I J C T A, 9(17) 2016, pp. 8665-8672

A Study of Big Data Analysis Using Apache Pig

Krati Bansal[*] and Priyanka Chawla[**]

ABSTRACT

Today Information Technology has achieved new hike by using the technology on Big Data. Analyzing data in tear bytes and beyond that –"Hadoop" is the ultimate solution. Nowadays, maintenance of data has become an integral part of life varied from personal data stored in mobiles to vast data stored on internet by millions of users. Organizations come across large data on daily basis which is either structure or most of the times unstructured. Maintenance and analysis of huge data is a big challenge in front of corporate companies which calls for use of Hadoop and using various components of Hadoop. Hadoop being at a developing stage utilizes familiar scripting platforms such as Apache Pig for reducing processing and generate faster results. Aim of this research study is identify Hadoop current shortcomings and benefits of using Pig on Hadoop for analyzing Big Data.

Keywords: Apache Pig, Pig Latin, Big Data, Hadoop, Map Reduce, Hadoop Distributed File System (HDFS), SQL, HBASE, OOZIE, SQOOP, Hive

I. INTRODUCTION

The idea of using Hadoop on Big Data is revolutionizing approach towards analysis, operation and maintenance of huge data. Information Technology companies are adopting various scripting platforms with Hadoop to reduce the time for analyzing and structuring the gigantic data. Hadoop is the need of the hour for Database companies which deal with enormous data at every step. Apache Pig is an important component of Hadoop Ecosystem which reduces the coding and development time for analyzing big data. This research study furnishes information on big data and various challenges related to big data are discussed in Section II. This paper in section III makes available information on Hadoop architecture and specifications of various components of different layers of Hadoop Ecosystem. Section IV of this research study explains Apache Pig and adoption of Apache Pig with Hadoop. Section V of this study explicates the comparison of scripting platforms and provides metaphor between 2 platforms Hive and Apache Pig. Existing work of big data analysis using Hadoop and Apache Pig has also been portrayed in section VI. Finally, section VII concludes the paper and future work has been discussed in section VIII.

II. BIG DATA

Big Data is an emerging phrase that represents cavernous heap of unstructured, semi-structured and structured data which can be mined for potential information [1]. Big Data can be signified by 3 V's, first V symbolizing enormous volume of data, second V indicates the velocity of processing this vast data and third V represents extensive variety of different types of Data. Figure 1 is a graphical representation of the characteristics of Big Data [2] [4].

[*] Department of Computer Science and Engineering, Student of M.Tech I.T.S Engg College, Greater Noida, Uttar Pradesh, India, *E-mail: kratibansl1992@gmail.com*

[**] Department of Computer Science and Engineering, NIET, Greater Noida, Uttar Pradesh, India, *E-mail: priyankachawla.cse@its.edu.in*

2.1. Challenges of Big Data

- Lack of resource availability (Data Scientists) is the biggest challenge of big data.
- Storage, Maintenance, Processing, Management are basic challenges during analyzing Big Data.
- Unstructured data in Big Data increase the complexity of processing data:-

LOW GRADE DATA + COMPLEXITY = BIG ISSUE

Low grade data which is also complex, affects the overall performance and quality of the projects. This also consumes lots of time to process and analyze the data, which gradually increase the processing time and eventually rising the overall costing [6] [8].

