I J C T A, 10(9), 2017, pp. 639-645 ISSN: 0974-5572

Survey on Use Cases of Apache Spark

Ramkrushna C. Maheshwar[1], Haritha Donavalli[2] and Bailappa Bhovi[3]

ABSTRACT

This paper mainly focuses on the advantages of Apache Spark and different use cases of Apache Spark. In Big Data Analytics, data can be processed in two ways, one is Disk Based Computation Model, second is In Memory Data Processing Model. Disadvantage with Disk based model is in each iteration of data processing temporary results are kept in disk. If want process next iteration we have to read data from disk, it consumes too much time for reading and writing data from the disk. Advantages with In Memory data processing model, the iterative results are stored in cache memory and for further iteration process data can be directly fetched for execution from cache memory. We could speed up the data processing as the data is directly fetched from cache memory. As Apache Spark is InMemory Data Processing Model. It is used when the system need high throughput with low latency. Apache Spark is a freeware basically used for iterative computations and real time processing on similar data.

Index Terms: Big Data Analytics, Apache Spark, Use cases of Apache Spark, Time Series Data Analytics, Hadoop Map Reduce.

1. INTRODUCTION

The Bigdata is collection of large amount of data, Processing on that been done in distributed manner. The multiple computers were used to process the data, which we normally call as commodity hardware. Big Data components are Hadoop, HDFS, Name Node, Data Node, Map Function and Reduce Function. Hadoop is distributed computing framework used for processing of large distributed data. The data is distributed over large collection of cluster or data is scattered in different data node. "HDFS is a Hadoop Distributed File System used to store the data across the network"[1]. Name node is also called as Job node. It is used to maintain directory of HDFS and it is centralized processing element, accepts processing request from clients or different different browser and divides the task into small task and it dispatched towards to the Task Node or Data Node for computation purpose. The Data node is responsible for processing on data, reverts its resultant results to HDFS. The processing of data will starts from collecting the data inputs by Name nodes from HDFS. This data is distributed to different data nodes then it performs mapping function on that data sets. Mapping is nothing but dividing data inputs into smaller chunks and assigning to different data nodes for computation purpose. After mapping, Shuffling will be performed, transporting data from one data node to another data node according to requirement. Then reducing function will perform on respective data node to get final results at Name Node. All these operations on data will be performed in parallel way in distributed manner. In Hadoop Map Reduce function if we want to perform any iterative operation at that time Map Reduce[13] function use to fetch the data input from HDFS, and perform iterative operations. While performing iterative operations temporary results are stored in HDFS. Here at each iteration Map Reduce function stores temporary results into HDFS and loads from HDFS[13]. So it consumes too much time for loading and storing data values from HDFS. Apache Spark is distributed framework used for processing distributed data over the Hadoop platform. It is faster than Hadoop Map Reduce because of its in memory computation. In Apache Spark if we want to process iterative operations, it performs operation through in memory or cache memory. First time it fetches input data values from

[1,3] Research Scholar at KL University and Worked at I2IT, Pune, *Emails: remomaheshwar1987@gmail.com, bhovibailu@gmail.com*

[2] Professor, Department of CSE KL University, Guntur, Andhra Pradesh, India, *Email: haritha_donavalli@kluniversity.in*

HDFS after first iteration temporary results were stored in memory itself or in buffer cache. Here utilizing processing time very effectively. In Map Reduce each time we were storing data in HDFS instead of buffer cache so becomes Apache Spark having very low latency. The operating system decides should be saved in disk or in buffer cache. For improving efficiency of Spark main role is Resilient Distributed Datasets we call it as RDD[1][4]. If different kinds of queries want to run on same data repeatedly at that time data is kept in memory for efficient execution times.

