Big Data Applications in Electric Power Distribution

Swargam, Prashanth
Indiana University Bloomington
107 S Indiana Ave
Bloomington, Indiana 47408
pswargam@iu.edu

ABSTRACT

Now-a-days, the process of storing the power measurements have changed. Conventional meters are replaced by the smart meters. New distribution management systems like SCADA and AMI are implemented to monitor power distribution. These smart meters record the readings and communicate the data to the server. However, these systems are designed to generate the readings very frequently i.e., 15 minutes to an hour. Upon that, smart meters are being deployed at every possible location to improve the accuracy of the data. This advancements in electric power distribution system results in enormous amounts of data which requires advance analytics to process, analyse and store data. This paper discusses about the implementation of Big Data technologies, challenges of implementing Big Data in Electric Power Distribution Systems.

KEYWORDS

Big Data, Power Distribution,Smart Power

1 INTRODUCTION

Volume of data is increasing. According to forbes, it is said that, world's data utilization will increase to 44 zettabytes from the current utilization of 4.4 zettabytes[12]. To process this data, Big Data analytics will be useful. But, instantiating a big data architecture is not easy task.

In electrical Power Distribution industry, data deluge is picking its pace. The data which was recorded for month, is now being noted for very small intervals. This quadruples the amount of data that should be process. There is a lot of potential work to be put in for designing a good Big Data architecture to process and analyse this data. Most of the power generation units are developing their infrastructure to support these designs.

1.1 4 v's in Big Data in Power Distribution System

Big Data is mostly described in 4 v's. Each of this V's are considerable factors in a Big Data Solution.[3]

Volume: The data is periodically generated by many data sources like smart meters, machines and other appliances.

Variety: Each data source in electric power distribution system is explicit to each other. Each source has its own frequency of data generation and its own method of data generation. Thus, the data is heterogenous.

Velocity: is the speed at which the data is available for the end user.

Veracity: It deals with the correctness of the data. As all the data collected by sensors, meter tend to have various losses, correction algorithms should be defined to find the accurate data. Their might be chances for data transfer losses.

2 DATA SOURCES

Smart meters which are placed at customer's vicinity will record the consumption of a specific group of customers. This data can be used to analyse the behaviour of customer for certain circumstances of weather and environment.

Distribution systems which manage the distribution of power, generate large amount of data related to voltages and currents at various levels of distribution. This data is very important in analysing the load level and demand for the distribution circle.[2]

Phasor measuring units at generation. This data is used to analyse the behaviour of generator and amount of power generation that will be required to supply enough power. This data will be used to decide the functioning of generators.[9]

Old market data will be used to analyse the pricing and marketing strategies. These data is more focused on users and their behaviour.

3 DATA INTEGRATION

3.1 Service Oriented Model

This model has a workflow which is defined in Business process Enterprise Language often referred as WS BPEL[6]. WS BPEL is used for is enterprise language used for automating a business process. BPEL files defines the process to be followed by a request from the web services. In this model, All the user requests are handled by services. These services either connect to the storage resources or calls the other services based on the process model defined in BPEL. This modelling ensures data is being utilised in a structural manner and analysed according to the process model.

Interfacing services: This service is used to manage the interfaces with the end user. This services generally initiates calls to a process defined in WS BPEL. After all the other processes which are defined in process model are completed, this service is used to project the analytical data to the user at the end of execution. In this case, this service receives data from one of the process models.[7]

Execution Service: This service is responsible for all the logic involved in modelling the data. For the common requests, these are well documented in BPEL files. These documents specify the set of instructions to be followed to model the data as per the request from the service. This service uses a Information management services to establish a data link to data storages.[7]

Pooling Services: All the data requests coming from Information management services are managed by pooling services. This service help the other services in establishing a dynamic connection to data storages. This service also handles one way communication between the data storages and Information management services.

This is called event-driven approach. All the activities like addition of data, removal of data in data storages are considered as event. This events are communicated to the information management services.

