ISSN 2278-3083
Volume 5, No.3, May - June 2016
International Journal of Science and Applied Information Technology
Available Online at http://www.warse.org/IJSAIT/static/pdf/file/ijsait01532016.pdf

Analysis of social networking data using Map Reduce and Hadoop

Sakshi Chauhan
Bahra University
Shimla Hills
Chauhansakshi47a@gmail.com

Anandita Singh Thakur
Bahra University
Shimla Hills
Anandita275@gmail.com

Madhusudan
Himachal Pradesh University
Summerhill Shimla
msudan_07@outlook.com

ABSTRACT

The increasing use of internet technologies with the convergence of computing machines, from mainframe to cellular devices and to multimodal HCIS, has resulted into tremendous amount of data rich in volume and variety, since all data is not always required ,but only the relevant one suited for particular information need, some methods are required to supply user with application specific data .The technique to this huge amount of data and to extract value out of this volume and variety rich data are collectively called Big data. Over the recent years, there has been an emerging interest in big data for social media analysis. The three V's of big data describe volume of data, variety types of data and velocity that defines the rate at which the data is processed by many social networking sites such as are considering big data. As big data is used when we speak usually in Peta-bytes and Exabyte's of data due to this problem it is difficult for these social networking data to keep up the integration. In this paper we discuss about the big data, Mapreduce which plays an important role that supports big data to understand the problems of social networking site. Hadoop which is used with mapreduce to increase the performance of the big data by creating clusters on different nodes.

Keywords: Big data, Mapreduce, Hadoop.

1. INTRODUCTION

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured [37] [38]. Big data are of particular interest in data science, although the discipline is not generally considered to be restricted to such big data, and big data solutions are often focused on organizing and pre-processing the data instead of analysis. The development of machine learning has enhanced the growth and important of data science[38].Data Scientist use their data and analytical ability to find and interpret rich data sources; manage large amount of data despite hardware, software and bandwidth constraints; merge data sources; build mathematical model using that data sources; and present and communicate the data insights/findings[39].

The growing number of people using social media to communicate with their peers and document their personal everyday feelings and views is creating data on an epic scale [1].In social networking sites data placement is a problem while we are talking about big data. Inserting more data creates difficulty in managing the highly efficient storage and space which directly affects the performance of the system. To deal with the fast query processing is also a problem faced which affects the time constraint. The basic requirements for data placement are Fast data loading, fast query processing, highly efficient storage and space utilization; and strong adaptability to highly dynamic patterns. To generalize all these requirements mapreduce show the support to big data and the challenges faced by big data regarding structured, semi- structured and unstructured data. As big data is measured in Peta-bytes and Exabyte's due to this reason it is difficult to manage the internet traffic and to keep the track of packets over the single nodes. To solve the problem of single node failure we discuss the use mapreduce, hadoop as it supports the parallel distribution of data so that problem of single node failure will be solved. Analysing and managing huge amount of data creates overhead so Mapreduce allows data to be clustered on different nodes using the programming paradigm Hadoop. This paper will review Map reduce and hadoop technologies that supports big data.

1.1 Big Data

Over the past 20 years, data has increased in a large scale in various fields. According to a report from

International Data Corporation (IDC), in 2011, the overall created and copied data volume in the world was 1.8ZB (≈ 1021B), which increased by nearly nine times within five years [2]. Big data is different from other datasets as it is not analysed by the relational database tool, Statistical analysis tool and visualisation aids Instead it needs fast parallel software's running on tens or even thousands of servers according to different cases Big data is a technology that processes high Velocity , high volume, high variety dataset to extract intended data to ensure high veracity Of original data obtained information that demand cost effective, innovative forms of data and information processing (analytics) For enhanced insight, decision making, and Processes control [3].

1.1.1 Characteristics of big data

Big data is a broad term for data sets so large or complex the traditional data Processing is inadequate to manage them.