2. APACHE SPARK

Apache Spark framework is a "Fast and general engine for large scale data processing" [2], [11], [15], [17]. It works in parallel, distributed way for effective data processing. It can be deployed in many ways "Apache Mesos, Apache Hadoop via Yarn, Apache Spark Desktop Manager" [3]. It enables Java, Python, R programming Language and Scala programming languages to design the applications. Basically it uses for streaming based application, iterative based applications, machine learning based applications. Apache Spark is highly recommended when requires low latency queries, real time processing on data and iterative computations. "Spark has a similar programming model to MapReduce but extends it with a data-sharing abstraction called "resilient distributed datasets," or RDDs" [12]. Apache Spark having many more features as follows: it is faster in processing that is it executes batch of jobs simultaneously, it provides real time system to process streamed data, it ensures lower latency computations by caching, it uses the RDD to store intermediate results[5],[6], it provides check points for recovery purpose, it provides fault tolerance that is if any process gets crash while execution, spark will continue where it let off, it performs graph related processing through GraphX. The name node accepts the application for execution. The first task of name node is partition the task into subtask and distributes to the data nodes for further processing. At the data node multiple tasks could be performed and there will be interleaved use of data values. In previously the intermediate result was stored in HDFS file system. The components of apache spark are Spark Core, Machine Learning Libraries (MLlib), GraphX[16], Spark Streaming is utility for incongestion of data, SparkSQL[7] is an interface for Relational Database, Data sources can referable Cassandra, HBase, JSON, elastic search, Comma Seperated Values (csv)[6],[8].

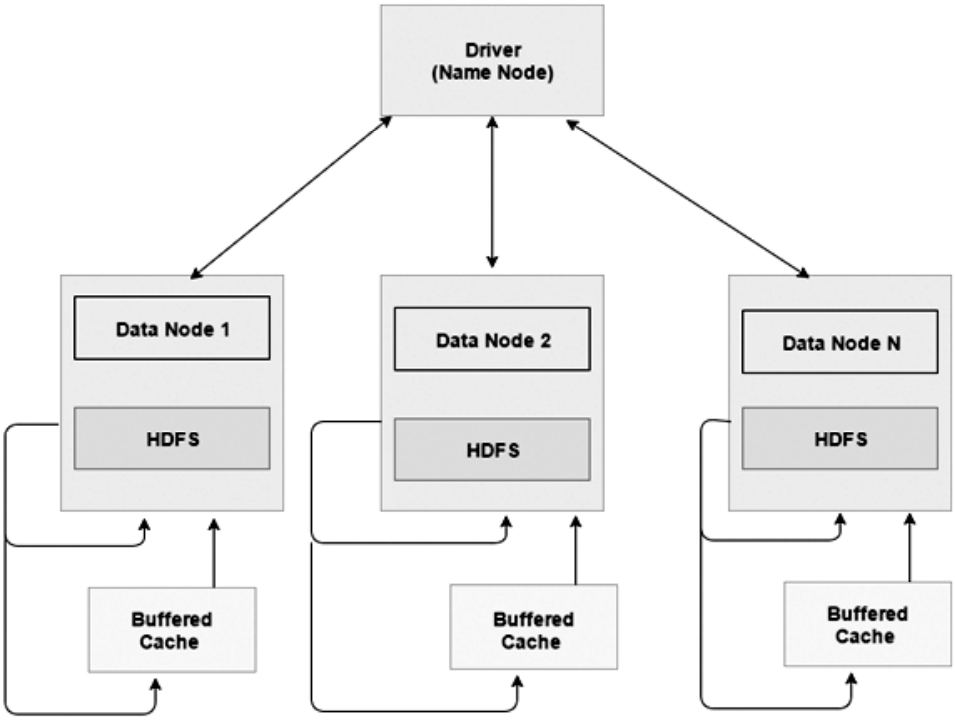

Figure 1: Execution of Application over Apache Spark

In Fig. 1 Apache Sparks system intermediated result are stored in buffered cache, which could help to access data stored in a buffered at a faster rate compared to store intermediate result data in HDFS file system. After completion of execution of task at data node, the result is sent back to the name node. After collecting all the result data from data node, name node will integrate the result and produces the final output.

3. USE CASES

Cancer Detection System, Call Drop Analytics, Driverless Car Analytics, Tax Compliances System, Churn prediction for feeling, Fraud detection for leading bank in India, Live Student performance predictor, Let's find missing child Application.