4 DATA STORAGE AND PROCESSING

4.1 Hadoop and MapReduce

Hadoop and MapReduce are prevalent technologies in storing and processing data. Hadoop has a database in file system called as HDFS. HDFS[4] and MapReduce is an Apache Project which is used to split the data into various segments and store the data in various commodity boxes. This boxes are clustered together to allow the flow of data between them.

As the data is generated at different physical locations, it will be easy to store data at different geographical locations. There will be minimal transmission of data. Changes in electrical grid doesn't require the change in entire data model. On addition or deletion of a electrical node, a new data storage can be added without any intervention to the existing data storages. This distributed model also ensures high availability. Availability of one data source will have minimal or no effect on the availability of the system thus reducing the downtime and business losses.

The data from various sources have different formats. This makes it difficult to store data in traditional relational databases because of type conversions and relational handling. Hadoop overcomes this problem by storing the data in filesystems. Data can be easily pre-processed and stored in the pictorial representations rather than in tables and schemas.

Mapreduce is a programming model. This has two components i.e., Jobtracker and Tasktracker. Jobtracker is a master process which is responsible for scheduling assigning the jobs to Tasktracker. Tasktracker is responsible for execution of the mapreduce jobs.A sample mapreduce[11] task takes has two phases. The first phase is a map phase, where the data is divided into several pieces. The second stage is reduce phase, where the data is processed to produce output.These mapreduce jobs are scheduled and run in batches.This is called Batch Processing.[8] This map and reduce functions are very reliable in analysing the nature and demand of customer from the data available from the most recent processed jobs. Mapreduce jobs run on static data. This will not serve the requests like load analysis, electrical machinery failure, metering failure, power loss which require real time data.

4.2 Apache Spark for Realtime data

Apache Spark[14] is a cluster computing model. It has capability to perform real time analysis of data. IT is nurtured with more enhanced machine learning algorithm and libraries. Spark SQL, MLlib, Spark streaming, GraphX are some of those. Spark framework contains data in distributed sets. It also has set of working programs on the distributed sets of data. This set of programs are called Resilient Distributed Dataset functions[10].

The dynamics of electrical properties changes in milliseconds. In order, to collect these dynamics, the power measuring systems have evolved . New instruments like phasor measurement units have evolved. These devices collect data at the rate of 20-40 readings per second. However, if there is any delay in processing such huge amount of data, then the collected data is not useful. Apache spark tackles this issue in two different approaches.

Streaming Approach:[14] Streaming approach reacts to the each and every event that occurs in the data. As soon as new data is injected, all the resilient distributed dataset functions are called. This function processes data and makes them into a usable format and stores them. This kind of approach is used in metering, billing and load management.

Iterative approach:[5] In this approach, spark offers in memory computing. The datasets are accessed in memory instead of the going to the physical database. All the phasor readings which are required by multiple requests to calculate state space estimation use the developed cache data on the servers instead of accessing them from the data storage. This make requests like state space calculation much lighter.

5 CHALLENGES IN IMPLEMENTING BIG DATA

5.1 Information Security

A large amount of customer electricity usage data is collected. This data must be protected from data leaks. Access control systems must be enhanced to restrict the access to the customer data. Leaked data can be exploited to trace the end user and his/her appliances.[13]

5.2 Asset Management

Assets are the power collection units. These are one of the important devices in the architecture. All the assets must be maintained properly to ensure the quality of data. If any of the power measuring unit goes down or malfunctions, there will be discrepancy in analysing data. This will lead to improper decisions.

5.3 Adaptability

The amount of data in increasing by many folds. In present world, Data Analytics has become a part of Electrical Industry. Though, Many Power Industries have implemented Big Data solutions, there are many industries which are yet to implement Big Data technologies. Most of the South asian countries still use SCADA for processing electrical Data.[1]

6 CONCLUSION

This paper briefly highlights the importance of Big Data Solutions in Power distribution systems. Firstly, Data sources for analytic systems in power distribution like smart meters, Phasor measurement units are briefed. Integration of Data from various sources using service oriented architecture and the important processes in the service oriented architecture are discussed. Later, Implementation of distributed file system i.e., HDFS with processing models like MapReduce and Apache Spark are discussed. At last, challenges like information security, asset management and adaptability of Big Data Technologies are discussed.

