Application of Big Data Analytics in Power Distribution Network

Foroogh Sedighi*, Mohammadreza Jabbarpour, Sheyda Seyedfarshi

Information and Communication Technology Department, Niroo Research Institute,Tehran, Iran
fsedighi@nri.ac.ir, mrjabbarpour@nri.ac.ir, sfarshi@nri.ac

Abstract

Smart grid enhances optimization in generation, distribution and consumption of the electricity by integrating information and communication technologies into the grid. Today, utilities are moving towards smart grid applications, most common one being deployment of smart meters in advanced metering infrastructure, and the first technical challenge they face is the huge volume of data generated from variety of smart devices including the meters. This data is beneficial for both customers and utilities, but only if the capability of using it and extracting knowledge and hidden patterns from data is exploited. In this article, a brief overview of data sources along with applications of big data analytics in power distribution networks and related analytical data models are presented. At the end, big data management tools and techniques applicable in power distribution networks are introduced.

Keywords: Power distribution networks; Big data; data analytics; Cloud computing; Fog computing.
Article history: Received 13-Aug-2018; Revised 20-Sep-2018; Accepted 04-Oct-2018.

1. Introduction

Integrating information and communication technologies into power grid, results in a sensor based network which provides the capability of monitoring the entire system and automation of the processes. These sensors and other smart devices including smart meters generate huge volume of data which has required characteristics (variability, variety and velocity) to be called Big Data. The main challenge in the distribution domain is how to collect and process large data sets generated from smart devices, in order to mine and detect patterns and make decisions in real-time or near real-time. In fact, one of the major factors in successful operation of today's utilities is their capability in accessing, analyzing and managing huge amount of data which might be generated in high speeds. Using result of this analysis helps utilities to improve their customer relationship, operational efficiency along with reducing their costs.

In this paper we provide brief overview of the main data sources in power distribution networks (PDNs) which for the purpose of clarity are divided them into two classes: field data (generated within the distribution grid) and external data (generated outside the distribution grid). Then current trend in application of big data analytics in PDNs and related investments are introduced. As the main part of this paper, important applications of big data analytics in PDN are investigated. Finally, platforms and techniques that are most appropriate to be applied in this domain are introduced.

2. Data Sources in Power Distribution Networks

Installation of new smart devices has resulted in sharp increase of the data volume of power distribution industry. These data can be called big data, since they have all the mentioned features related to big data. These data come from various resources some of which are: Supervisory Control and Data Acquisition (SCADA) system, Energy Management System (EMS), Wide Area Measurement System (WAMS), Outage Management System (OMS), Distribution Management System (DMS), Advanced Metering Infrastructure (AMI), Enterprise Asset Management System (EAMS), power quality monitoring system, Meter Data Management

System (MDMS), Management Information System (MIS), Enterprise Resource Planning (ERP), Geographic Information System (GIS), Weather Forecast System (WFS) [1].

Considering variety of distribution networks' data sources and wide range of devices used to collect them, data fusion and integration are vital aspects in distribution big data analytics. Collected data are usually heterogeneous, independent, and irrelevant. In addition, the data structure, format and quality may vary widely. These complications result in challenges for data fusion and analytics. For these reasons, it is important to identify and classify the types of big data sources in the PDN. Our proposed classification for big data sources in PDN is as follows:

A) Field Data

Field data come from devices that are designed to be used in different parts of electric power network, especially in distribution networks. Field data sources could be classified into 5 categories:

Smart Meters: These devices are believed to be the starting point to move towards smart grid and are the most important device in the customer side. Traditional meters are replaced with smart meters to enhance meter-to-cash operations and provide power-quality measurements including line voltage, current and frequency in predefined time intervals. These data can improve grid troubleshooting, maintenance, load planning, demand response, customer satisfaction, theft identification, and load prediction. Fault detection, revenue protection and operating efficiencies can also be provided using smart meter data analytics.

Sensors: Along with smart meters, sensors are used to collect network data from transformers, voltage detection devices, power lines and demand-side management equipment. Sensors provide a bird view of the grid state along with information regarding overall operating parameters. Most of the sensors are composed of three parts, namely, transducer, CPU and communication module for receiving, processing and sending information. Sensors are likened as the eyes and ears of the utilities in the grid.

