduce unscheduled downtime and reduce equipment cost, improve grid efficiency, and increase the reliability and scalability of the grid.

Acknowledgements

This publication was made possible by National Priorities Research Program (NPRP) grant (NPRP10-0101-170082) from the Qatar National Research Fund (a member of Qatar Foundation) and the co-funding by IBERDROLA QSTP LLC. The findings achieved herein are solely the responsibility of the author(s).

References

Acquisto, G., Domingo-Ferrer, J., Kikiras, P., Torra, V., de Montjoye, Y.A. and Bourka, A. (2015). Privacy by Design in Big Data: An Overview of Privacy Enhancing Technologies in the Era of Big Data Analytics. arXiv preprint arXiv:1512.06000 (2015).

Adiba, M., Castrejon-Castillo, J.-C., Espinosa Oviedo, J. A., Vargas-Solar, G. and Zechinelli-Martini, J. L. Netherlands, (2016). *Big Data Management Challenges, Approaches, Tools and their Limitations*. Networking for Big Data.

Afrati, F.N., Borkar, V., Carey, M., Polyzotis, N., Ullman, J. D. (2011). Map-reduce extensions and recursive queries. In: *Proceedings of the 14th International Conference on Extending Database Technology*, Uppsala, Sweden, 22–24 March 2011, pp. 1–8.

Alexandros, L., Jagadish, H. V. (2012). Challenges and Opportunities with Big Data. *Journal Proceedings of the VLDB Endowment*, 5(12), pp. 2032–2033.

Béjar Alonso, J. (2013). Strategies and Algorithms for Clustering Large Datasets: A Review.

Ben Ayed, A., Ben Halima, M. and Alimi, M. (2014). Survey on clustering methods: towards fuzzy clustering for big data. In: 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), IEEE, Tunis, Tunisia, 11–14 August 2014, pp. 331–336.

Beyer, M. (2011). *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data*. Gartner. Archived from the original on 10 July 2011 [Retrieved 13 July 2011].

Bonomi, F., Milito, R., Zhu, J. and Addepalli, S. (2012). Fog computing and its role in the internet of things. In: *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing; MCC* '12, New York, NY, USA, ACM, 2012, pp. 13–16.

Borkar, V. R., Carey, M. J. and Li, C. (2012a). Big data platforms: what's next? XRDS Crossroads, *the ACM Magazine for Students*, 19(1), pp. 44–49.

Borkar, V., Carey, M. J. and Li, C. (2012b). Inside 'Big Data Management': Ogres, Onions, or Parfaits?

Bredillet, P., Lambert, E. and Schultz, E. (2010). CIM, 61850, COSEM standards used in a model driven integration approach to build the smart grid service oriented architecture. In: *2010 First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Gaithersburg, MD, USA, 4–6 October 2010, IEEE, pp. 467–471.

Cagri Gungor, V., Sahin, D., Kocak, T., Ergut, S., Buccella, C., Cecati, C. and Hancke, G. P. (2013). A Survey on Smart Grid Potential Applications and Communication Requirements. *IEEE Transactions on Industrial Informatics*, 9(1), pp. 28–42.

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *SIGMOD Record*, 39(4), pp. 12–27.

Chandarana, P. and Vijayalakshrni, M. (2014). Big data analytics framework. In: *Proceedings of the International Conference on Circuits, System, Communication and Information Technology*

Applications (CSCITA), Mumbai, 4–5 April 2014, IEEE, pp. 430–434.

Chandarana, P. and Vijayalakslnni, M. (2014). Big data analytics framework. In: *International Conference on Circuits, System, Communication and Information Technology Applications,* Mumbai, India, 4–5 April 2014, IEEE.

Demchenko, Y., Grosso, P., De Laat, C. and Membrey, P. (2013). Addressing big data issues in scientific data infrastructure. In: *Collaboration Technologies and Systems (CTS), 2013 International Conference onx*, San Diego, CA, USA, 20-24 May 2013, IEEE, pp. 48–55.

Electric Power Research Institute (EPRI). (2009). Report to NIST on the Smart Grid Interoperability Standards Roadmap. June 2009.