3.1. eBay eCommerce data analysis using Apache Spark

In eBay Company, Apache Spark with Shark used for ecommerce data analysis to give imperative offers, move execution, optimize the performance of server and client encounters[3]. Apache spark deployed at eBay through Hadoop with Yarn. Yarn deals with the Hadoop Cluster nodes and permits Hadoop to reach out past standard guide and decay occupations by utilizing Yarn holders to run non specific tries. Through the Hadoop Yarn system, "eBay's Spark clients can affect packs drawing nearer the degree of 2000 nodes, 100TB of RAM, and 20,000 cores"[3]. eBay's Spark clients were written their application in Scala and uses Apache Spark Machine Learning Library (MLib), to maintain Cluster performance using KMeans[14] and the trader, clients and product based information is secured in HDFS. These occupations are supporting exposure through cross examination of complex information, information delineating, and information scoring among other utilize cases. Shark is designed for interactive SQL on hadoop system. eBay's basic algorithm for data analytics using Apache Spark:

1) Scan input files from HDFS and switch into usable records
2) Build coaching knowledge set from sample and outline knowledge.
3) Train the model
4) Score the information
5) Store resultant knowledge in computer file.

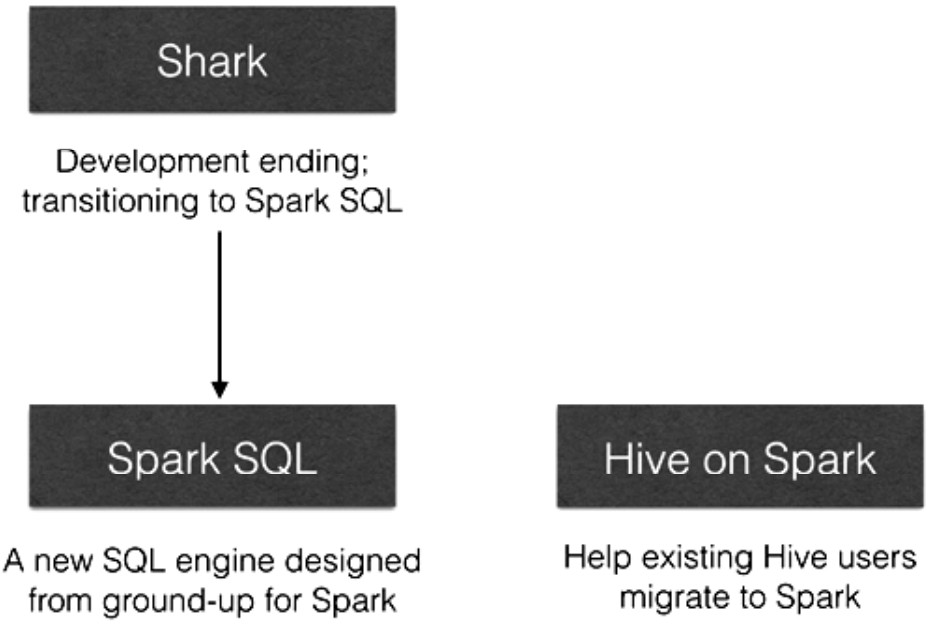

Figure 2: Shark interactive SQL

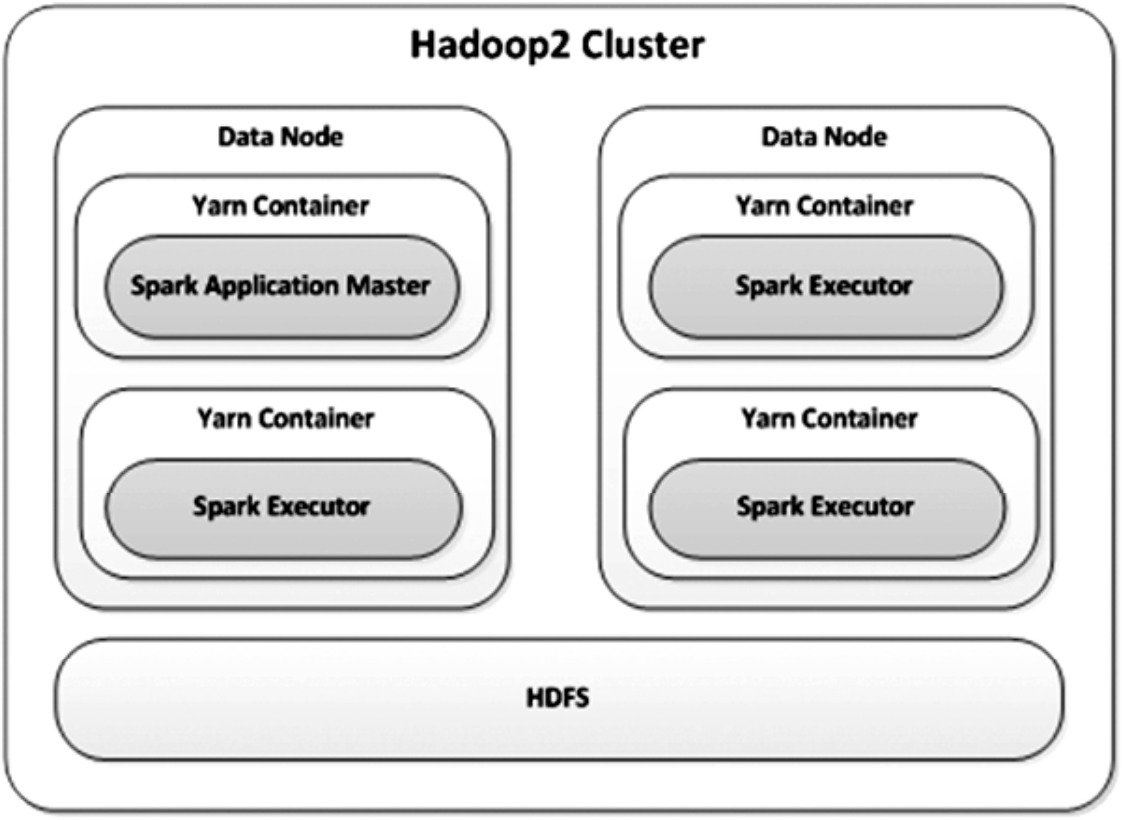

Figure 3: Apache Spark Deployment through Yarn

In eBay have started utilizing Spark with Shark[4] to enliven their Hadoop SQL execution, Shark request are feasibly running 5X speedier than Hive. Fig. 2 Shark interactive queries gets transition into spark SQL and existing hive users can also migrate to spark. eBay making forecasting through available information, along these forecasting what's to come is breathtaking