Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 21 (2015) 171 – 178

Procedia Technology

SMART GRID Technologies, August 6-8, 2015

Apache Spark a Big Data Analytics Platform for Smart Grid

Shyam R[a*], Bharathi Ganesh HB[a], Sachin Kumar S[a], Prabaharan Poornachandran[b], Soman K P[a]

[a]*Centre for Excellence in Computational Engineering and Networking, Amrita Vishwa Vidyapeetha,, Coimbatore – 641112, India*
[b]*Amrita Center for Cyber Security Systems and Networks, Amrita Vishwa Vidyapeetham, Kollam, India*

Abstract

Smart grid is a complete automation system, where large pool of sensors is embedded in the existing power grids system for controlling and monitoring it by utilizing modern information technologies. The data collected from these sensors are huge and have all the characteristics to be called as Big Data. The Smart-grid can be made more intelligent by processing and deriving new information from these data in real time. This paper presents Apache spark as a unified cluster computing platform which is suitable for storing and performing Big Data analytics on smart grid data for applications like automatic demand response and real time pricing.

Peer-review under responsibility of Amrita School of Engineering, Amrita Vishwa Vidyapeetham University

Keywords: Smart Grid; Data Analytics; Big Data; Apache Spark;

1. Introduction

With the power of emerging technologies in information and communication systems, the present day power grid scenario is evolving into a sensor-embedded network (Smart Grid), which has the ability to control and automate the entire processes. Sensors of various types are deployed across the length and breadth of the smart grid. All these sensors produce different types of data (Heterogeneous), which are then collected at the utility data-centers. The amount of data collected from all the different connected components and sensors in a very short time interval is huge. Therefore, it is out of scope of usual storing techniques and computing facilities which are available. The variability, variety and velocity of the data make it into a very special category called “Big Data”. The data collected can be analyzed to reveal the knowledge of unseen patterns which are hidden in large datasets and utilized for

*Corresponding author: Tel.: 8098007545;
E-*mail address:* shyam.neezhoor@gmail.com

Peer-review under responsibility of Amrita School of Engineering, Amrita Vishwa Vidyapeetham University
doi:10.1016/j.protcy.2015.10.085

making strategic, tactical and operational level decisions. In data science, the term data analysis, data mining and textmining refers to the same technique of deriving hidden information using various machine learning algorithms from the data acquired. Among the data mining methods the most widely used are:

- Pattern Matching and Associative rule: It involves learning the frequently occurring trends in the data to definerules for future decision making. Pattern matching algorithms includes Eclat, FP-Tree, so on.
- Classification: A supervised machine learning method in which the data is divided into training and testingsets. Then a classifier model is trained using training set in-order to predict the class labels for the given test data. Some of the most widely used classification algorithms are Decision tree, K-Nearest Neighbour (K-NN), Support Vector Machine (SVM) and Naive-Bayes.
- Clustering: It is an unsupervised learning process. The goal of clustering is to group data points in the datasettogether into a number of groups, depending upon its distribution in higher dimensional space. The choicefor number of clustersdepends on data and problem definition. Some of the most commonly used clusteringalgorithms are k-means, Expectation Maximization and Hierarchical clustering.
- Regression: It is process of identifying the relationship model between independent and dependent variables using the given data. The model is then used to predict the forthcoming values for the upcoming independentvariable. It is a supervised machine learning method. Some of the widely used regression algorithms includeLogistic Regression, Support Vector Regression (SVR) and Gauss-Newton algorithm.

This paper details about the role of Apache Spark in Big Data analytics and it is organized into three sections. Section 2 describes about various sources of data in power grid and asurvey on data analysis techniques that can be performed on smart grid data. The Big Data processing techniques that can be applied to smart grid data are explained in section 3. A real case smart grid application using Apache Spark is detailed in section 4. Section 5 suggests a unified Big Data processing platform suitable for smart-grid applications

2. Smart Grid Data Flow

Apart from the electrical aspect, Smart gird is now becoming an interesting research area for Data scientists. Smart grid data are broadly classified into Generation data, Transmission/Distribution data and Consumer data. Data-centric sensor network view and Distribution of data in smart grid is visualized in the Fig.1. A lot of data analysis can be done over this data to make the Grid more intelligent and smart.