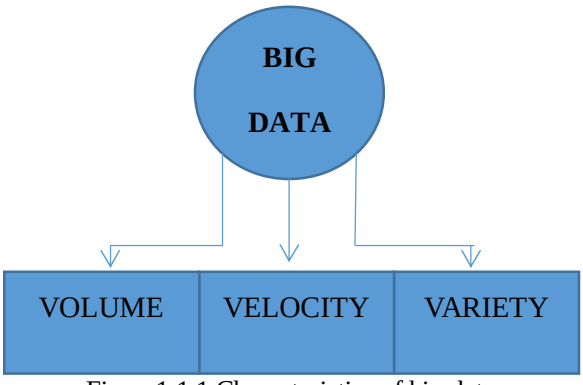

Figure1.1.1 Characteristics of big data

The three V's of big data is characterized as [4] [5].

- Volume: The quantity of generated data is important in this context. The size of the data determines the value and potential of the data under consideration, [4] and whatever it can actually be considered big data or not.
- Variety: the type [5] of content, and an essential fact that data analyst must know. This helps people who are associated with and analyse the data to effectively use the data to their advantage and thus uphold its importance.
- Velocity: In this context, the speed at which the data is generated and processed to meet the demands and the challenges [4] that lies in the path of growth and development. Which are processed and optimized to bring out the fourth V namely veracity.

Veracity: veracity defines that the data is being stored, and mined meaningful to the problem being analysed it refers to the noise and abnormality of data [36].

The characteristics of big data lead us to focus on the complexities that need to be handled with such a huge data.

1.1.2 Challenges of big data

Big Data has some inherent challenges and problems that can be primarily divided into three groups [6]: data, processing and management challenges at (fig 1.1.2) large amount of data usually face such challenges regarding volume, variety and velocity. As these big data characteristics are examined in scientific literature [8][9][10]. Volume refers to the large amount of data, especially, machine-generated. This characteristic defines a size of the data set that makes its storage and analysis problematic utilizing conventional database technology. Variety is related to different types and forms of data sources: structured (e.g. financial data) and unstructured (social media conversations, photos, videos, voice recordings and others).Multiplicity of the various data results in the issue of its handling. Velocity refers to the

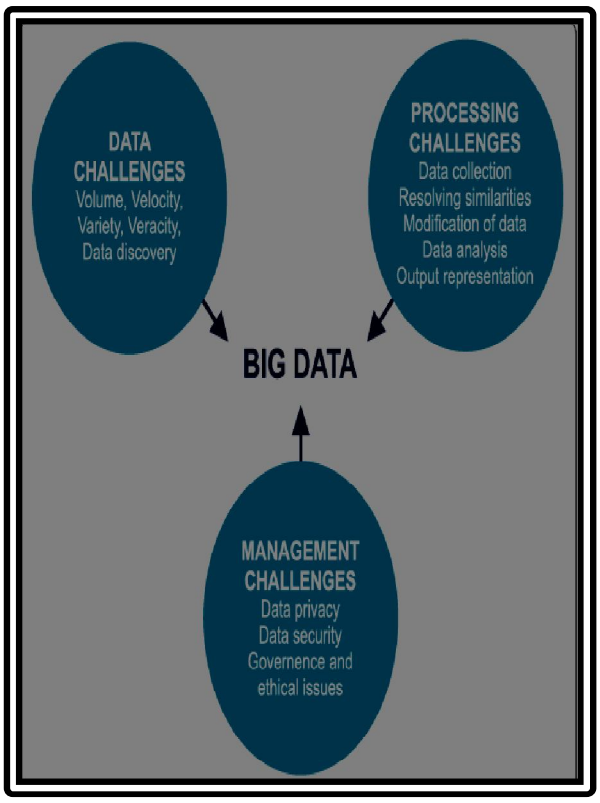

Figure 1.1.2 big data challenges [7].

speed of new data generation and distribution. This characteristic requires the implementation of real-

time processing for the streaming data analysis (e.g. on social media, different types of transactions or trading systems, etc.). Veracity refers to the complexity of data which may lead to a lack of quality and accuracy [9]. This characteristic reveals several challenges: uncertainty, imprecision, missing values, misstatement and data availability.