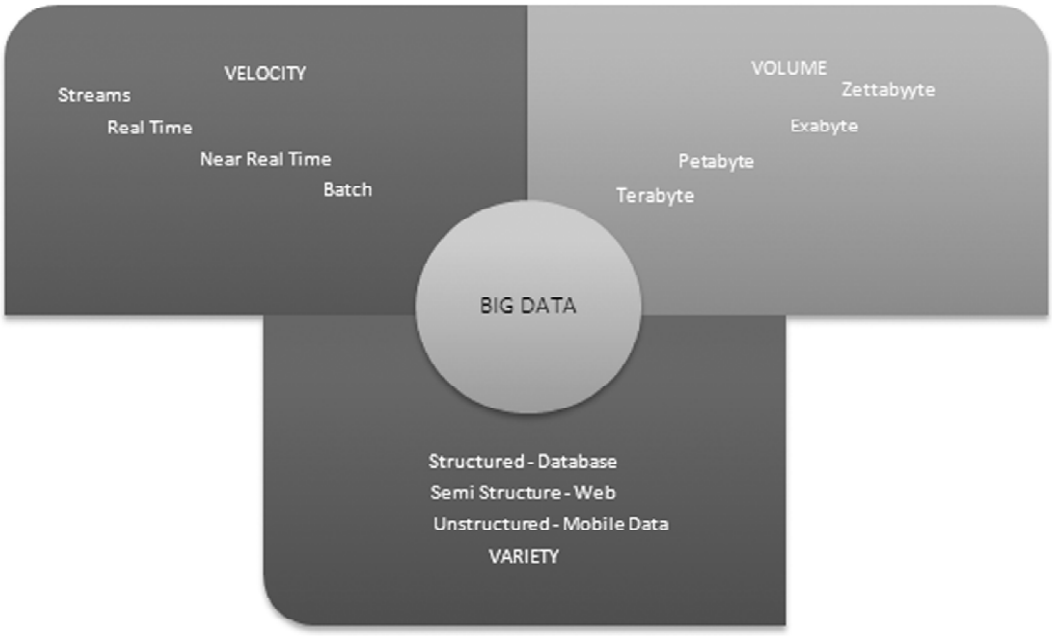


Figure 1: Characteristics of Big Data

Figure 1 describes Big Data concept and also the characteristics of Big Data. Challenges and issues related to maintenance and analysis of Big Data can be reduced using Hadoop Architecture [3].

III. HADOOP ARCHITECTURE

In 2009, Apache developed an open source software language for distributed data processing on large amount of data and named it as Apache Hadoop. Hadoop was developed for computation large amount of data in distributed mode and storage of data in multiple data nodes [2]. Big Data can be refined by adopting Hadoop. Hadoop architecture contains two important layers: Hadoop Distributed File System (HDFS) layer and Map Reduce layer. Hadoop is also related to the components like Apache Pig, Hive, HBase, H CATALOG, Oozie, Sqoop, Ambari, Zoo Keeper and Mahout which are applied at top most layer of Hadoop. The lowest layer of Hadoop can also run on two more components, first is Amazon S3 which is based on cloud and second being Map R which is based recovery time and automatic data backup [7]. Various layers of Hadoop are used for maintenance and analyzing Big Data. Components of Hadoop Ecosystem as especially top most layer of Hadoop such as Ambari, Hive, Pig, Sqoop, Zookeeper and Oozie are utilized for analyzing Big Data [8]. Hadoop Disturbed File System (HDFS) section provides information on the Master node, Slave mode and also functions of HDFS. HDFS is the lowest layer of Hadoop Ecosystem. HDFS is the fault tolerant system of Hadoop [8] [9].