Elluri, V. R. and Salim, A. (2016). A comparative study of various clustering techniques on big data sets using Apache Mahout. In: *3rd MEC International Conference on Big Data Smart City*, Muscat, Oman, 15–16 March 2016, IEEE.

Fadnavis, R. A. and Tabhane, S. (2015). Big data processing using hadoop. *International Journal of Computer Science and Information Technologies*, 6(l), pp. 443–445.

Fahad, A., Alshatri, N., Tari, Z., ALAmri, A., Zomaya, A. Y., Khalil, I., Sebti, F. and Bouras, A. (2014). LOOKING BACK of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3), pp. 267–279.

Fang, X., Misra, S., Xue, G. and Yang, D. (2012). Smart Grid—The New and Improved Power Grid: A Survey. *IEEE Communications Surveys & Tutorials*, 14(4), pp. 944–980.

Ferreira Cordeiro, R. L., Traina, C. Jr., Traina, A. J. M., López, J., Kang, U. and Faloutsos, C. (2011). Clustering very large multi-dimensional datasets with MapReduce. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Diego, California, USA, 21–24 August 2011,pp. 690–698.

Gungor, V., Lu, B. and Hancke, G. (2010). Opportunities and Challenges of Wireless Sensor Networks in Smart Grid. *IEEE Transactions on Industrial Electronics*, 57(10), pp. 3557–3564.

Hoffmann, L. (2013). Looking back at big data. *Communications of the ACM*, 56(4), pp. 21–23.

Jararweh, Y., Jarrah, M., Alshara, Z., Alsaleh, M. N. and Al-Ayyoub, M. (2014). Cloudexp: A Comprehensive Cloud Computing Experimental Framework. *Simulation Modelling Practice and Theory*, 49, pp. 180–192.

Kersten, M. L., Idreos, S., Manegold, S. and Liarou, E. (2011). The Researcher's Guide to the Data Deluge: 'Querying a Scientific Database in Just a Few Seconds. *Proceedings of the VLDB Endowment*, 4(12), pp.

Kleiner, A., Jordan, M., Ameet, T. and Purnamrita, S. (2012). The big data bootstrap. In: *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland.

Kusum, M. and Rupali, M. (2013). A Review on Various Classification Algorithms for An Incremental Spam Filter. *International Journal of Application or Innovation in Engineering and Management*, 2(11), pp. 325–331.

Mandai, B., Sahoo, R. K. and Sethi, S. (2015). Architecture of efficient word processing using hadoop for big data applications. In: *International Conference on Man and Machine Interfacing*, Bhubaneswar, India, 17–19 December 2015, IEEE.

Markovic, D., Zivkovic, D., Branovic, I., Popovic, R. and Cvetkovic, D. (2013). Smart Power Grid and Cloud Computing. *Renewable and Sustainable Energy Reviews* 24, pp. 566–577.

Mohan, C. (2013). History repeats itself: sensible and NonsenSQL aspects of the NoSQL hoopla. In: *Proceedings of the 16th EDBT International Conference on Extending Database Technology (EDBT'13)*, Genoa, Italy, 18–22 March 2013.

Nagpal, P. B. and Mann, P. A. (2011). Survey of Density Based Clustering Algorithms. *International Journal of Computer Science and its Applications*, 1(1), pp. 313–317.

Oracle Corporation. (2011). Oracle NoSQL Database Compared to MongoDB. White-Paper.

Park, K., Nguyen, M. C. and Won, H. (2015). Web based Collaborative Big Data Analytics on Big Data as a service platform. In: International Conference on Advanced Communication Technology (ICACT), Seoul, South Korea, 1–3 July 2015.

Reinprecht, N., Torres, J. and Maia, M. (2011). IEC CIM architecture for Smart Grid to achieve interoperability. In: 5th Grid Interop Meeting (Grid Interop), Phoenix, USA, 2011.

Rohr, M., Osterloh, A., Gründler, M., Luhmann, T., Stadler, M. and Vogel, N. (2011). Using CIM for Smart Grid ICT Integration. *IBIS,* 11(2011), pp. 45–61.