There is also a challenge regarding data discovery that is related to the search of high quality data in data sets. The second [8] branch of Big Data challenges is called processing challenges. It includes data collection, resolving similarities found in different sources, modification data to a type acceptable for the analysis, the analysis itself and output representation, i.e. the results visualization in a form most suitable for human perception. The last [10] type of challenge offered by this classification is related to data management.

Management challenges usually refer to secured data storage, its processing and collection.

Here the main focuses of study are: data privacy, its security, governance and ethical issues. Most of them are controlled based on policies and rules provided by information security institutes on state or international levels.

1.1.3 BIG DATA IN SOCIAL NETWORKS

In the last decades, the world has become connected to an extent that today people know the minute detail of what everyone is doing around them. Be it their close ones, their friends or even their favourite celebrities; all thanks to facebook and twitter. But in the process of being connected, everyone is constantly contributing to one common entity, "big Data". Big Data is the vast sets of information gathered by researchers at companies like facebook, Google and Microsoft from patterns of cell phones calls, text message and internet clicks by millions of users around the world. Companies often refuse to make such information public, sometimes to protect customer's privacy. The challenging part about big data is not the collection, but the management of the data. With the deceasing cost of storage (<$0.1 for 1 gigabytes). Anyone can gather data. The tough part is however, analysing the data and making sense of it. Data collected from social networks is highly unstructured and has to be cleaned before being used for any kind of analysis. After being cleaned, the big data can be used for various purposes ranging from making network graphs to study user activity, to helping big brands in knowing what the customer cares about. The worldwide average of time spent daily by a user on social media is two and a half hours, and information on their activity help social network deliver personalized content, and help advertiser hyper-target users.

Report that shows data each social network is collecting

- Facebook's interest/social graph: the world's largest online community collects more data via its API than any other social network. Facebok's "like" button is pressed 2.7 billion times every day across the web, revealing about what page care about.
- Google+'s relevance graph: the number of "+ 1s" a doyher google+ data are now a top factor in determining how a webpage ranks in Google search results.
- Linkedln's talent graph: At least 22% of Linedln users have firdt degree connections between 500-900 on the social network, and 19% have between 301-499. This rich professional's data is helping linkedln build a "talent graph".
- Twitter's new graph: at its peak late last year the social network was processing 143,199 tweets per second globally. This fire hose of tweets provide a real- time window into the news and information that people care about. Fifty two percent of twitter users in the U.S. consume news on the site (more than the percent who do so on facebook), according to pew.
- Youtube's entertainment graph: What music, shows, and celebrities do we like? YouTube reaches more U.S. adults aged 18 to 34 than any single cable network, according to Nielsen. YouTube knows what they like to watch.

2. TECHNIQUES TO PROCESS BIG DATA

2.1 MapReduce

MapReduce is the distributed compute component of Hadoop. MapReduce jobs are controlled by software known as the Job Tracker. A job is a full MapReduce program, including a complete execution of Map and Reduce tasks over a dataset. The MapReduce paradigm relies on master/slave architecture. Job Tracker runs on the master node and assigns Map and Reduce tasks to the slave nodes in the cluster. The slave nodes run software called the Task Tracker that is responsible for actually instantiating the Map or Reduce tasks and reporting the progress back to the Job Tracker. A MapReduce job consists of two parts, a map phase, which takes raw data and organizes it into key/value pairs, and a reduce phase which processes data in parallel [11].Mapreduce is the heart of hadoop. It is the programming paradigm that allows scalability across hundreds of servers in a hadoop cluster. Mapreduce is simple to understand for those who are familiar with clustered data processing solution. Mapreduce is the data processing

scheme used in hadoop that includes breaking the entire task into two parts known as mappers and reducers. At a high level stage, mappers read the data from HDFS process and generate intermediate result to the reducers. Reducers on the other hand are used to aggregate the intermediate result to generate the final output which is written again to HDFS [12]. A mapreduce map phase takes raw data and organise it into key/value pairs, and a reduce phase which processes data in parallel. The two phase maps and reduces the input to map reduce.