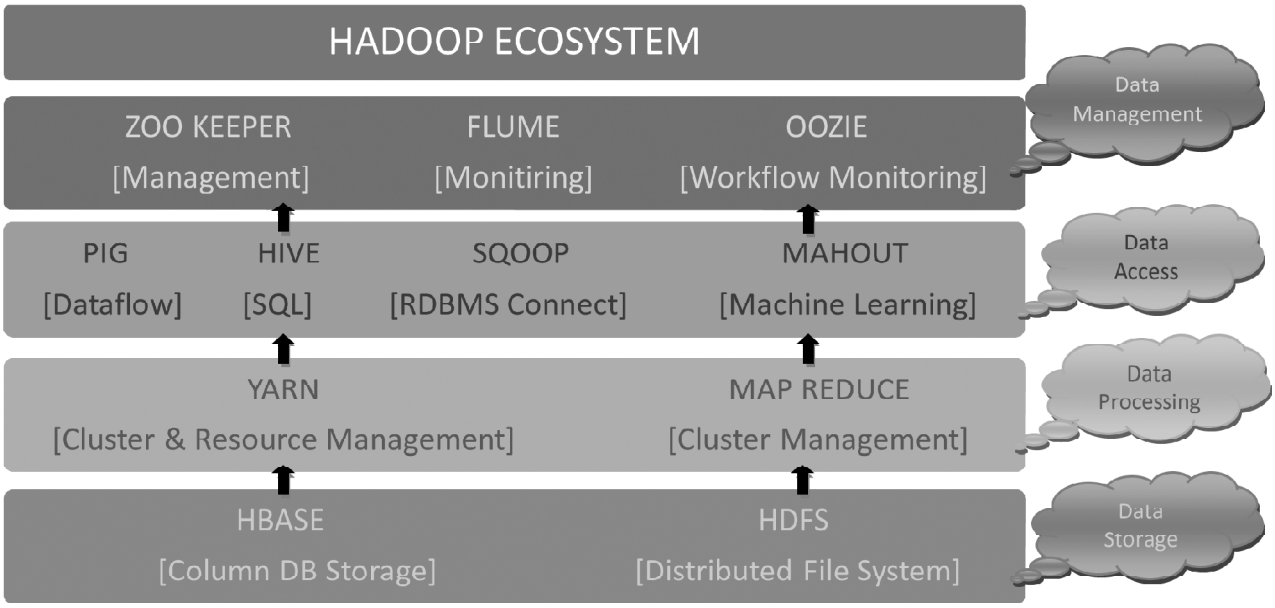


Figure 2: Components of Hadoop Ecosystem

(A) Hadoop Distributed File System (HDFS)

Distributed File System of Hadoop which is also an extended version of Google File System is called as Hadoop Distributed File System (HDFS). HDFS accumulates huge amount of data. Basically Hadoop consist of two nodes: Data Node (Slave Node) and Name Node (Master Node) which is written in java on Hadoop. Data nodes are used in cluster mode which enables the user to work on multiple machines from a single data node. There is a rare case when 2 data nodes run on single machine [13]. HDFS works on huge data sizing up to peat bytes. HDFS does the partitioning and schedules the task on the machine after distribution of data nodes which creates a large file system. HDFS is the lowest layer of Hadoop [20]. Figure 2 represents the Work flow of Nodes [18].

