Technologies to Handle Big Data: A Survey

Sabia[1] and Love Arora[2]
[1,2]*Department of Computer Science & Engineering,*
Guru Nanak Dev University, Regional Campus, Jalandhar, India
E-mail: [1]sabiajal@gmail.com, [2]aroralove7@gmail.com

***Abstract*—Big data came into existence when the traditional relational database systems were not able to handle the unstructured data (weblogs, videos, photos, social updates, human behaviour) generated today by organisation, social media, or from any other data generating source. Data that is so large in volume, so diverse in variety or moving with such velocity is called Big data. Analyzing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, Apache Hive, No SQL and HPCC. These technologies handle massive amount of data in MB, PB, YB, ZB, KB and TB. In this research paper various technologies for handling big data along with the advantages and disadvantages of each technology for catering the problems in hand to deal the massive data has discussed.**

Keywords: Big Data, Hadoop, Map Reduce, Apache Hive, No SQL

I. INTRODUCTION

With the growth of technological development and services, the large amount of data is formed that can be structured and unstructured from the different sources in different domains. Massive data of such sort is very difficult to process that contains the information of the records of million people that includes everyday massive amount of data from social sites, cell phones GPS signals, videos etc. Big data is a largest buzz phrases in domain of IT, new technologies of personal communication driving the big data new trend and internet population grew day by day but it never reach by 100%. The need of big data generated from the large companies like facebook, yahoo, Google, YouTube etc for the purpose of analysis of enormous amount of data which is in unstructured form or even in structured form. Google contains the large amount of information. So; there is the need of Big Data Analytics that is the processing of the complex and massive datasets This data is different from structured data (which is stored in relational database systems) in terms of five parameters –variety, volume, value, veracity and velocity (5V's). The five V's (volume, variety, velocity, value, veracity) are the challenges of big data management are [1]:

1. **Volume**: Data is ever-growing day by day of all types ever MB, PB, YB, ZB, KB, TB of information. The data results into large files. Excessive volume of data is main issue of storage. This main issue is resolved by reducing storage cost. Data volumes are expected to grow 50 times by 2020.
2. **Variety**: Data sources (even in the same field or in distinct) are extremely heterogeneous [1]. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more. The varieties are endless, and the data enters the network without having been quantified or qualified in any way.
3. **Velocity**: The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive. Some organisations data velocity is main challenge. The social media messages and credit card transactions done in millisecond and data generated by this putting in to databases.
4. **Value**: Which addresses the need for valuation of enterprise data? It is a most important v in big data. Value is main buzz for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.
5. **Veracity**: The increase in the range of values typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data. Big data and analytics technologies work with these types of data.

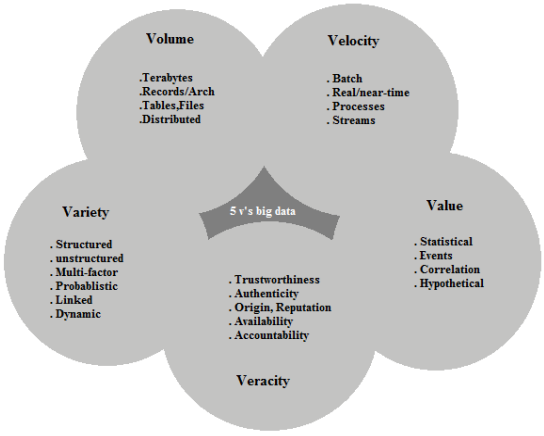

Fig. 1 Parameters of Big Data

Huge volume of data (both structured and unstructured) is management by organization, administration and governance. Unstructured data is a data that is not present in a database. Unstructured data may be text, verbal data or in another form. Textual unstructured data is like power point presentation, email messages, word documents, and instant massages. Data in another format can be.jpg images, .png images, audio files (.mp3, .wav, .aiff) and video files that can be in