2.1.1 Mapreduce Architecture

The processing domain in the hadoop system is the mapreduce framework. The framework allows operation to be applied on the huge dataset, divide the problem and run them parallel. The architecture defines how the Mappers and reducers work (see at fig 1.2.1)

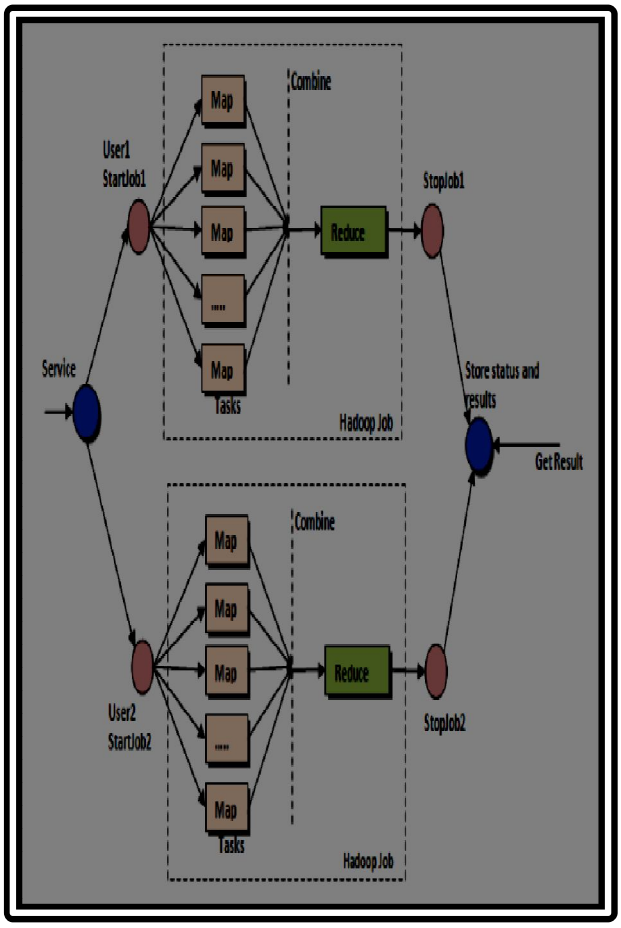

Figure 2.1.1 Map reduce Architecture [13]

Processing on mapreduce can be performed on data stored in a file system (unstructured) or in a database (structured). Mapreduce can also take advantage of locality of data, processing it on near the storage assets in order to reduce the distance over which the data is to be transmitted. Mapping and reducing is performed in three steps [13].

Map: node applies map () function is applied to the local data, and output is written to a temporary storage Shuffle: nodes redistribute data based on the output keys such that data belonging to the same key is located on the same node.Reduce: nodes now process each group of data, including per key in parallel.

2.1.2 Social Media and Mapreduce

Mapreduce systems plays important role in supporting big data analytics to understand the user trends in typical web services and social networking data. Based on our observations and analysis of Facebook production systems, we have characterized four requirements for the data placement structure: (1) fast data loading, (2) fast query processing, (3) highly efficient storage space utilization, and (4) strong adaptively to highly dynamic workload patterns [14]. A critical challenge in implementing data placement structure in mapreduce system is to address the above four requirements. In simple dataset three data placement structure has been widely used.

1) Horizontal row-store structure [15]
2) Vertical column-store structure [16] [17][18][19]
3) Hybrid PAX store [20][21]

These data placement techniques worked well with the common dataset. As big data holds data in large amount these data placement techniques carries some of the drawbacks according to the requirement of social media data.

RC file (record columnar) according to the need of social media data was introduced as it removed the drawback of the data placement techniques of common dataset RC File applies the concept of"first horizontally-partition, then vertically-partition" from PAX. It combines the advantages of both row-store and column-store. First, as row-store, RC File guarantees that data in the same row are located in the same node, thus it has low cost of tuple reconstruction. Second, as column-store, RC File can exploit a column-wise data compression and skip unnecessary column reads. It performs various steps as stated; Data layout and compression, data appending, data read and lazy compression and row group size [14].