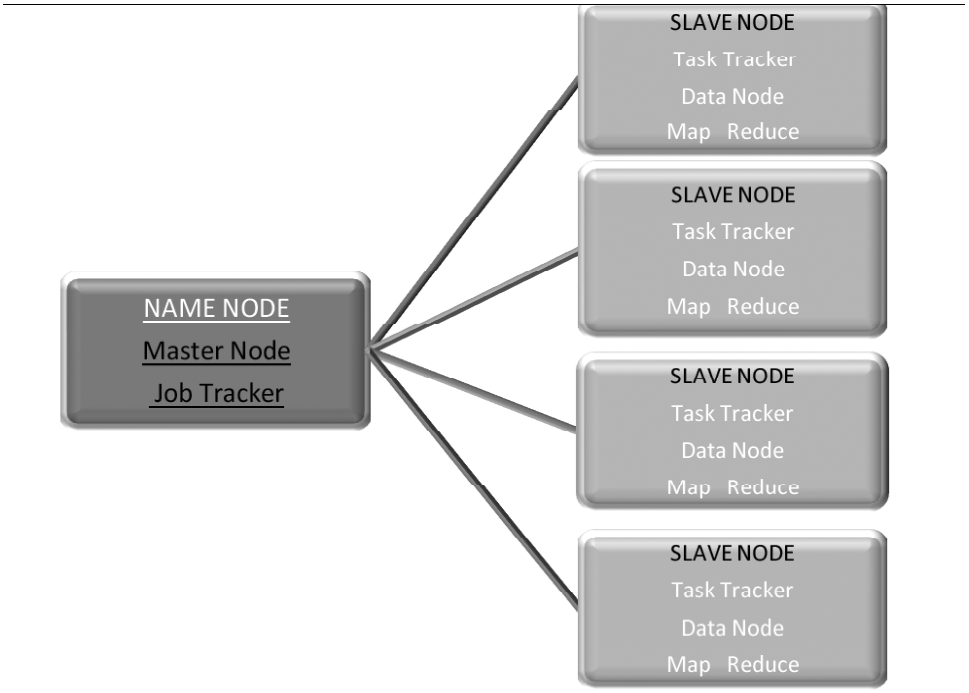


Figure 3: Work Flow of Nodes

(B) Map Reduce

Data processing component of Hadoop is commonly known as Map Reduce. Map Reduce is the second layer of Hadoop which is above HDFS. Map Reduce comprises of 2 functions: - Map function and Reduce function. Map Reduce phase produce result into the single Key. Map function divides the data into chunks for processing, post that reduce function merges the data and produces same key in the form of results. Functionality of Map Reduce can be described as first phase of Map Reduce is Map function which is used for mapping the data. Second phase is Reduce function which is utilized for processing intermediate data and for appropriate final outcome [17] [22].

IV. APACHE PIG

Apache Pig is a scripting platform created by Yahoo later taken over by Apache Foundation which works on data flow language. Apache Pig platform is used for analyzing the large data sets. Apache Pig creates a mechanism for executing data flows parallel to analysis of data on Hadoop [17]. Pig Latin scripting language is the backbone of Apache Pig, for expressing these data flows. Pig is Apache Pig runs on Hadoop by using of Map Reduce for data processing and also Hadoop Distributed File System, HDFS. Apache Pig satisfies the Pig Latin scripts that users have written into a series of one or multiple Map Reduce [28]. Pig Latin language has no If statements or for loops because it only focuses on execution of data flow [14]. Search Engine giant Yahoo is one of the companies who have adopted Apache Pig technology to computing data and also for resolving issues via Apache Pig platform. Apache Pig is implied on distributed environment which utilizes Map Reduce for providing useful results on large data sets. This unable to resolve data sets problems at real time [27].

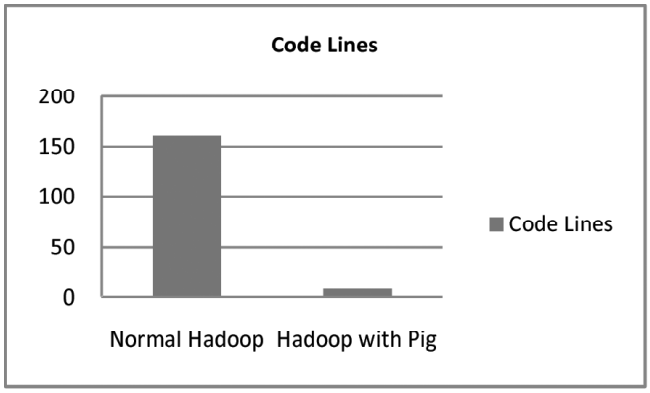
Apache Pig is most widely being adopted follow cases by-

(i) Web Search Platform
(ii) Ado Queries
(iii) Web log processing
(iv) Data sets processing
(v) Iterative data
(vi) Replicated data
(vii) Unstructured data

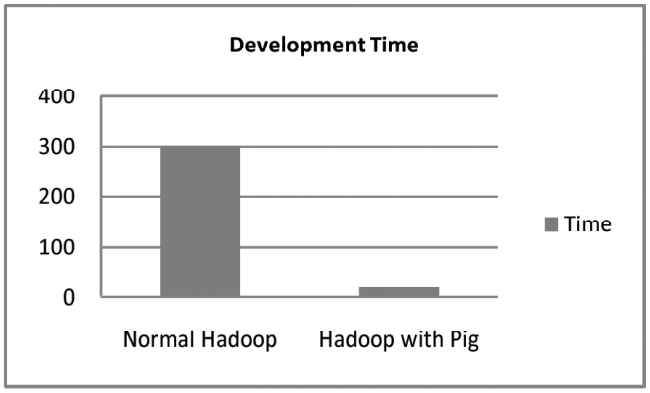
Apache Pig can be considered as one of the best scripting platform which analysis unstructured and inconsistent data in very less time by using two important layers of Hadoop: Map Reduce and HDFS. Pig Latin is one of the powerful and valuable languages for Hadoop. Apache PIG platform and Pig Latin together are one of the most suitable tools and technology which work on scripting language which is similar to SQL based tools [9] [28].