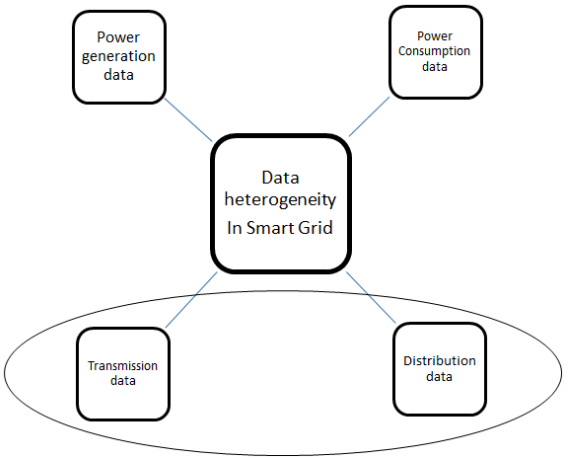

Fig. 1. Heterogeneity in Smart Grid data

2.1. Power generation oriented data

In power plants, electricity is generated from different sources like water, coal, tides, wind, nuclear etc. Swartz et al. [1] proposed a wireless sensors and actor networks which can be employed seamlessly over wind farms to extract the data about the dynamic state behavior of the wind turbines. A coal-based power plant fault analysis and diagnosis system using Association rule mining is mentioned in Li et al.[2]. Foreseeing the predicted load paves the way for power plants, to plan for their future needs. This helps for utility industries to save millions of dollars a year. A number of works based on different types of load forecasting exists in literature. Multiple Linear Regression was used by Hong[3]. Short-term load forecasting based on artificial neural network was proposed by Zhang et al. [4].

2.2. Power Transmission Distribution data

Once the power has been produced, it is feed to step-up transformer and intensified into high voltage, and is transmitted to multiple substations. At each substation, the high voltage electricity is reconstructed to a low state, which is best suited for real time consumption. This is done with a step-down transformer. Further it is distributed to the consumers for utilization. In General, distributed control system (DCS) and supervisory control and data acquisition (SCADA) are propriety control and monitoring systems used in power transmission and distribution. Kaplan etal. [6] provides a list of such monitoring sub systems, which can be used for monitoring the grid activities. In recent years a number of grid monitoring technologies have been developed. Some of them are listed below.

- Grid monitoring system: Phasor measurement unit (PMU) enabled with GPS (global positioning system) measures the spontaneous magnitude of voltage and current from selected grid locations. It is then transmitted to the servers with timestamps. This opens up a global as well as dynamic perspective of the power system.
- Backscatter radio: contributes improved data. Cautions about of transmission and distribution component failing.

The data collected from these sources can be utilized to analyze power system state estimation, a real-time stability determination and commence significant sequence of action. Power system state estimation is used to ensure the stableness of the grid and prevent blackouts. Likewise Chen et al. [5].discuss calculation of power system state prediction via weighted least-square method. When some part of the network are detached from the grid Islanding will occurs and such events can leads to stability issues in grid. In [7] Naive-Bayes classifier is used to for islanding detection. A real time power quality assessment is much needed for power systems in which power disturbances like harmonics, swell and sag can affect the performance. Hongke and Linhai [8]developed a data analysis framework for power quality, using SQL Server and OLAP, an Online Analytical Processing [9]. In power systems grid fault identification, failure cause identification is other major issues. Mori[11] made a survey over the smart power grid oriented papers, which deals with various applications over data mining on the power systems, like failureclassification; examine the transient faults, etc.

2.3. Power Consumption data

The distributed electricity will be consumed by consumers from various zones like Residential (Individual houses and Apartments), Commercial (e.g., Insurance), Industrial (e.g., Factories), Transportation (Railways), Emergency services (e.g., hospitals) and governmental services (e.g., school), etc. Smart meters are equipped at customer end points, which sense and broadcast utilization data to the service providers at regular interval of period. These data are of two types, either in disaggregated data (break up data for every single component or group of numerous components in the single oriented electrical circuits) or in aggregated data (collective data of all appliances). These data are first aggregated at data concentrator. It is then transmitted to central servers. . Advanced metering techniques and IP-based smart meters and appliances enable the data flow in smart grid in more fast and efficient way.