Internet traffic measurement and analysis have been usually performed on a high performance server that collects and examines packet or flow traces. However, when we monitor a large volume of traffic

data for detailed statistics, a long period or a large-scale network, it is not easy to handle Tera or Peta-byte traffic data with a single server [23] for this purpose mapreduce flow based technique is used in order to remove the problem of single node failure as mapreduce provide us with the parallel distribution of data on different nodes. Flow data for a large-scale network, we need to handle and manage a few Tera or Peta-byte packet or flow files simultaneously. Google has first developed the MapReduce [24] programming model for page ranking or web log analysis. MapReduce is a software framework that supports distributed computing with two functions of map and reduce on large data sets on clusters. After Google announced the MapReduce model, Yahoo has released an open-source system for the cloud computing platform, called Hadoop [25].

2.2 HADOOP

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. We will focus on Hadoop MapReduce, which is the most popular open source implementation of the MapReduce framework proposed by Google [27]. The Hadoop MapReduce framework utilises a distributed file system to read and write its data. Typically, Hadoop MapReduce uses the Hadoop Distributed File Sys-tem (HDFS), which is the open source counterpart of the Google File System [28].

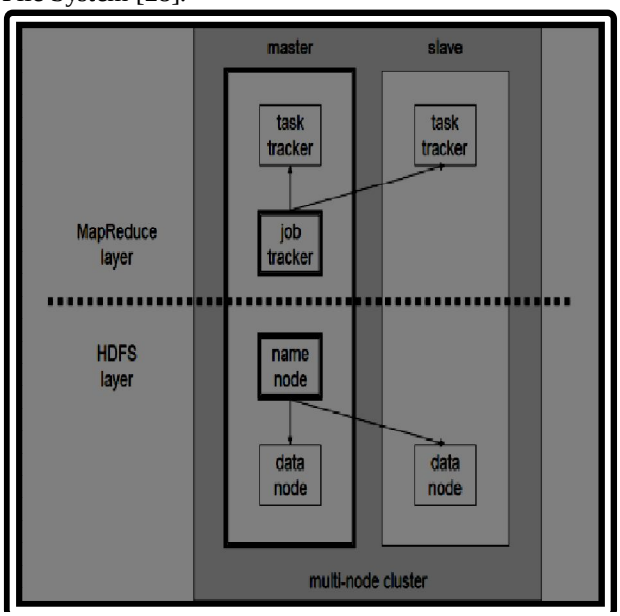

Figure 2.2 Hadoop Architecture [29]

Hadoop is a programming paradigm which supports processing of large datasets on distributed computing .Hadoop ecosystem consist of Mapreduce and HDFS layer (hadoop distributed file system) the above figure explains that HDFS [30] is a distributed file system designed to run on top of the local file system of the cluster node and store Extremely large files suitable for streaming data access. HDFS is highly fault tolerant and can scale up from a single server to thousands of machines, each offering local computation and storage .HDFS consists of two types of nodes, namely, a name node called “master” and several data nodes called “slaves.” HDFS can also include secondary name nodes. The name node manages the hierarchy of file systems and director namespace (i.e., metadata). MapReduce [31] is a simplified programming model for processing large numbers of datasets pioneered by Google for data-intensive applications. The MapReduce model was developed based on GFS [32] and is adopted through open-source Hadoop implementation, which was popularized by Yahoo. A platform the MapReduce framework, several other current open-source Apache projects are related to the Hadoop ecosystem, including Hive, Hbase, Mahout, Pig, Zookeeper, Spark, and Avro. Twister [33] provides support for efficient and iterative MapReduce computations. Hadoop schedules reduce tasks at requesting nodes without considering data locality leading to performance degradation. Locality-Aware Reduce Task Scheduler (LARTS), [34] a practical strategy for improving MapReduce performance. LARTS attempts to collocate reduce tasks with the maximum required data computed after recognizing input data network locations and sizes. LARTS adopts a cooperative paradigm seeking a good data locality while circumventing scheduling delay, scheduling skew, poor system utilization, and low degree of parallelism. Moving data repeatedly to distant nodes is becoming the bottleneck [35].To remove this bottleneck Locality- Aware Reduce Task Scheduler (LARTS), a practical strategy was proposed [34] that leverages network locations and sizes of partitions to exploit data locality. In particular, LARTS attempts to schedule reducers as close as possible to their maximum amount of input data and conservatively switches to a relaxation strategy seeking a balance between scheduling delay, scheduling skew, system utilization, and parallelism. Evaluations demonstrate LARTS’s outperformance over native Hadoop. In this [35] work we make the following contributions: We propose a novel strategy, LARTS, which applies data Locality to reduce task scheduling in MapReduce. We empirically analyse Hadoop’s erformance and network traffic. We observe that the process of interleaving the execution of map tasks with the shuffling of partitions employed by native Hadoop improves performance but increases network traffic.