flash format, .mkv format or .3gp format. According to the "IDC Enterprise Disk Storage Consumption Model" report released in year 2009, in which the transactional data is proposed to raise at a composite yearly growth rate (CAGR) of 21.8%, it's far outpaced by a 61.7% CAGR calculation for unstructured data [3]. From last twenty years, the data is mounting day by day across the world in every domain. Some distinct facts about the data are, there are about 277,000 tweets per minute, 2 million queries approximately on Google every minute in all domains, 75 hours of new videos in different formats are uploaded to YouTube, More than 100 million emails are sent via Gmail, yahoo, rediff mail and many more, 350 GB of data is dealing out on facebook every day and more than 576 websites are created every minute. During the year 2012, 2.5 quintillion bytes of data were created every day. Big data and its depth analysis is the core of modern science, research area and business areas. Huge amount of data is generated from the distinct various sources either in structure or unstructured form. Such form of data stored in databases and then it become very complex to extract, transform and make in use [8]. IBM indicates that 2.5 Exabyte data is created everyday which is very difficult to analyze in various aspects. The estimation about the generated data is that till year 2003 it was represented about 5 Exabyte, then until year 2012 is 2.7 Zettabyte and till 2015 it is expected to boost up to 3 times [10].

This paper is organised as follows. In section II literature survey have been described along with advantages and disadvantages of the paper. In section III the various big data techniques has been discussed. Future Scope has been discussed in section IV for direction to emerging researchers and Final section gives a conclusion of the paper.

II. LITERATURE SURVEY

1. John A. Keane [2] in 2013 proposed a framework in which big data applications can be developed. The framework consist of three stages (multiple data sources, data analysis and modelling, data organization and interpretation) and seven layers (visualisation/presentation layer, service/query/access layer, modelling/ statistical layer, processing layer, system layer, data layer/multi model) to divide big data application into blocks. The main motive of this paper is to manage and architect a massive amount of big data applications. The advantage of this paper is big data handles heterogeneous data and data sources in timely to get high performance and Framework Bridge the gap with business needs and technical realities. The disadvantage of this paper is too difficult to integrate existing data and systems.
2. Xin Luna Dong [5] in 2013 explained challenges of big data integration (schema mapping, record linkage and data fusion). These challenges are explained by using examples and techniques for data integration in addressing the new challenges raised by big data, includes volume and number of sources, velocity, variety and veracity. The advantage of this paper is identifying the data source problems to integrate existing data and systems. The disadvantage of this paper is big data integration such as integrating data from markets, integrating crowd sourcing data, providing an exploration tool for data sources.
3. Jun Wang [17] in 2013 proposed the Data-g Rouping-Aware (DRAW) data placement scheme to improve the problems like performance, efficiency, execution and latency. It could cluster many grouped data into a small number of nodes as compared to map reduce/hadoop. the three main phases of DRAW defined in this paper are: cluster the data-grouping matrix, learning data grouping information from system logs and recognizing the grouping data. The advantage of the paper is improve the throughput up to 59.8%, reduce the execution time up to 41.7% and improve the overall performance by 36.4% over the Hadoop/map reduce.
4. Yaxiong Zhao [7] in 2014 proposed data aware caching (Dache) framework that made minimum change to the original map reduce programming model to increment processing for big data applications using the map reduce model. It is a protocol, data aware cache description scheme and architecture. The advantage of this paper is, it improves the completion time of map reduce jobs.
5. Jian Tan [6] in 2013 author talks about the theoretical assumptions, that improves the performance of Hadoop/map reduce and purposed the optimal reduce task assignment schemes that minimize the fetching cost per job and performs the both simulation and real system deployment with experimental evolution. The advantage of this paper is improves the performance of large scale Hadoop clusters. The disadvantage of this paper is environmental factors such as network topologies effect on a reduce task in map reduce clusters.
6. Thuy D. Nguyen [4] (2013) author solve the multilevel secure (MLS) environmental problems of Hadoop by using security enhanced Linux (SE Linux) protocol. In which multiple sources of Hadoop applications run at different levels. This protocol is an extension of Hadoop distributed file system (HDFS).
   The advantage of this paper is solving environmental problems without requiring complex Hadoop server components.
7. Keith C.C. Chan [15] 2013 author describes large amount of structured and unstructured data

collection, processing and analysis from hospitals, laboratories, pharmaceutical, companies or even social media and also discus about how to collect or analyse huge volume of data for drug discovery. The advantage of this paper is how big data analytics contributes to better drug safety efficacy for pharmaceutical regulators and companies. The disadvantage of this paper it needs the algorithms that are simple, scalable, efficient and effective for data discovery process.