Control devices: Bidirectional communication is one of the most primary features of the smart grid. It means devices can send their sensed data and react to the received commands from central system. These devices enable the grid to distribute the load responsively, preserve grid stability to manage complex Distributed Energy Resources (DERs) and response to unexpected grid stability issues. Reaching to the state of self-healing is the main goal of using control devices in the smart grid. The deployment of control devices in different areas can provide significant improvement for PDNs in counteracting the dynamic and continuous changes of the load in the network. The presence of control devices are essential to many processes such as monitoring the grid, automation, regulation of power disturbances, facilitation of remote repairs, and provision of command and control from a centralized management system. Post-processing analytics, which can be reconstructed through sensors and control devices in the grid, enable utilities to identify trends and demands.

Distributed Energy Resources (DERs): The growing penetration of DERs such as renewable energy resources, micro-grids, electric vehicle networks and storages in the smart grid enhances the disturbance possibility. Real-time information, privileged situational intelligence, knowledge about weather conditions, data fusion capability are essential for making quick and accurate decisions regarding frequency control management, power quality and other operational parameters. Successful integration of renewable energy requires that the utilities have access to weather forecasts and take into account wind, cloud cover and other environmental variables in management of these resources. These factors can change instantaneously, and utilities should be able to increase the level of reliability and precision of their predictions in order to adjust demand with capacity.

Intelligent Electronic Devices (IEDs): IEDs serve as grid controllers in the network and can issue control commands based on received information. Tripping circuit breakers based on voltage, current, or frequency irregularities and the ability to function as protective relaying devices including capacitor bank switches, on load-tap changers, reclosers, circuit breakers, and voltage regulators are the common forms of IEDs in PDNs. IEDs function in variety of roles in the grid including protection, control, monitoring and measurement. As IEDs provide extensive and useful information, their data are essential for root-cause and troubleshooting analysis.

B) External Data

External data come from equipment which are not explicitly designed for smart grid, but their data can be collected and used for smart grid applications and analytics.

Customer Side Devices: Internet of Thing (IoT) technology enables any device including electric devices in the customer side, to connect and exchange data. Gathering and modeling these data is a starting point for utilities that seek to increase trust and satisfaction among customers and reduce their business risks by providing new

products and services. Customer data analytics, addresses many problems, including management of interconnections between micro- and nano-grids, performance surveillance, advanced demand response, dynamic pricing and management of electric vehicles. Combining these information with available data sources about customers, including population and housing statistics, behavioral web and social data, and financial records, can provide deep insights regarding customer pattern of electricity usage.

Historical data: Obviously there is a need to preserve data and improve access to them. However, keeping all generated data could result in increased storage and management costs. There are two other factors which should be considered: compliance and privacy. The usefulness of historical data is directly related to how they are aggregated, organized and stored. Creating energy data center under certain and specific rules can solve these problems by providing data access for both customers and utilities.

Third-Party Data: Combining third-party data like weather, demographic information, mobile and GIS data with field data sources could result in enhanced customer classification, demand prediction, and fraud identification. The main concern of third-party data sharing is related to the sharing of customer information that are collected by the utilities, in particular billing and smart meters data that specify the exact amount of customer consumptions.

3. Current Trend

European Union, USA, China, South Korea, Australia, India, Brazil and Japan are among the pioneer countries in the field of smart grid. Based on their projects and investments, smart meters are the first and main big data sources in PDNs. For example, the total investment of European Union in smart grid technologies will reach 56.5 billion euros until 2020 [2] and the number of deployed smart meters will reach to 240 million. Categorizing existing projects can be a way of identifying the future path and the key issues and challenges in the smart grid. Many categories are introduces in this field, but one of the most complete ones is presented by European Union where the projects are classified into six domains, namely Smart Network Management (SNM), Demand-Side Management (DSM), integration of Distributed Generation and Storage (DG&S), Electric mobility (E-mobility), integration of Large-scale Renewable Energy Sources (L_RES) and other [2]. Based on the existing reports, SNM is the most targeted domain in Europe and 34% of investments are allocated to this domain. SNM projects emphasize on Smart grid assets and implementation and development of its functionalities. Reducing the operational and planning costs are the main goals of SNM domain projects. In order to achieve this goal, network monitoring and control equipment such as smart meters and IEDs, should be installed throughout the grid to obtain accurate and continuous data. From the investment perspective, DSM projects are in the second places taking 25% of total investments. This domain focuses on demand response and energy consumption reduction projects. The figure for the sum of investments for SNM and DSM projects means that almost 60% of the investments in smart grid projects is related to big data analytics in distribution system. Nevertheless, some of these projects can be beneficial for other systems including generation and transmission. We will provide a comprehensive overview on big data analytics applications in PDN in Section 5. The total investment per smart grid domain is illustrated in "**Error! Reference source not found.**".