Load forecasting for huge economic and industrial consumers performs a vital part in optimizing electricity consumption for their future development. Edwards et al. [10] compares the efficiency of popular data mining methods for prediction of load in large scale industries and apartments. Customer profiling is helpful in their

behavior prediction and in providing dynamic pricing that meets their requirements. A list of consumer based applications which perform the analysis of utilization data has been described byZeyarAung [12]. Some of them are Real-time pricing, Load control, Metering information and energy analysis via website, outage detection and notification, metering aggregation for multiple sites and facilities, unification customer-owned generation, theft control [12].

Various kinds of individual applications listed in [14], which could be performed by utilizing the data generated from power grid, by applying various Machine Learning techniques, were discussed above. As the data becomes 'Big data', the storage as well as the processing becomes crucial issue. This demands a unified frame-work, which can handle both data and the processing capabilities required for smart-grid data analysis

3. Big Data Analytics on Smart Grid

Performance of Big Data analytics and machine learning upon sensor and operational data from power grid generation and utilization systems should be precise. In order to achieve this, the system requires the flexibility to evaluate the collection of all the power system and consumption data in near real time. When we aggregate all consumer and operational grid systems data from systems like PMU's (Phasor Monitoring Units), Wide Area Measurements(WAM), Advanced Metering Infrastructure systems, IP-based appliances, etc. the dataset tend to easily attain a Peta-byte scale range. Then these datasets will grow in size of hundreds of gigabytes per day. The Big Data analytic platforms are designed with huge power and flexibility to meet all such requirements. The main types of processing techniques employed in Big Data analysis are batch, stream and iterative processing. Thus we require such a platform to store this vast distributed data and to perform all these types of analysis.

3.1. Batch Processing

Initially, Apache Hadoop is utilized as the Big Data Analysis platform for smart grid data. Apache Hadoop is an open source Big Data framework maintained by the Apache Software Foundation (ASF) for processing Big Data. Architecture of Apache Hadoopfocuses on Map-Reduce concept initiated by Google [16]. This includes the original Hadoop Map-Reduce paradigm, Hadoop Distributed File System (HDFS) and a resource scheduler (Hadoop YARN)[17]. Hadoop Map-Reduce is a batch processing programming model which is primarily used for the analysis of large pool of static and empirical data.

In this model, a very large dataset is divided into numerous small sets for processing. Computation is done parallel on all these tiny units of data. For example, A Map-Reduce job can be used to obtain the minimal readings amongst all smart meters. Each instance of the "map" task takes a small set of the smart meters readings and finds the minimum amongst them. These smaller results are "reduced" to an overall minimum value by the reducer job. Map-Reduce can be applied to compute customer usage analysis, energy savings measure modeling and other such analytics which are done over static data. Hence Map-Reduce are used specifically for periodic batch processing but inappropriate for frequent reprocessing of enormous dynamic datasets. It cannot be used for real-time sensor data and streaming data processing. Hence it is not suited for many smart grid real-time analytic processes such as demand response, short-term load forecasting, AMI (Advanced Metering Infrastructure) operations, real time usage and pricing analysis, real-time customer analytics, on-line grid control and monitoring, etc.

Apart from the traditional batch processing technique (Map-Reduce), inability to perform on-line and streaming data analysis is a major drawback for Apache Hadoop [15]. The set of machine learning algorithms provided by Apache Hadoop is also not enough to meet the requirements of Smart grid data analysis due to which, Apache Hadoop is not an apt choice for Big Data analytics on smart grid systems.

Apache Spark [13] is another general purpose cluster computing platform, which delivers flexibility, scalability and speed to meet the challenges of Big Data in smart grid. Other than the usual batch processing, Apache spark has the ability to perform iterative and streaming process. Apache spark has more efficient set of machine learning Algorithms and enhanced linear algebra Libraries. Fig.2 shows the Apache Spark stack with its components.