8. Sagiroglu, S. [8] (2013) offered the big data content, its scope, functionality, data samples, advantages and disadvantages along with challenges of big data. The critical issue in relation to the Big data is the privacy and protection. Big data samples describe the review about the environment, science and research in biological area. By this paper, we can conclude that any association in any domain having big data can take the benefit from its careful investigation for the problem solving principle. Using Knowledge Discovery from the Big data convenient to get the information from the complicated data records.
   The overall appraisal describe that the data is mounting day by day and becoming complex. The challenge is not only to gather and handle the data but also how to extract the useful information from that collected data records. In accordance to the Intel IT Center, there are several challenges related to Big Data which are rapid data growth, data infrastructure, and variety of data, visualization and data velocity.
9. Garlasu, D. [10] (2013) discussed the enhancement about the storage capabilities, the processing power along with handling technique. The Hadoop technology is widely used for the simulation purpose. Grid Computing provides the notion of distributed computing using HDFS. The benefit of Grid computing is the maximum storage capability and the high processing power. Grid Computing makes the big assistance among the scientific research and help the researcher to analyze and store the large and complex data in various formats.
10. Mukherjee, A. [11] (2012) The Big data analysis define the large amount of data to retrieve the useful information and uncover the hidden information. Big data analytics refers to the Map Reduce Framework which is discovered by the Google. Apache Hadoop is the open source platform which is used for the purpose of simulation of Map Reduce Model. In this the performance of SF-CFS is compared with the HDFS with the help of the SWIM by the facebook job traces. SWIM contains the workloads of thousands of jobs with complex and massive data arrival and computation patterns.
11. Aditya B. [12] (2012) defines big data Problem using Hadoop and Map Reduce" reports the experimental research on the Big data problems in various domains. It describe the optimal and efficient solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage data and Map Reduce framework for parallel processing to process massive data sets and records.

III. BIG DATA TECHNOLOGIES

Big data is a new concept for handling massive data therefore the architectural description of this technology is very new. There are the different technologies which use almost same approach i.e. to distribute the data among various local agents and reduce the load of the main server so that traffic can be avoided. There are endless articles, books and periodicals that describe Big Data from a technology perspective so we will instead focus our efforts here on setting out some basic principles and the minimum technology foundation to help relate Big Data to the broader IM domain.

A. Hadoop

Hadoop is a framework that can run applications on systems with thousands of nodes and terabytes. It distributes the file among the nodes and allows to system continue work in case of a node failure. This approach reduces the risk of catastrophic system failure. In which application is broken into smaller parts (fragments or blocks).Apache Hadoop consists of the Hadoop kernel, Hadoop distributed file system (HDFS), map reduce and related projects are zookeeper, Hbase, Apache Hive. Hadoop Distributed File System (HDFS) consists of three Components: the Name Node, Secondary Name Node and Data Node [15]. The multilevel secure (MLS) environmental problems of Hadoop by using security enhanced Linux (SE Linux) protocol. In which multiple sources of Hadoop applications run at different levels. This protocol is an extension of Hadoop distributed file system (HDFS) [12]. Hadoop is commonly used for distributed batch index building; it is desirable to optimize the index capability in near real time. Hadoop provides components for storage and analysis for large scale processing [1]. Now a day's Hadoop used by hundreds of companies.

The advantage of Hadoop is Distributed storage & Computational capabilities, extremely scalable,

optimized for high throughput, large block sizes, tolerant of software and hardware failure.

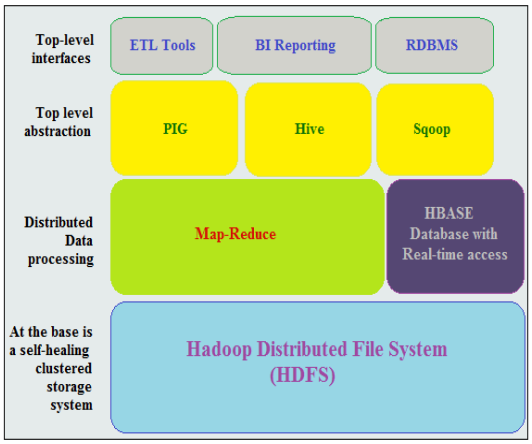

Fig. 2 Architecture of Hadoop

The disadvantage of Hadoop is that it is master processes are single points of failure. Hadoop does not offer storage or network level encryption, inefficient for handling small files.