Sadalage, P. J. and Fowler, M. (2012). *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot*. Upper Saddle: Addison Wesley.

Sagiroglu, S. and Sinang, D. (2013). Big Data: A Review. IEEE, 2013.

Saha, B. and Srivastava, D. (2014). Data Quality: The other face of big data. In: *Proceedings of the 2014 IEEE 30th International Conference on Data Engineering (ICDE)*, Chicago, IL, USA, 31 March–4 April 2014.

Shahrivari, S. (2014). Beyond Batch Processing: Towards Real-Time and Streaming Big Data. *Computers*, 3(4), pp. 117–129.

Sherin, A., Uma, S., Saranya, K. and Vani, S. (2014). Survey On Big Data Mining Platforms, Algorithms and Challenges. Journal of Computer Science & Engineering Technology, 5(9), pp. 854–862.

Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y. and Herawan, T. (2014). Big data clustering: a review. In: *International Conference on Computational Science and Its Applications*. Springer, Cham International Publishing, pp. 707–720. 2014.

Shyam, R., Kumar, S., Poornachandran, P. and Soman, K. P. (2015). Apache Spark a Big Data Analytics Platform for Smart Grid. *Procedia Technology*, 21(2015), pp. 171–178.

Srinivas, B. and Togiti, B. (2015). Analysis of Mining on Big Data, *International Journal of Research and Computational Technology*, 7, pp. 1–10.

Stonebraker, M., Abadi, D. and DeWitt, D. (2010). MapReduce and parallel DBMSs: friends or foes? *Communications of the ACM*, 53(1), pp. 64–71.

Thakur, B. and Mann, M. (2014). Data Mining for Big Data: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(5), pp. 469–473.

Ward, J. S. and Barker, A. (2013). Undefined by Data: A Survey of Big Data Definitions. arXiv preprint arXiv:1309.5821.

Wei, C., Fadlullah, Z. M., Kato, N. and Stojmenovic, I. (2014). On Optimally Reducing Power Loss in Micro-Grids with Power Storage Devices. *IEEE Journal on Selected Areas in Communications*, 32(7), pp. 1361–1370.

Weilki, J. (2013). Implementation of big data concept in organizations – possibilities, impediments and challenges. In: *Proceeding of 2013 Federated Conference on Computer Science and Information Systems*, IEEE, pp. 985–989.

Wu, X., Zhu, X., Wu, G. Q. and Ding, W. (2014). Data Mining with Big Data. *IEEE Transactions in Knowledge and Data Engineering*, 26(1), pp. 97–107.

Xhafa, F., Naranjo, V. and Caballe, S. (2015). A software chain approach to big data stream processing and analytics. In: *International Conference on Complex Intelligent and Software Intensive Systems*, Blumenau, Brazil, 8–10 July 2015, IEEE.

NIST. (2010). Office of the National Coordinator for Smart Grid Interoperability, National Institute of Standard and Technology, U.S. Department of Commerce, "*NIST Framework and Roadmap for Smart Grid Interoperability Standard, Release 1.0*," NIST Special Publication 1108 on the January 2010.

Yadav, C., Wang, S. and Kumar, M. (2013). Algorithm and Approaches to Handle Large Data – A Survey. *International Journal of Computer Science and Network*, 2(3), pp. 2277–5420.

Yan, Y., Qian, Y., Sharif, H. and Tipper, D. (2013). A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements, and Challenges. IEEE Communications Surveys & Tutorials, 15(1), pp. 5–20.

Zerhari, B., Lahcen, A. A. and Mouline, S. (2015). Big data clustering: algorithms and challenge. In: *Proceedings of the International Conference on Big Data, Cloud, and Applications (BDCA'15)*.

Zhang, D. (2013). Inconsistencies in Big Data. In: *Proceeding of IEEE international conference on Cognitive Informatics and Cognitive Computing, IEEE*.

Zikopoulos, P., DeRoos, D., Parasuraman, K., Deutsch, T., Giles, J. and Corrigan, D. (2013). *Harness the Power of Big Data*. McGraw-Hill.