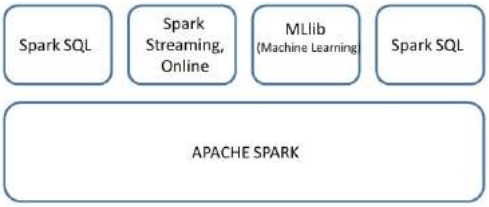

Fig. 2. Apache Spark framework

3.2. Stream Processing

Stream processing involves calling dependent logic on each new data instance, rather than waiting for the next batchof data and then reprocessing everything. In Apache Spark, more analytics are carried using stream processing thanbatch processing. This avoids unnecessary repetition of reprocessing the data. Stream processing provides timely andmore accurate results when compared to batch processing. This model enables Apache Spark to perform analyticson the data with dynamic behavior. This becomes significant when the data comes continuously in real time fromnumerous data sources.

For example, in order to monitor and predict the stability of the entire power grid, using the PMU data, we needto calculate and keep track of stability index of grid. The stability index has to be calculated without any latency,from the data available from PMU's, which are installed across the grid in real time. This can protect the grid frompotential threats like Islanding, blackouts and so on. This makes stream processing very useful smart grid applicationslike Real-time pricing, Real-time theft identification and Power grid cyber security related problems.

3.3. Iterative Processing

Apart from batch and stream processing methods there are many Big Data analytic problems which will not come under the scope of these techniques. Iterative processing is another category of processing methodologies used to solve such kind of problems. Main characteristic of Iterative processing is processing of all variety of data types frequently. In general, Iterative processing is time consuming, due to their frequent read and write operations (each iteration) which involves more I/O transfer. Apache Spark is the current leading framework used for iterative processing. It has the power to process and hold data in memory across the cluster. Once the data is loaded into the framework, the data is written back, only after completion of all iterative process. This ensures Apache Spark 10x to 100x faster than Map Reduce framework, which involves reading and writing data from the disk during each iteration.

4. Apache Spark in Action

In power transmission and distribution, propriety systems like DCS (distributed control system), SCADA (Supervisory control and data acquisition) are used for the control and monitoring purposes. The power systems are verydynamic in nature and disturbances can occur within few milliseconds. The traditional monitoring and measurementinfrastructure has the ability to sense data once in 4-6 seconds without any time synchronization. Real time processing using SCADA data may leads to wrong estimates.

A system which can capture the real time grid information at much faster rate is need to be analyzed for powersystem dynamics and to take control actions. The advent of Synchrophasors enables the collection of data with accurate timestamps, in a rapid manner. PMU(Phasor Measurement Unit) is a device that produces synchronized phasor,frequency and rate of change of frequency estimates from the voltage and current signals with the GPS timestamps.

PMU is mainly used to measure the voltage and current phasors, voltage and current sequence components, frequency,rate of change of frequency and circuit breaker status. The number of data frames transmitted from PMU's varies from10–60 frames per second which is much faster than the default data rate by SCADA systems and deployed acrossentire transmission and distribution centers. The arrival of PMU's results in collection of huge amount of data atvery high speed which are accumulated at Phasor Data concentrators(PDC), where it is gets stored and analyzed byEMS (Energy management system). Tiny latency between the data collection and processing leads to wrong estimateswhich ensure the need of platform for the data to be processed much quicker.

This time synchronized data from all PMUs gives an accurate view of entire power systems. Big Data analysis canbe performed on this collected data to derive information about the state of the system, stability of the grid, etc. Real-time processing of PMU data helps in identifying the existences of any kind of disturbances in the grid and furthertimely actions can be initiated. Apache spark can be utilized effectively for processing the PMU data for variousapplications.

Apache spark paves the way to process the PMU data without much latency to give an accurate view ofthe grid. Fig.3 lists out some important applications of Synchrophasor which can be done using the Apache Sparkframe work. The PDCs are arranged in a hierarchical structure, to collect phasor data, discrete event data from PMUsand other PDCs. So PDCs in each stage can perform a particular time duration analysis suitable for that level.