Components of Hadoop [8]:

- **HBase**: It is open source, distributed and Non-relational database system implemented in Java. It runs above the layer of HDFS. It can serve the input and output for the Map Reduce in well mannered structure.
- **Oozie**: Oozie is a web-application that runs in a java servlet. Oozie use the database to gather the information of Workflow which is a collection of actions. It manages the Hadoop jobs in a mannered way.
- **Sqoop**: Sqoop is a command-line interface application that provides platform which is used for converting data from relational databases and Hadoop or vice versa.
- **Avro**: It is a system that provides functionality of data serialization and service of data exchange. It is basically used in Apache Hadoop. These services can be used together as well as independently according the data records.
- **Chukwa:** Chukwa is a framework that is used for data collection and analysis to process and analyze the massive amount of logs. It is built on the upper layer of the HDFS and Map Reduce framework.
- **Pig**: Pig is high-level platform where the Map Reduce framework is created which is used with Hadoop platform. It is a high level data processing system where the data records are analyzed that occurs in high level language.
- **Zookeeper**: It is a centralization based service that provides distributed synchronization and provides group services along with maintenance of the configuration information and records.
- **Hive**: It is application developed for data warehouse that provides the SQL interface as well as relational model. Hive infrastructure is built on the top layer of Hadoop that help in providing conclusion, and analysis for respective queries.

B. Map Reduce

Map-Reduce was introduced by Google in order to process and store large datasets on commodity hardware. Map Reduce is a model for processing large-scale data records in clusters. The Map Reduce programming model is based on two functions which are map() function and reduce() function. Users can simulate their own processing logics having well defined map() and reduce() functions. Map function performs the task as the master node takes the input, divide into smaller sub modules and distribute into slave nodes. A slave node further divides the sub modules again that lead to the hierarchical tree structure. The slave node processes the base problem and passes the result back to the master Node. The Map Reduce system arrange together all intermediate pairs based on the intermediate keys and refer them to reduce() function for producing the final output. Reduce function works as the master node collects the results from all the sub problems and combines them together to form the output [19].

Map(in_key,in_value)---
>list(out_key,intermediate_value)

Reduce(out_key,list(intermediate_value))---
>list(out_value)

The parameters of map () and reduce () function is as follows:

map (k1,v1) ! list (k2,v2) and reduce (k2,list(v2)) ! list (v2)

A Map Reduce framework is based on a master-slave architecture where one master node handles a number of slave nodes [18]. Map Reduce works by first dividing the input data set into even-sized data blocks for equal load distribution. Each data block is then assigned to one slave node and is processed by a map task and result is generated. The slave node interrupts the master node when it is idle. The scheduler then assigns new tasks to the slave node. The scheduler takes data locality and resources into consideration when it disseminates data blocks.

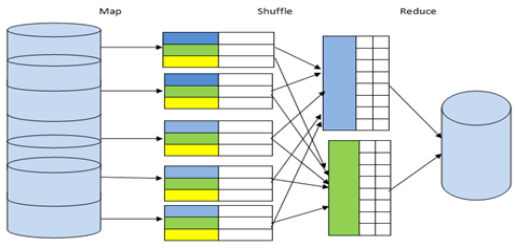

Fig. 3 Architecture of Map Reduce

Figure 3 shows the Map Reduce Architecture and Working. It always manages to allocate a local data block to a slave node. If the effort fails, the scheduler will assign a rack-local or random data block to the slave node instead of local data block. When map() function complete its task, the runtime system gather all intermediate pairs and launches a set of condense tasks to produce the final output. Large scale data processing is a difficult task, managing hundreds or thousands of processors and managing parallelization and distributed environments makes is more difficult. Map Reduce provides solution to the mentioned issues, as is supports distributed and parallel I/O scheduling, it is fault tolerant and supports scalability and it has inbuilt processes for status and monitoring of heterogeneous and large datasets as in Big Data [18]. It is way of approaching and solving a given problem. Using Map Reduce framework the efficiency and the time to retrieve the data is quite manageable. To address the volume aspect, new techniques have been proposed to enable parallel processing using Map Reduce framework [13]. Data aware caching (Dache) framework that made slight change to the original map reduce programming model and framework to enhance processing for big data applications using the map reduce model [16].