for Spark at eBay. They are turning upward their own Spark by connecting with outside of Hadoop clusters. These bunches of cluster may affect more particular rigging or be application-particular. Various people are joining eBay's beginning now solid information arranges language ventures into the Spark model to make it altogether less hard to effect eBay's information inside Spark.

When user submits the Spark job to Hadoop. In the Fig. 3 Apache Spark[17],[18],[19] application with in Yarn container Name node is initiated, then spawn Spark executors begins working with the Yarn resource manager with requested number of users. These Spark executors uses specified amount of memory and number of CPU cores to run the Spark application. In this case, the cluster's data residing in HDFS is able to read and written by Spark application. This model of running Spark provides a singular, foundational platform for data processing over shared data on Hadoop.

3.2. Big Data Analytics for Smart Grid System using Spark

Smart grid system is automation system; consist of large collection of sensors in power grid system for monitoring and controlling to grid system. Here smart deals with complete automation of monitoring and controlling the power grid system. Data collection through the sensors in every time slots so here will get real time data. On this data using smart grid system process and derive new information for the applications like "utomatic demand response and real time pricing forecasting, online grid monitoring , fault identification, peak time load balancing" [10]. It is capable of analyzing the both slowly and rapidly changing data. It is combination of iterative processing, batch processing and real time data processing. It uses the Machine Learning Library Algorithm on datasets for rapidly expanding its nature. Smart grid data consist of consumer