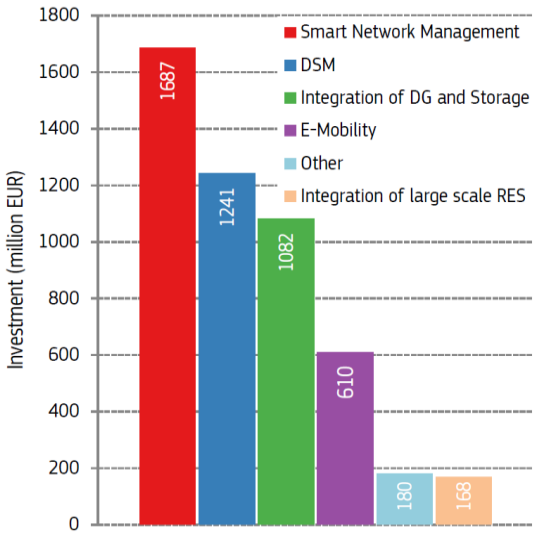

Fig. 1. Total investment per smart grid domain [2]

4. Applications of Big Data Analytics in Power Distribution Networks

This section provides comprehensive overview of big data analytics applications in PDN. These applications are classified into 9 categories. **Error! Reference source not found.** divides them based on two main stakeholders in power system, namely, customers and PDNs. This figure represents that some of the applications are beneficial only for PDNs, but there are some other applications that are useful for both customers and PDNs.

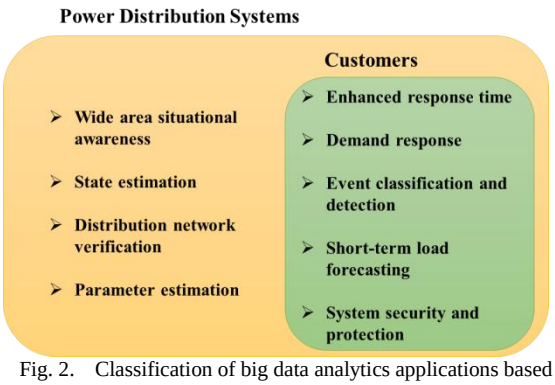

Fig. 2. Classification of big data analytics applications based on stakeholders

Enhanced response time: Today's customers expect accurate and immediate response to any question and problem, and delays could lead to customer dissatisfaction. Utilities take the results from data analysis to fully operationalize the smart metering process from meter installation to billing, redesigning to reduce costs, detecting previously unclear anomalies, reducing downtime (outage management), and automate many exceptions by eliminating manual processes. This type of applications is an immediate win-win for both utilities and their customers.

Offering the right program to the right customer (demand response): The ability to accurately respond to demands or provide the right program for each customer will sharply reduce the power consumption and costs especially an in peak time which is beneficial for both utilities and customers. Applying data analysis results, not only increases the accuracy of planning, but also reduces the costs by avoiding the need for new power plant and substation construction.

Wide area situational awareness: This process has three phases: The first phase, perception, is to understand the heterogeneous data that could be obtained from the SCADA system or embedded equipment, such as IEDs and PMUs. The second phase, comprehension, is to comprehend what the perceived data mean in relation to system oscillation or instability, which requires data processing and storage resources, and the knowledge to extract information. The last phase, projection, is to predict the system future behavior using information from two previous phases. With this process, operators of the control room have enough time to prevent events that may cause problems for system [3].

State estimation (SE): SE means minimizing errors and information defects. With the emergence of smart cities and big data, new algorithms and techniques are being introduced and deployed in this field. There are two major challenges in state estimation deployment, filtering bad data and dimensionality reduction of collected big data. Failure of metering devices, noises, and electromagnetic interferences can produce bad data. SE outcomes are used in power system control centers [3].

Event classification and detection: System faults, transmission line problems, load shedding, generation problems, and electrical fluctuations are some factors that cause disturbance in PDNs. Categorizing these events through big data analytics helps to make appropriate decisions to deal with the disruptions [3].