The advantage of map reduce is a large variety of problems are easily expressible as Map reduce computations and cluster of machines handle thousands of nodes and fault-tolerance.

The disadvantage of map reduce is Real-time processing, not always very easy to implement, shuffling of data, batch processing.

Map Reduce Components:

1. **Name Node**: manages HDFS metadata, doesn't deal with files directly.
2. **Data Node**: stores blocks of HDFS—default replication level for each block: 3.
3. **Job Tracker**: schedules, allocates and monitors job execution on slaves—Task Trackers.
4. **Task Tracker**: runs Map Reduce operations.

C. *Hive*

Hive is a distributed agent platform, a decentralized system for building applications by networking local system resources [8]. Apache Hive data warehousing component, an element of cloud-based Hadoop ecosystem which offers a query language called HiveQL that translates SQL-like queries into Map Reduce jobs automatically. Applications of apache hive are SQL, oracle, IBM DB2. Architecture is divided into Map-Reduce-oriented execution, Meta data information for data storage, and an execution part that receives a query from user or applications for execution.

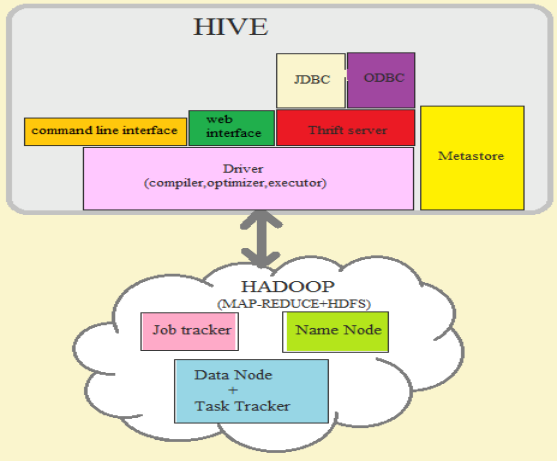

Fig. 4 Architecture of HIVE

The advantage of hive is more secure and implementations are good and well tuned.

The disadvantage of hive is only for ad hoc queries and performance is less as compared to pig.

D. *No-SQL*

No-SQL database is an approach to data management and data design that's useful for very large sets of distributed data. These databases are in general part of the real-time events that are detected in process deployed to inbound channels but can also be seen as an enabling technology following analytical capabilities such as relative search applications. These are only made feasible because of the elastic nature of the No-SQL model where the dimensionality of a query is evolved from the data in scope and domain rather than being fixed by the developer in advance. It is useful when enterprise need to access huge amount of unstructured data. There are more than one hundred No SQL approaches that specialize in management of different multimodal data types (from structured to non-structured) and with the aim to solve very specific challenges [5]. Data Scientist, Researchers and Business Analysts in specific pay more attention to agile approach that leads to prior insights into the data sets that may be concealed or constrained with a more formal development process. The most popular No-SQL database is Apache Cassandra.

The advantage of No-SQL is open source, Horizontal scalability, Easy to use, store complex data types, Very fast for adding new data and for simple operations/queries. The disadvantage of No-SQL is Immaturity, No indexing support, No ACID, Complex consistency models, Absence of standardization.