We show how LARTS manages to maintain the advantage of the interleaving process besides diminishing network traffic. We [35][34] implemented LARTS in Hadoop 0.20.2 and conducted extensive experimentations to evaluate its potential. We [35] found that LARTS improves node-local, racklocal, and off-rack traffic by 34.45%, 0.32%, and 7.5%, on average, versus native Hadoop. In summary, LARTS outperforms native Hadoop by an average of 7%, and up to 11.6%.

3. CONCLUSION

There are lot of challenges in big data processing and analysis which leads to difficulties in the extraction of data. In this paper we reviewed big data and its techniques namely Mapreduce and Hadoop work as support to big data analysis. Mapreduce helps in solving the data placement problem of social networking site using RC file as well as provided a better flow analysis framework to remove the internet traffic. Hadoop is a programming paradiagm of mapreduce, moving data to distant nodes is a problem faced via hadoop, mapreduce to overcome this problem LARTS was proposed.

Future work can focus on processing of batch jobs to develop convient flow tool based on mapreduce and on the iterative data that is a problem of mapreduce so that convenient support can be given to analysis of big data while dealing with social sites.

REFERENCES

[1] P. Anderson, “what is web 2.0? Icha technologies and implications for education,” in JISC online report. Available at www JISC.ac.UK/media/documents/technologies ltsw07010b.pdf, 2007.

[2] Fact sheet: Big data across the federal government (2012).http//www.Whithouse.gov/sites/default/files/bidata_fact_sheet_3_29_2012.pdf.

[3] Demchenko, Y, De. Laat C, memberery P. Defining architecture components of the big data eco-system. In Providian’s of international conference on collaborations technologies and system (CTS), IEEE; 2014.PP 104-/2.

[4] Hilbert, Martin. “Big data for development: A review of Promises and challenges. Development Policy Review.” Martinhilbert.net. Retrieved 2015-10-07.

[5] Hilbert, M (2015). Digital Technology and social change [Open Online Course at the University of California] (freely available). http://www.youtube.com/watch?v=XRVIh47sA&index=51&list=PLtjBSCvVVCU3rNm46D3r85efN0hrzjuAlg Retrieved from https://canavas.instructure.cpm/course/949415.

[6] Akerkor R.Big data computing.Bora Raton, FL.CRC press, Taylor and Francis group; 2013.

[7] Visualizing BD with augmented and virtual reality:challenges and research agenda ekaterira Oeshamikova, Aleksadr Ometoo, Xevgeni koucheryavy and thomsas Olson.

[8] Kaisler S; armever F, Esperord JA, Money W. Big data: issues and challenges moving forward In: providing of 46th Hasidi international conference on sydtem scores CHICSS, IEEE, 2013 P. 995-1004.

[9] Tale A A, et.al. Big data challenges database system J. 2013; 4(3): 31-40.

[10] Chen M, Ma b S, Chang Y, Leung Vc. Big data: related technologies: challenges and future properties: Springer; 2014.

[11] Arantxa Duque Barrachima and Aisling O driscoll: A big data methodology for categorising technical support requirements using hadoop and Mahout.

[12] Dilpreet singh and Chandan K reddy: A survey on platform for big data analytics.

[13] Harshawardhan S. Bhosale, Prof. Devendra P. Godekekar: A review paper on Big data and Hadoop.