Figure 4: Hadoop work flow using Apache Pig

Analysis reveals that Apache Pig is an important component of Hadoop ecosystem, as Apache Pig uses both the lower level components of Hadoop which are Map Reduce and HDFS for providing appropriate results. Pig Latin the scripting language used on Pig platform is also unique as it focuses on dataflow whereas other languages focus on control flow and data flow [11].



Graph 1: Comparison on Code Lines between Normal Hadoop and Hadoop with Pig [33]



Graph 2: Comparison on Development time between Normal Hadoop and Hadoop with Pig [33]

Graph 1 replicates that Pig requires only 1/20th lines of the code as compared to Hadoop without using Pig. Graph 2 provides graphical representation that Pig takes on 1/16th time for development as compared to normal Hadoop.

V. COMPARISON OF FRAMEWORKS

Hadoop utilizes Pig and Hive frameworks at top most layers. Hive is a data warehouse application whose infrastructure is compatible for usage with Hadoop top most layer and for providing analysis, summarization and query. Pig is a high level scripting platform which is used to analyze data using Pig Latin scripting language. Pig is also utilized on the top most layer of Hadoop ecosystem. Table 1 provides comparison between Pig and Hive [10] [13].

Table 1
Comparison of two Frameworks: Hive, Apache Pig [29] [30] [32]

PIG	HIVE
• Founder: Facebook	• Founder: Yahoo
• Language: Pig Latin a procedural dataflow scripting language.	• Language: HiveQL a declartive SQL subset language.
• Backed Language: Backed works on SQL.	• Backed Language: Backed works on Map Reduce.
• Utilization: Pig is ultized for programming.	• Utilization: Hive is utlized for generating reports.
• Utilized by: Researchers and Programmers used Pig.	• Utilized by: Data Analyst used Hive.
• Operational Use: Pig operates on any cluster on the client side.	• Operational Use: Hive operates on any cluster on the server side.
• Compartibility: Pig works on unstructured and schema less data.	• Compartibility: Hive doe not work on unstructured and schema less data.
• Performance: Pig performance is not good as compared to HIve.	• Performance: Hive gives better performance as compared to Pig.
• Meta Data: Pig does not have a dedicated Meta Data database.	• Meta Data: Hive stores the Meta Data in any local database.
• Data Partition: Pig defines the data in the script, hence Pig is able to do data partitioning .	• Data Partition: Hive partitions the data into the tables and futher the tables are subdivided into buckets.
• Avro: Avro is compartible with Pig.	• Avro: Hive does not support Avro.
• Learning: Pig has many varitions which need time and efforts, hence Experts use Pig.	• Learning: Hive being SQL subset language, hence easy to learn for Data Analyst.
• Recommendation: Recommended for streaming satellite, real world base data for processing.	• Recommendation: Recommended for databases with structured data .

VI. LITERATURE SURVEY

This section discuss about existing research studies conducted on maintaining BIG Data and challenges for analyzing Big Data.

Bo Li, "Survey of Recent Researches Progress and Issue in Big Data"

In this research paper authors have furnished information on Big Data and various characteristics of Big Data. Big Data cannot be managed by SQL or normal analytic tools. Previously Big Data worked on Classic big data but now it works on network, big data in cloud, data engineering and benchmarking approach mobile big data network.

Authors have described various challenges and intercept the data which is necessary to implement but these fields are latest in trends [26].

Ajay Kumar, Seem, "Distributed and Big Data Management in Grid Computing",

Authors have described that grid computing need data storage Management and researcher have proposed new architecture which is Dynamic & Scalable storage management (DSSM). In DSSM Grid storage are divided into multiple geographically domain. In each phase it explains the algorithm [23].