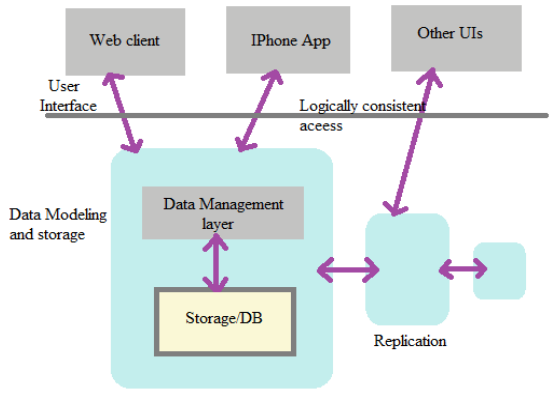

Fig. 5 Architecture of No SQL

E. *HPCC*

HPCC is an open source platform used for computing and that provides the service for handling of massive big data workflow. HPCC data model is defined by the user end according to the requirements. HPCC system is proposed and then further designed to manage the most complex and data-intensive analytical related problems. HPCC system is a single platform having a single architecture and a single programming language used for the data simulation.HPCC system was designed to analyze the gigantic amount of data for the purpose of solving complex problem of big data. HPCC system is based on enterprise control language which has the declarative and on-procedural nature programming language the main components of HPCC are:

- *HPCC Data Refinery*: Use parallel ETL engine mostly.
- *HPCC Data Delivery*: It is massively based on structured query engine used.
- Enterprise Control Language distributes the workload between the nodes in appropriate even load.

IV. Future Scope

There is nothing concealed that big data significantly influencing IT companies and through development new technologies only we can handle it in a managerial way. Big data totally change the way of organizations, government and academic institution by using number of tools to make the management of big data. In future Hadoop and NoSQL database will be highly in demand moving forward. The amount of data produced by organizations in next five years will be larger than last 5,000 years. In the upcoming years cloud will play the important role for private sectors and organisations to handle the big data efficiently.

V. Conclusion

In this paper we have surveyed various technologies to handle the big data and there architectures. In this paper we have also discussed the challenges of Big data (volume, variety, velocity, value, veracity) and various advantages and a disadvantage of these technologies. This paper discussed an architecture using Hadoop HDFS distributed data storage, real-time NoSQL databases, and MapReduce distributed data processing over a cluster of commodity servers. The main goal of our paper was to make a survey of various big data handling techniques those handle a massive amount of data from different sources and improves overall performance of systems.

References

[1] Yuri Demchenko "The Big Data Architecture Framework (BDAF)" Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.
[2] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), "Big Data Framework" 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499.
[3] Margaret Rouse, April 2010 "unstructured data".
[4] Nguyen T.D., Gondree M.A., Khosalim, J.; Irvine, "Towards a Cross Domain MapReduce Framework" IEEE C.E. Military Communications Conference, MILCOM 2013, 1436 – 1441
[5] Dong, X.L.; Srivastava, D. Data Engineering (ICDE)," Big data integration" IEEE International Conference on , 29(2013) 1245–1248
[6] Jian Tan; Shicong Meng; Xiaoqiao Meng; Li ZhangINFOCOM, "Improving ReduceTask data locality for sequential MapReduce" 2013 Proceedings IEEE ,1627 - 1635
[7] Yaxiong Zhao; Jie Wu INFOCOM, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework" 2013 Proceedings IEEE 2013, 35 - 39 (Volume 19)
[8] Sagiroglu, S.; Sinanc, D.,"Big Data: A Review",2013,20-24
[9] Minar, N.; Gray, M.; Roup, O.; Krikorian, R.; Maes, "Hive: distributed agents for networking things" IEEE CONFERENCE PUBLICATIONS 1999 (118-129)
[10] Garlasu, D.; Sandulescu, V.; Halcu, I.; Neculoiu, G,"A Big Data implementation based on Grid Computing", Grid Computing, 2013, 17-19
[11] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, "Shared disk big data analytics with Apache Hadoop", 2012, 18-22
[12] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", 2012, 6-8
[13] Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1,2010, 72-77.
[14] Chan,K.C.C. Bioinformatics and Biomedicine (BIBM), "Big data analytics for drug discovery" IEEE International Conference on Bioinformatics and Biomedicine 2013,1.
[15] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013.
[16] Wang, J.; Xiao, Q.; Yin, J.; Shang, P. Magnetics, "DRAW: A New Data-gRouping-AWare Data Placement Scheme for Data Intensive Applications With Interest Locality"IEEE Transactions (Vol: 49), 2013, 2514 – 2520
[17] HADOOP-3759: Provide ability to run memory intensive jobs without affecting other running tasks on the nodes.