Short-term load forecasting: Smart grid big data can be used for short-term load prediction. The main technique in this approach is the classification of load patterns by association and clustering analysis based on smart metering data along with historical data and environmental data such as temperature, wind condition, rainfall and humidity. By using accurate spatial and temporal data, complex techniques such as regression tree learning and artificial neural networks can be used to achieve precise predictions [3].

Distribution network verification: Another challenge for power systems is to ensure the correct operation of the distribution network. GIS data are used for this purpose, however, the verification due to low accuracy of this kind of data is not reliable. Therefore, some methods based on big data analytics and processing are proposed to correct operation of distribution networks. Statistical correlation algorithm along with big data is a successful use case to verify the distribution network topology in the smart grid, especially for underground feeders, which are difficult to verify [4].

Parameter estimation (PE) for distribution system: Grid development and utilization planning, and finding solutions to improve the security and performance of the distribution system require extensive and systematic studies. The main stage is the network modeling, which requires precise information on transformers and lines impedance parameters. PE is a process in which one or more network parameters' values are estimated. The accuracy of PE strongly depends on the accuracy of the measurements. Big data generated by metering equipment can increase the accuracy of the measurements, and consequently, the accuracy of PE for distribution network [3].

System security and protection: Cyber-attack is the biggest threat to smart grids due to the connections and interactions between components of power grids and communication networks. Existing solutions suffer from scalability and flexibility issues. Deploying a specification-based hybrid Intrusion Detection System (IDS) by considering multiple data sources in its design, can be a proper solution. Big data analytics can provide

better solutions including big data oriented cryptosystems, big data oriented anomaly detection, and big data oriented intelligent applications for comprehensive protection of power systems and customer privacy [3].

5. Big Data Analytics in Power Distribution Networks

A) Analytical Data Models

Models are the main component of advanced analytics. The purpose of analytics is to extract, interpret and communicate meaningful patterns hidden in data using different algorithms and statistics. Purpose of performing data analytics in power distribution networks is to transform it into an intelligent, efficient and gainful network.

In a PDN, analytics can be divided into different classes such as signal analytics, event analytics, state analytics, engineering operations analytics and customer analytics [5]. To build a valuable model for power distribution networks, it is required to select appropriate data sources, algorithms, variables and techniques in order to come up with a solution for business challenges. It is also required that the analyst have enough domain knowledge of power systems [6]. In PDNs, smart devices such as smart meters and sensors can all be connected through digital platforms that exploit advanced analytics to overcome challenges in real-time. There are four main analytical models: descriptive, diagnostics, predictive and prescriptive. The purpose of different types of analytics along with an example of their application in PDNs is depicted in "**Error! Reference source not found.**" [6, 7].

B) Big Data Management

As provided in "**Error! Reference source not found.**", there are variety of tools and techniques proposed by Big Data technologies. To reach the desired result, utilities should deploy suitable platforms and tools according to their needs. In this section some of the platforms and techniques to manage huge volume of data in power network are introduced.

Platforms: Generated data in power networks are usually stored in data centers which could be based on cloud. According to requirements, two approaches called cloud and fog computing can be applied. Many challenges related to management of big data in power systems can be solved using cloud computing platform. It provides flexibility, agility and efficiency in terms of saving cost, energy and resources for utilities. Cloud computing provides on-demand access to remote shared resources and data and is based on a service driven model which can be offered as private, public and hybrid [5]. As stated in [3], there are variety of industrial cloud platforms that can be exploited in PDNs including Microsoft Azure, Holm, Google PowerMeter, InterPSS, Smart-Frame and etc. Sample of cloud computing platform is displayed in "**Error! Reference source not found.**". Fog computing which is an extension of cloud computing provides local computation, storage and networking services. In other words, it gathers data from end users (for example sensors of PDN) and stores and computes them locally before transferring them to the cloud that is, after performing the computation processes, data will be transferred to cloud for storage.

Table.1.
Analytical Models

Analytic Approach	Purpose	Example
Descriptive	To provide information about what happened and it contains the initial step in detecting beneficial information/data for further processing.	Collect data about customers participated in demand-response programs and gain an initial understanding of customers' behavior.
Diagnostics	To figure out the cause of events and behavior of the system and to detect challenges and opportunities.	Determine the cause of certain behaviors from customers and the basis of their decision making in terms of their participation in demand-response programs.
Predictive	To determine what might happen in the future by making probabilistic predictions to detect trends.	To predict the customer behavior in certain conditions with placement of smart meters in their homes (customer segmentation).
Prescriptive	To detect the most appropriate outcome of the events by using parameters of the system and plan strategies to deal with same events that might happen in the future.	Achieve success by understanding the best next steps according to knowledge gained from customers who participates in demand-response programs and present appropriate engagement strategies.