Chantal, Shelling, Manor, "Algorithm and Approaches large data base –A survey"

In this paper Researcher explained about the data mining which extract the knowledge for the data and analysis. This research study provides information on Storage System, Handling, and Analysis which were used in previous part of data. Authors have defined cluster based master slave architecture and use the Apriority Algorithm which produced the result for the truncation time on the data [25].

M.M.Hansen, T.Miron –Shatz, "Big Data in Science and Health Care" this paper provides information on machine learning and predication algorithm. It also defines the idea of connectivity brain model tasks from neural network in machine learning form. This paper furnishes new way of big data in term of scientific and healthcare research. Challenges related the data privacy, confidentiality learning and analytics are published in this research study [24].

Jurmo Mehine, Satish Srirama, Pelle Jakovits "Large Scale Data Analysis Using Apache Pig" this research paper demonstrates use of Apache pig on real world problem. This research paper also provides analysis on data collect from news web links for creating new ways to showcase the news updates. This paper provides information on representing news updates in various categories. Authors have showcased analysis on utilization of Apache Pig tool and RSS as input tool for appropriate results [15].

Dave Jaffe "Three Approaches to Data Analysis with Hadoop" this research study provides information on analysis of large scale data by adopting three tools HIVE, PIG and Map Reduce with Hadoop, it take web logs, which contain customer habit like shopping social media network on the basis of that data result Authors have indicated that HIVE and Pig are more quick to develop whereas Map Reduce takes longer time to run [20].

VII. CONCLUSION

Big Data being gigantic in size analysis and structuring of Big Data is a big challenge in front of researchers. Implementing of Hadoop on Big Data has provided solutions to maintain Big Data. Apache Pig works on data flow and provides appropriate results which are easy to understand with space and time complexity. Pig is a user friendly tool which makes user to access more information in one place. Adopting Apache Pig with Hadoop can provide appropriate results for analyzing and structuring Big Data with less coding lines as compared to traditional coding. Analyses also reveal that Pig is one of the most suitable scripting platforms

for analyzing and structuring of Big Data with lesser development time. Pig Latin is also one of the most compatible scripting languages on Hadoop.

VIII. FUTURE WORK

Research study conducted on analyzing Big Data through Hadoop has various challenges. Hence new tools and applications such as Pig, Hive on implemented with Hadoop to reduce processing time and development time. Researchers have only focused on using Pig on Hadoop for basic analysis of Big Data. In our future endeavor we will conduct research to create an advanced application using Pig on Hadoop which analysis and structuring Big Data. We will utilize information of professional misconducts conducted by doctors in United States for analyzing and advance sorting of Big data which is user friendly. Our main goal is to generate appropriate results by adopting this advance application on Hadoop using Pig.

REFERENCES

[1] Rini T. Kaushik,Milind Bhandarkar,Klara Nahrstedt, "Evaluation and Analysis of Green HDFS". IEEE Paper, 5 August 2012

[2] Kurnal Dave Mr. JigneshVania, "Servey on Big Data Processing using Hadoop Component", IJSRD vol.3, Issue 01, 2015

[3] Steve Lohr "How Big Data Became So Big" The New York Times, 11 August 2012

[4] Stephen Kaisher, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issue and Challenges Moving Forword", 46th Hawaii International Conference on Sysytem Science, 2013

[5] Nawsher, Ibrar, Ibrahim Abaker Targio Hashem, Zakir Inayat "Big Data :Survey, Technologies, Opportunities, and Challenges", C Volume, 2014

[6] Divyakant Aggarwal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwas Dayal, "Challenges and Opportunities on Big Data" Cyber Centre Technical Report, 1 January 2011.

[7] Daniel. F "Extract, Transform, Load Big Data with Apache Hadoop" Intel, 19 July 2013.