[14] Yongauang He, Rubao Lee, Yin Huai, Zheg shao, Namil Jain, Xiaoelong Zhang, Zhiwei Yu; RC file: A fast and space efficient data placement structure in Mapreduce based workhouse system.

[15] R. Ramakrishnan and J. Gehrke, “Database management system”, MC Grave-Hice, 2003.

[16] G.P Copeland and S Khoshafuon, “A decomposition storage model”, in SIGMOD conference, 1985, PP 268-279.

[17] M. Stonebraker, D.J. Abadi A Batkin, X. Chen, m. Cherneick, M Kerrerd, A cin, S Maddan, E J . O neil, P.E O ‘Neil A. Pasins, N Tran, and S.B zdonik, “C- store. A column –orientation DBMS”, in VLDB, 2005, PP 553-564.

[18] D.J abadi, S Maddan, and N. Hachen, "Column stores vs row stores: How different are they really?" in SIGMOD conference 2008.

[19] A.L. Holloway and D.J Witt, "dead optimized database, in depth," PVLDB, vol.1, no .1, PP. 502-513, 2008.

[20] A. Ailamake, D.J Dewitt, M.D Hill, and M. Skounakis, "weaving relation for cache performance ", in VLDB, 200', PP. 169-180.

[21] D. Tseriegianmis, s. Harizopoulies, M.A shalr, J.L. Wiever, and G. Graefe, "Query processing techniquies for solid states devices", in SIGMOD conference, 2009, PP 59-72.

[22] Yongauang He, Rubao Lee, Yin Huai, Zheg shao, Namil Jain, Xiaoelong Zhang, Zhiwei Yu; RC file: A fast and space efficient data placement structure in Mapreduce based workhouse system.

[23] Youngseak lee, Wonchulkeng, Hylongee son; an internet traffic analysis method with mapreduce.

[24] J. Dean and S.Ghemawat, Mapreduce: simplified Data processing on large cluster, OSDI, 2014.

[25] Hadoop, http://hadoop.apache.org/.

[26] Youngseak lee, Wonchulkeng, Hylongee son; An internet traffic analysis method with mapreduce

[27] J.Dean and S.Ghemawat, Mapreduce: a flexible data processing tool CACM, 53(1): 72-77, 2010.

[28] S.Ghemwat, H. Gobioff, and S-T . Leung. The google file system .In SOSP, pages 29-43, 2003.

[29] Harshawardhan S. Bosale, Prof. Devendra, P. Gadikar. A Review paper on Big data and Hadoop.

[30] K, shvachko, K. Haronz, S .Radia, R. Chansler, The Hadoop Distributed File System, Mass storage systems and Technologies(MSST), 2010 IEEE 26th Symporuenson, 2010, PP-1-10.

[31] J.Dean and S.Ghemawat, Mapreduce:Simplified Data processing on large clusters, Commun, ACH S1(2008) 107-113.

[32] S.Ghemwat, H. Gobioff, and S-T. Leung. The google file system, ACM SIGOPS oper. Syst, Rev. ACM 37(5)/2003/29-43.

[33] J. Ekanayake, H. Li, B. Zhang, t. Gunarathne, S-H, bae, J, Qui, G, Fox, Twister: A runtime for iterative mapreduce, in : proceedings of the 19th ACM international Symporeoium on high performance distributed computing, ACM, 2010, PP. 810-818.

[34] Mohmmad Hammoud and MajdF. Sakr; Locality –Aware Reduce Task for Scheduling.

[35] A. Szalay, A. Bunr, J . Greey and I, Rauill, "The importance of data locality in distributed computing applications," NSF Workflow work-shop, 2006.

[36] http://insidebigdata.com/2013/09/12/beyond-volume-velocity-issue-big-data-veracity.

[37] Dhar, V. (2013)." Data Science and predictive". Communications of the ACM56 (12):64.doi; 10.114512500499.

[38] Jeff Luk (2015-12-12). "The Keyword in "Data Science" is not Data, it is Science"Simply statistics.

[39] N gugen, Thomoson. "Data Scientists vs. data Analyst: why the distinction matters" retrieved 2 Oct 2015.