REFERENCES

[1] ABB.COM. 2002. ABB SCADA system to automate power for Hyderabad & Secunderabad and streamline APCPDCL electrical distribution network. (2002). http://www.abb.com/cawp/seitp202/bdaf43d0073a9eb965256cb60021c734.aspx

[2] A. B. M. Shawkat Ali (Ed.). 2013. *Smart Grids Oppurtunities,Developments, and Trends*. Springer, School of Information and Communication Technology, Central Queensland University, North Rockhampton, QLD Australia.
[3] Amr A.Munshi and Yasser A.-R.I. Mohamed. 2017. Electric Power Research Systems. *Elsevier* 151 (2017), 68–85.
[4] Apache Hadoop. 2017. Apache Hadoop. (2017). https://en.wikipedia.org/wiki/Apache-Hadoop
[5] Justin Kestelyn. 2013. Putting Spark to Use: Fast In-Memory Computing for Your Big Data Applications. (11 2013). https://blog.cloudera.com/blog/2013/11/putting-spark-to-use-fast-in-memory-computing-for-your-big-data-applications/
[6] MANAGEMENT MANIA. 2015. WS-BPEL (Web Services Business Process Execution Language). (2015). https://managementmania.com/en/ws-bpel-web-services-business-process-execution-language
[7] Jyotishman Pathak, Yuan Li, Vasant Honavar, and James McCalley. 2003. A Service Oriented Architecture for Electric Power Transmission System Asset Management. In *Service-Oriented Computing ICSOC 2006: 4th International Conference, Chicago, IL, USA, December 4-7, 2006, Workshops Proceedings*. Springer-Verlag Berlin Heidelberg 2007, Springer, Berlin, Heidelberg, Chicago, IL, USA,, 26–37.
[8] Shyam R, Bharathi Ganesh HB, Sachin Kumar S, Prabaharan Poornachandran, and Soman K P. 2015. Apache Spark a Big Data Analytics Platform for Smart Grid. *Procedia Technology* 21, Supplement C (2015 2015), 171–178. SMART GRID TECHNOLOGIES.
[9] Abu-Rub Shady S. Refaat, Haitham, Rub, and Mohamed Amira. 2016. Big Data Better Energy Management and Control Decisions for Distribution Systems in Smart Grid. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, Washington, DC, USA, 1–6. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7840966
[10] Chunming Tu, Xi He, Zhikang Shuai, and Fei Jeing. 2016. Big data issues in smart grid, In 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA). *Renewable and Sustainable Energy Reviews* (2016 2016), 1007–1012.
[11] Tutorialspoint. 2017. Hadoop - MapReduce. (2017 2017). https://www.tutorialspoint.com/hadoop/hadoop-mapreduce.htm
[12] EMC Digital Universe with Research & Analysis by IDC. 2014. The Digital Universe of Oppurtunities: The Rich Data and Increasing value of Internet of Things. (04 2014). https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
[13] Nanpeng Yu, Sunil Shah, Mingguo Hong, and Kenneth Loparo. 2015. Big Data Analytics in Power Distribution Systems. In *2015 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, Washington ,DC, USA. http://ieeexplore.ieee.org/document/7131868/
[14] Matei Zaharia. 2017. Apache Spark. (2017). https://en.wikipedia.org/wiki/Apache-Spark

7 BIBTEX ISSUES

Warning–empty publisher in ShadyS.Refaat2016

Warning–no number and no volume in Tu2016

Warning–page numbers missing in both pages and numpages fields in Yu

(There were 3 warnings)

8 ISSUES

8.1 Formatting

DONE:

Incorrect number of keywords or HID and i523 not included in the keywords