[8] Lev Manovich, "Trending: The Promises and Challenges of Big Data Social Data", in proceedings at Debates in the Digital Humanities, Univesity of Minnesota Press, January 2012

[9] Sam Madden, "From Data bases to Big Data", IEEE Computer Society, 2012

[10] Sanjeev Dhawan, Sanjay Rathee, "Big Data Analytics using Hadoop Component Like Hive and Pig", American International Journal of Researching Science, Technoloy, Engineering & Mathematics, pp88-93, March-May 2013

[11] Vahid Jalali , David Leake, "Manual for Bear Big Data Ensemble of Adaptations for Regression Version 1.0", General Public License Version 3, 5 October 2015

[12] EMC2 "Data Lake For Data Science" EMC White Paper, May 2015

[13] Dr. E.Laxmi Lydia , Dr. M.Ben Swarup, "Analysis of Big Data Through Hadoop Ecosysytem Component Like Flume , Hive ,Pig and Mapreduce", International Journal of Computer Science Engineering , Volume 5, 01 January 2016

[14] J.Ramsingh , Dr. V. Bhuvaneswari, "An Insight on Big Data Analytics Using Pig Script", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 4 Issue 6, pp 84-90, November-December 2015

[15] Jurmo Mechine, Satish Sriama, "Large Scale Data Analysis Using Apache Pig", Master Thesis, Tartu, 2011

[16] Casey Stella "Apache Pig for Data Science" in proceeding at Linuz Foundation, 9 April 2014

[17] Keren Ouaknine, Michale Carey, Scott Kirkpatrick "The Pig Mix Benchmark on Pig , Map Reduce , and HPCC System"IEEE International Congress on Big Data, 2015

[18] Willam M.Buros, Guan Cheng Chen, Mei-Mei Fu, Anne E. Gattiker, Fadi H.Gebara, Ahamed Gheith, H. Peter Hofstee, Damir A. Jamesek, Thomas G. Lendacky, Jian Li, Yan Li, John S. Poelman, Steven Pratt, Ju Wei Shi, Evan Speight, Peter W. Wong, "Understanding System Architecture for Big Data", IBM, 7 March 2013

[19] Christopher, Benjamin, Utkarsh, Ravi, Andrew, "Pig Latin :A Not –so –Foregin Language for Data Processing", ACM, June 2008

[20] Dave Jaffe, "Three Aproach to Data Analysis with Hadoop", Dell Technical White Paper, November 2013.

[21] Pradeepa A, Dr. Antony Selvadoos, "Significant Trends of Big Data Analytics in Social Network", IJARCSSE, Volume 3, Analysis, 8 August 2012

[22] The Hadoop Ecosystem Table https://hadoopecosystemtable.github.io/

[23] Ajay Kumar, Seema Bawa "Distributed and Big Data Storage Manangement in Grid Computing", International Journal of Grid Coputing & Applications, Vol. 3 Issue 2, p 19, June 2012

[24] M.M.Hansen, T. Miron –Shatz, "Big Data in Science and Health Care", IMIA and Schattauer Gmbh, IMIA Year Book of Madical Informatics, 2014

[25] Chanchal Yadav,Shuliang Wang,Manoj Kumar, "Algorithm and Approaches large data base –A servey", International Journal Computer Science and Network, Volume 2, Issue 3, 2013, ISN -2277-5420

[26] Bo Li, "Survey of Recent Researches Progress and Issue in Big Data"

[27] Welcome to Apache Pig! <https://pig.apache.org/>

[28] Apache Pig – Hortonworks <http://hortonworks.com/apache/pig/>

[29] Hive – Introduction <http://www.tutorialspoint.com/hive/hive_introduction.htm>

[30] Tutorial - Apache Hive - Apache Software Foundation <https://cwiki.apache.org/confluence/display/Hive/Tutorial>

[31] Hadoop Tutorial - YDN – Yahoo <http://developer.yahoo.com/Hadoop/tutorial>

[32] Processing frameworks for Hadoop - O'Reilly Media <https://www.oreilly.com/ideas/processing-frameworks-for-hadoop>

[33] An Introduction to Pig -StratApps <http://www.stratapps.net/intro-pig.php>