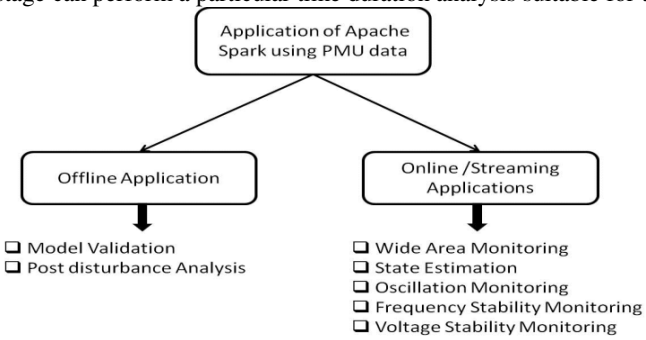

Fig. 3. Applications of Apache Spark using Synchrophasor technology

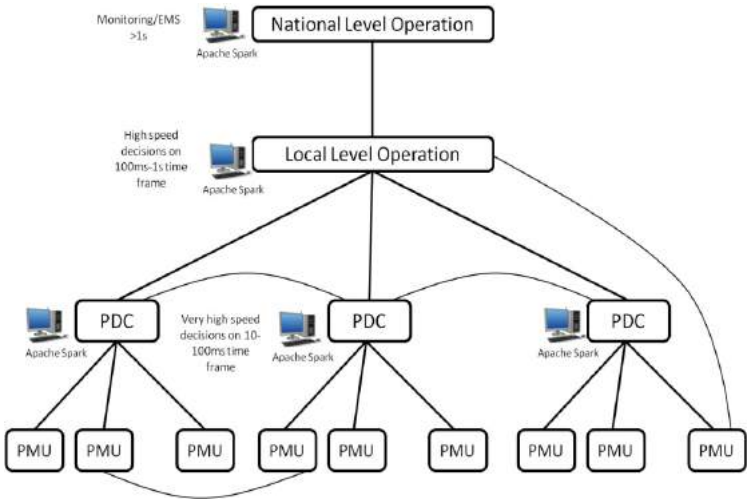

Fig. 4. PDUs embedded with Apache Spark at different levels

Apache Spark can be in-cooperated on these PDC systems for efficient real time data analytics. Fig.4 represents the operational level hierarchies of PDCs achieved using Apache Spark. Power grid can be better monitored by facilitating better control. This can be achieved using Apache Spark.The computational requirement of solving optimizationproblems [18] arises in smart grid networks can also be solved in Apache Spark. Hence it ensures solution to the problemslike optimal state estimation and optimal power flow.

3.4. An experimental setup using Apache Spark

Four systems each having specifications (16 GB RAM,2 TB hard disk, I7-4790K processor, MSI Z87-GD65) were usedto set up an Apache Spark cluster over 1Gbps Ethernet network. Fig.5 shows the set up done in the laboratory.

Fig. 5. Experimented Spark Cluster

An hourly based PMU data from different locations, taken from Texas Synchrophasor Network is available. A streaming analysis on this time-series data is performed using Apache spark cluster.Fig.6 shows the architecture of experimented Spark system. The Apache Cassandra database is used to store the data.

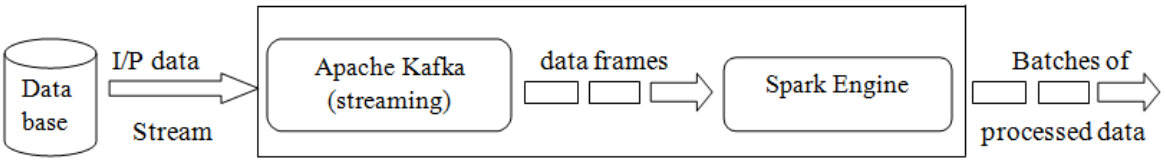

Fig. 6. Architecture of Spark Stream

It one of best NoSQlBig Data database suited for time-series data. Apache Kafka is the used to provide a windowed streaming data to Apache sparkengine. Computation is done parallel across the cluster machines. Experiments are done on using a Synchrophasorstream splitter which splits the same data into multiple times (as we need), to test the stability and performance ofApache spark.