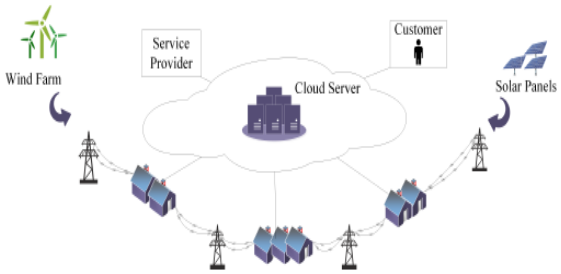

Fig. 3. Cloud Computing Platform [9]

It acts just like a bridge between the power network and cloud. Thus, in some cases that low latency, efficient privacy and locality become issue in communication and computation processes, fog computing is considered an alternative to cloud computing [9]. Example of Fog computing platform is shown in **Error! Reference source not found.**.

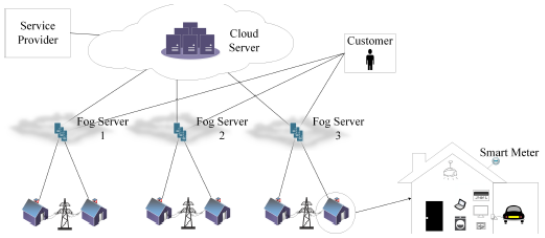

Fig. 4. Fog Computing Platform [9]

Table.2.
Big data management in power distribution networks

Big Data Management in Power Distribution Networks			
Platforms		***Techniques***	
Cloud Computing	***Fog Computing***	***Batch Processing***	***Stream Processing***
On-demand access to computing resources	Low latency, highly distributed model	Suitable for static and non-real time applications	Suitable for both real time and non-real time applications
No maintenance Cost	No need of data transfer to cloud	Analyzes historical data	Ideal for sensors and big data streams
Service driven model	Data is processed by devices at the edge of network	Splits data sets into smaller sets and process the concurrently on multiple machines	Highly scalable and fault tolerant

Techniques: Two main processing techniques as provided in ***Error! Reference source not found.*** are batch and stream processing which can be deployed in big data management and analysis in power networks. In real world combination of these two; hybrid processing is in use.

Batch processing is an approach for processing huge amount of data over a period of time. In this type of processing, batch results are produced after collecting, entering and processing data. Apache Hadoop is a suitable choice for batch analytics in PDNs. As shown in "**Error! Reference source not found.**", it includes Hadoop Distributed File System (HDFS) which provides high-throughput access to application data, a resource scheduler (Hadoop YARN) and Hadoop Map-Reduce paradigm. Hadoop Map-Reduce which can be deployed for parallel processing of large data sets is mainly used for static and historical data [11]. In this approach, large data sets are divided into smaller sets and process is done on each unit in parallel. In PDNs, Map-reduce can be applied in extracting customer behavior analysis, modeling the measured energy savings and etc. using static data. Thus this approach is not appropriate for real-time analytics process [8].

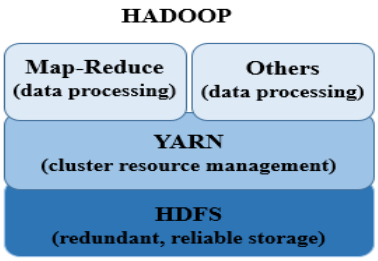

Fig. 5. Apache Hadoop Framework [12]

For real-time applications, stream processing is used. In contrast with batch processing, a continual input, process and output of data is involved in stream or real-time processing and data should be processed in real-time or near-real time [10]. Comparing to batch processing, stream processing provides more accurate and timely results. This becomes very important in a network like PDN, where data is generating continuously from variety of data sources [8].

To perform on-line and streaming data analysis in power networks, Apache Spark [8] can be used. As in "**Error! Reference source not found.**", Apache Spark framework includes Structured Query Language (SQL), Spark streaming, Mllib for machine learning and GraphX.