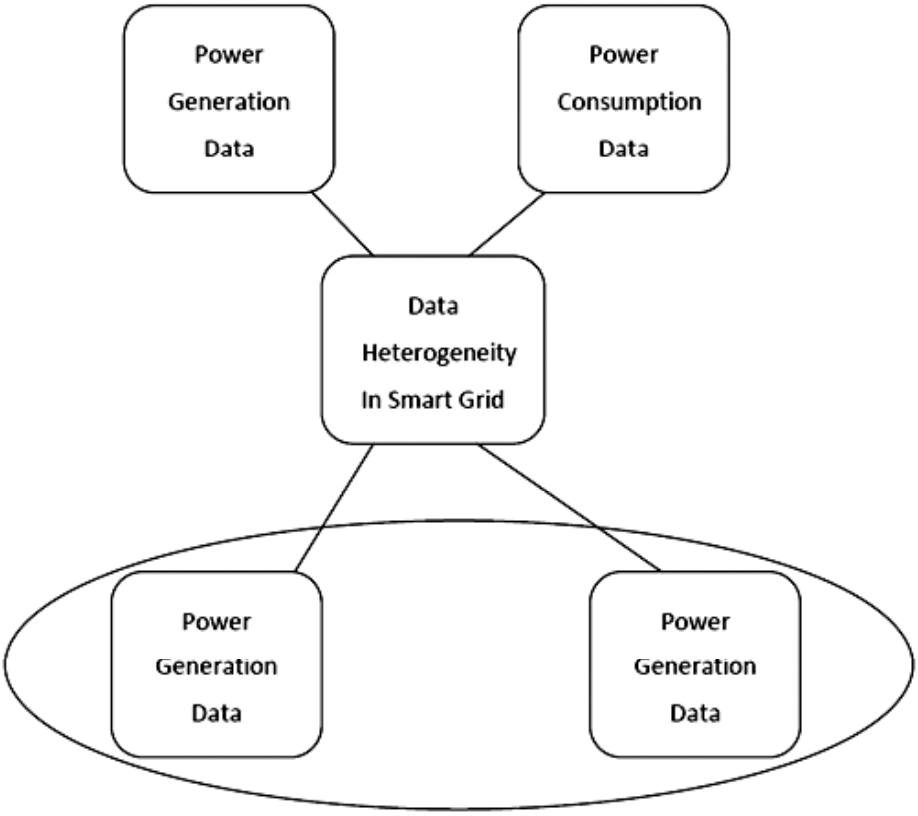

Figure 4: Data Heterogeneity

data, Transmission data, Distribution data, and Generation data. Distribution of data in smart grid is visualized using Heterogeneity data. A data analysis is to make system smart and more intelligent.

In Fig.4, large datasets is divided to small datasets for processing. Its computation is completely performed through distributed, parallel processing. It uses different kinds of processing techniques like the Batch Processing, Stream Processing and Iterative Processing. In this online streaming application designed for wide area monitoring, state estimation, oscillation monitoring, frequency stability monitoring, voltage stability monitoring, Offline application for model validation, post distribution analysis purpose. Both the online as well as offline application developed using the synchrophasor technology. Cluster consists of "PMU (Phasor Measurement Unit) and PDC (Phasor Data Concentrators)"[10]. PMU acts as Name node and PDC acts as Data node. PMU used to measure voltage and current phasors, frequency, circuit breaker status. PDC used to collect phasor data, discrete event data.

3.3. Time Series Document Analysis of Research Papers Citation using Graph Processing

In this use cases time series data and forecasting methods used to "capture the interlinked documents and to predict the citations of paper would receive in future"[11]. Here interlinked documents are World Wide Web pages or online citation indices of papers. To predict the number of citations in future edges represents that hyperlinks or citations for that document or paper. So extracting meaning full information from graph is referred as Hyperlink Analysis. The system's time series analysis categorized into kinetic model and dynamic model. Kinetic Model data is measured using function of the time for different observations, xt=f(t). Dynamic model data is measured using xt = f(xt-1, xt-2, xt-3…). While decomposing the time series function we have to take care of long term monotonic change, long wave trade cycles, fluctuation behaviour, influencing observations.