5. Conclusion

Apache spark is emerging as the cluster computing platform for smart grid Big Data applications. Present day smart grid utility demands a wide range of Big Data processing methodologies. Simplistic data processing patterns like Map-Reduce are insufficient to address the complexity of the analytic and data processing requirements of the smart grid. A data analytics software platform for the smart grid should be capable of analyzing both slowly and rapidly changing data. This can be done by incorporating the combination of batch and real-time data processing techniques. It also requires utilization of machine learning algorithms on dataset that are characterized by large, dynamic, and rapidly expanding nature. These requirements necessitate a framework, which have the power of Map-Reduce, Stream, and Iterative processing. Apache Spark is the integrated platform that seamlessly combines batch, real-time and iterative data processing requirements. It delivers advanced analytics and machine learning methodologies for the power grid. This can be effectively used for applications like real time price forecasting, automatic demand response systems, peak time loadbalancing, fault-identification and on-line grid monitoring

References

[1]. Swartz R.A, Lynch J.P, Zerbst S, Sweetman B, Rolfes R. Structural monitoring of wind turbines using wireless sensor networks. Smart Structures and Systems 6, 114, 2010.
[2]. Li J q, Wang S l, Niu C l, Liu J z. Research and application of data mining technique in power plant. In Proceedings of the 2008 InternationalSymposium on Computational Intelligence and Design (ISCID), vol. 2, p. 250253, 2008.
[3]. Hong T. Short term electric load forecasting. Ph.D. thesis, North Carolina State University, USA, 2010.
[4]. Zhang H T, Xu F Y, Zhou L. Artificial neural network for load forecasting in smart grid. In Proceedings of the 2010 International Conferenceon Machine Learning and Cybernetics (ICMLC), vol. 6, p. 32003205, 2010.
[5]. Chen Y, Huang Z, Liu Y, Rice M J, Jin S. Computational challenges for power system operation. Proceedings of the 2012 Hawaii InternationalConference on System Sciences (HICSS), p. 21412150, 2012.
[6]. Kaplan S M, Sissine F, Abel A, Wellinghoff J, Kelly S G, Hoecker J J. Smart Grid: Modernizing Electric Power Transmission and Distribution;Energy Independence, Storage and Security; Energy Independence and Security Act of 2007 (EISA); Improving Electrical Grid Efficiency,Communication, Reliability, and Resiliency; Integrating New and Renewable Energy Sources. TheCapitol.Net, Inc., 2009.
[7]. Najy W, Zeineldin H, Alaboudy A K,Woon W L. A Bayesian passive islanding detection method for inverter-based distributed generation usingESPRIT, IEEE Transactions on Power Delivery 26, 26872696, 2011.
[8]. Hongke H, Linhai Q. Application and research of multidimensional data analysis in power quality. In Proceedings of the 2010 International Conference on Computer Design and Applications (ICCDA), vol. 1, p. 390393, 2010.
[9]. Hart D G. Using AMI to realize the Smart Grid. In: Proceedings of the Conference on Power and Energy Society General Meeting – Conversion and Delivery of Electrical Energy in the 21st Century, pp. 2024, and 2008.
[10]. Edwards R E, New J, Parker L E. Predicting future hourly residential electrical consumption: A machine learning case study. Energy and Buildings 49, 591-603, 2012.
[11]. Mori H. State-of-the-art overview on data mining in power systems. In: Proceedings of the 2006 IEEE PES Power Systems Conference and Exposition (PSCE), p. 3334, 2006.
[12]. ZeyarAung. Database Systems for the Smart Grid. In Smart Grids, p. 151-168. Springer London, 2013.
[13]. ZahariaMatei, MosharafChowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: cluster computing with working sets. InProceedings of the 2nd USENIX conference on Hot topics in cloud computing, p. 10-10. 2010.
[14]. Carol L Stimmel. Big Data Analytics Strategies for the Smart Grid. CRC Press, 2014.
[15]. Vijay Agneeswaran. Big Data Analytics BeyondHadoop, 2014.
[16]. Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters, Communications of the ACM 51, no. 1 :p. 107-113, 2008.
[17]. Sanjay Ghemawat, Howard Gobioff and Shun-Tak Leung. The Google file system. In ACM SIGOPS operating systems review, vol. 37, no. 5,p. 29-43. ACM, 2003.
[18]. Lanchao Liu and Zhu Han. Multi-block ADMM for Big Data optimization in smart grid. In International Conference on Computing, Networkingand Communications (ICNC), p. 556-561. IEEE, 2015.