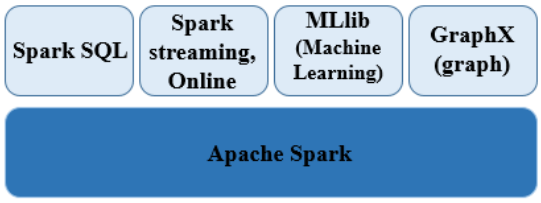

Fig. 6. Apache Spark Framework [8]

Apache Spark is an open source framework used for big data computing. As stated in [13], Spark runs programs up to 100x faster than Hadoop

MapReduce in memory, or 10x faster on disk. In addition to Apache Spark, other solutions such as S4, Splunk, and Storm can be used for real time processing in PDNs.

In PDNs, in order to monitor the stability of the system, PMU data can be used. In order to achieve best results, this process should be done with no latency and real-time. Except online monitoring of power system, it also prevents unwanted events in the system such as blackouts. Other benefits of stream processing for PDNs include real-time pricing, real-time fraud detection, and so on.

6. Conclusion

Big data analytics is one the most trending concepts in power distribution networks. In this paper, different aspects of big data analytics in PDNs have been surveyed. To investigate different data sources in power distribution networks we divided them into two groups of field data and external data. Field data sources include smart meters, sensors, control devices and Intelligent Electronic Devices (IEDs). Distributed Energy Resources (DERs), consumer side devices, historical data and third-party data are grouped into external data sources. We also presented an overview of different applications of big data analytics in PDNs. Finally, we went through different platforms and techniques that best suits the power distribution domain.

According to findings of the project, smart meters are the main big data sources in power distribution networks and analytics of collected data from smart devices brings opportunities for utilities to enhance customer satisfaction, reliability, operational efficiency and safety. To achieve this, variety of platforms and techniques can be used. Nowadays in many utilities, collected data from different part of the network are stored in cloud based data centres. As discussed, when low latency and security become issues in communication and computational requirements, fog computing can be used as an alternative for cloud computing. Fog computing uses the same resources and attributes with the cloud computing. The additional capability of fog computing is associated with locality, proximity to end users, privacy, latency and geo-distribution [9]. In terms of techniques, according to findings of the project and the need in power distribution networks to address the complexity of the analytic and data processing requirements, cluster computing platforms such as Apache Spark are the best solution for this environment comparing to approaches like Map-Reduce. A combination of batch and stream data processing techniques can fulfill the requirements of data analytics in a power distribution network. Investigating the current condition of big data analytics in power distribution networks and feasibility of applying existing applications in the power distribution networks of Iran can be considered as future works.

References

[1] D. Cai, H. Tian, Y. Wang, H. Wang, H. Zheng, K. Cao, et al., "Electric Power Big Data and Its Applications," in International Conference on Enery, Power and Electrical Engineering (EPEE), 2016.

[2] F. Gangale, J. Vasiljevska, C. F. Covrig, A. Mengolini, and G. Fulli, "Smart grid projects outlook 2017," Joint Research Centre of the European Commission: Petten, The Netherlands, 2017.

[3] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid–A review," Renewable and Sustainable Energy Reviews, Vol. 79, 2017.

[4] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart meter data analytics for distribution network connectivity verification," IEEE Transactions on Smart Grid, Vol. 6, 2015.

[5] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, and A. Dahbi, "Big Data management in smart grid: concepts, requirements and implementation," Journal of Big Data, Vol. 4, 2017.

[6] C. L. Stimmel, Big data analytics strategies for the smart grid: Auerbach Publications, 2016.

[7] I. S. G. Big. (2018). Big Data Analytics in the Smart Grid. Available: https://smartgrid.ieee.org/images/files/pdf/big_data_analytics_white_paper.pdf

[8] R. Shyam, B. G. HB, S. Kumar, P. Poornachandran, and K. Soman, "Apache spark a big data analytics platform for smart grid," Procedia Technology, vol. 21, 2015.

[9] F. Y. Okay and S. Ozdemir, "A fog computing based smart grid model," in Networks, Computers and Communications (ISNCC), International Symposium on, 2016.

[10] M. Walker. (2013). Batch vs. Real Time Data Processing. Available: https://www.datasciencecentral.com/profiles/blogs/batch-vs-real-time-data-processing

[11] A. Hadoop. (2014). What Is Apache Hadoop? Available: http://hadoop.apache.org/

[12] S. P. Bappalige. (2014). An introduction to Apache Hadoop for big data. Available: https://opensource.com/life/14/8/intro-apache-hadoop-big-data

[13] Apache Spark. Available: http://spark.apache.org