Smoothing techniques used for reducing of cancelling the effect due to random behaviour. Average methods used for average of all past observations equally associates:

$$x = 1/n \sum xi = (1/n)\ x1 + (1/n)\ x2 + \ldots\ldots + (1/n)\ xN$$

Exponential smoothing is combination of smoothing technique and Average methods. It is begins with the S2 to Y1 here S stands for smoothing observations and y represents for original observations. The basic equation of computing t is time period, S computing power and á represent for dampling. The main moto of the system is to predict for given time series data x1, x2, x3…xN; is to calculate future data values such as xN+K i.e. x(N,K) where k is lead time and at time N for K steps ahead. Forecasting can be done in two ways Automatic and Non Automatic forecasting [9], [11]. Automatic forecasting does not includes human intervention its completely system oriented. Non Automatic forecasting deals with subjective input from forecasters. Straight line forecasting model for time series data can written using E(Yt) = β0+β1there Yt, to time t, and at least square lines is used to forecast future values using Yt. The model for forecasting is follows,

$$Predicte_{dt}+1 == Predicte_{dt} + 0.3*9(Actual_{t}-Predicte_{dt}).$$

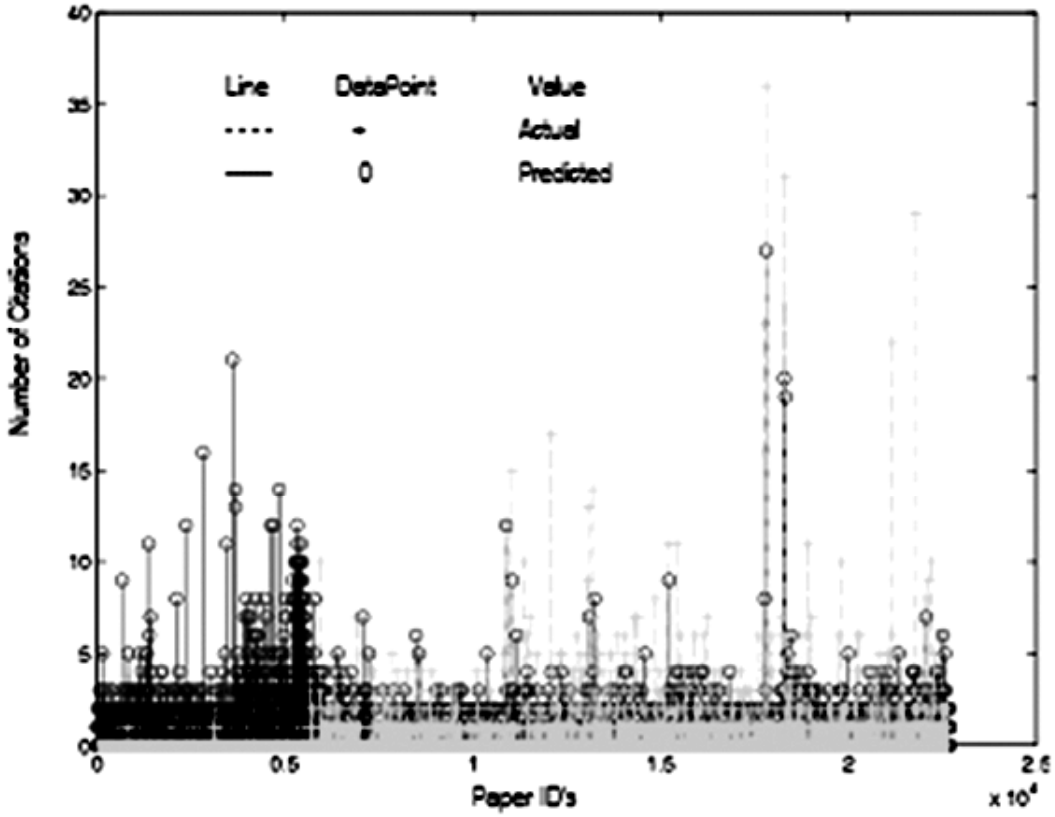

Figure 5: Number of Citations Vs Paper ID's

The system result was not highly accurate, can increase degree by high accuracy. Future work suggested to estimate number of downloads for paper in first three months of publications. That is predicting number of downloads by its popularity of author. In the Fig. 5 Number of Citation Vs Paper ID's graph depicts that actual and prediction of number citation based on paper ID.

4. CONCLUSION

As the conclusion of the survey on different use cases of Apache Spark Framework initiate me to perform the prediction analysis through time series data. Now I got to know that on which domain can perform predictive analysis using time series data. In eBay eCommerce data analysis, they had initiated the prediction for providing different offers for customer, optimize the server performance, and improve search engine efficiency through spark. In data analytics for grid system, they had implemented an automation system that monitors and controls the power grid system. In time series document analysis, they had captured the

interlinked documents and to predicted the citations of paper would receive in future and plotted the graph of citation of papers.

REFERENCES

[1] Dilpreet Singh and Chandan K Reddy, "A survey on platforms for big data analytics", Journal of Big Data a Springer open Journal, Published on Oct 2014.

[2] Tao, Hang, Bin Wu, and Xiuqin Lin. "Budgeted mini-batch parallel gradient descent for support vector machines on Spark", 2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS), 2014.

[3] eBay Global Data Infrastructure Analytics Team, "Using Spark to Ignite Data Analytics", http://www.ebaytechblog.com/2014/05/28/using-spark-to-ignite-data-analytics/, on 05/28/2014.

[4] Reynold Xin,Shark, "Spark SQL, Hive on Spark, and the future of SQL on Apache Spark", ENGINEERING BLOG, July 1, 2014.

[5] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma,Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica, "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing", NSDI'12 USENIX Symposium on networked design and implementation with ACM SIGOCOMM and ACM SIGOPS, SAN-JOSE,CA, April 25-27, 2012.

[6] Abdul Ghaffar Shoro & Tariq Rahim Soomro, "Big Data Analysis: Ap Spark Perspective", Global Journal of Computer Science and Technology: C Software & Data Engineering Volume 15 Issue 1 Version 1.0 Year 2015.

[7] Michael Armbrusty, Reynold S. Xiny, Cheng Liany, Yin Huaiy, Davies Liuy, Joseph K. Bradleyy,Xiangrui Mengy, Tomer Kaftanz, Michael J. Franklinyz, Ali Ghodsiy, Matei Zahariay, "Spark SQL: Relational Data Processing in Spark", AMPLab, UC Berkeley, 2015.

[8] Zhijie Han, Yujie Zhang, "Spark: A Big Data Processing Platform Based On Memory Computing", IEEE Xplore: 21 January 2016.

[9] (Accessed on 26th October 2016) Predictive analytics Comparisons[Online]Available: http://butleranalytics.com/enterprise-predictiveanalytics-comparisons-2014/

[10] Shyam R., Bharathi Ganesh H.B., Sachin Kumar S.,Prabaharan Poornachandran, and Soman K.P."Apache Spark a Big Data Analytics Platform for Smart Grid",Procedia Technology, 2015.

[11] Prasanna Desikan, Srivastava,"Time Series Analysis and Forecasting Methods for Temporal Mining of Interlinked Documents", University of Minnesota, Data Analyst Research Group, 2003.

[12] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, And Ion Stoica,"Apache Spark: A Unified Engine for Big Data Processing", DOI:10.1145/2934664 | VOL. 59 | NO. 11 | Communications Of The Acm, November 2016.

[13] (Accessed on 23rd October 2016) MapReduce [Online] Available: https://github.com/RevolutionAnalytics/rmr2/blob/master/docs/tutorial.md

[14] Shwet Ketu,Sonali Agarwal, "Performance Enhancement of Distributed K-Means Clustering for Big Data Analytics Through In-Memory Computation", IEEE Xplore: 07 December 2015.

[15] Nabi, Zubair. "Introduction to Spark", Pro Spark Streaming, 2016.

[16] Hameeza Ahmeda, Muhammad Ali Ismaila,*, Muhammad Faraz Hydera, Syed Muhammad Sheraza, Nida Fouqa., "Performance Comparison of Spark Clusters ConfiguredConventionally and a Cloud Service ", Science Direct, Procedia Computer Science 82 (2016) 99–106.

[17] (Accessed on 3rd November 2016)" 6 Sparkling Features of Apache Spark!" [Online]Available: https://dzone.com/articles/6-sparkling-features-apache

[18] (Accessed on 5th November 2016) "Spark Tutorial" October 20-22,2014 University of Maryland, College Park [Online]Available: http://lintool.github.io/SparkTutorial/

[19] (Accessed on 7th November 2016) "5-things-one-must-know-about-spark" [Online]Available: www.edureka.co/blog/webinars/5-things-one-must-know-about-spark/+&cd=1&hl=en&ct=clnk&gl=in