Implementation of Big Data: A Comparative Study

Neetu Sharma[1], Dr. Ashish Sharma[2]
Department of Computer Science and Engineering
Jodhpur National University, Jodhpur, India
[1]neetusmart30@gmail.com
[2]aashishid@gmail.com

Abstract

The term big data refers to a large amount of data which is present over the internet. The size of this data is so large that it cannot be stored, handled, analyzed and processed by the traditional database systems. This is because when size of data increases, then its complexity also increases. There are various sources which are responsible for generation of Big Data such as news media, social networking sites, business applications and many more. Efficient management, proper storage, availability, scalability and processing are some of the issues that creates some problems while dealing with Big Data. Thus to handle this big data, new techniques, tools and architecture is required. This paper discusses some tools and techniques that are commonly used for handling Big Data as they overcome the traditional difficulties and opens a new way for the researchers to draw their attention towards the tools which can be best chosen for maintenance of Big Data.

Keywords*: Big Data, Hadoop, Map Reduce, HDFS, Grid Computing, Big Table.*

I. INTRODUCTION

"Big Data" in today's Information Technology world is a very hot and interesting topic to discuss about. We are living in an era where everything is getting digitized whether it is an image, text, audio, video, etc. in other words we can say that we are living in a digital world. Most of the data is produced, stored, exchanged and processed over the internet, leading to the increase in size of data every day. This large amount of data present over the internet is referred to as "Big Data".

Earlier the size of data was referred in megabytes and gigabytes, but now it is referred in petabytes and zettabytes. Big Data can be characterized by 3 V's introduced by Gartner analyst, Doug Laney [1] – Volume, Velocity and Variety.

Volume: Volume refers to the size of data over the internet. It is currently in petabytes and is expected to be raised to zettabytes.

Velocity: Velocity of data refers to the speed of data generation and data processing. For example the data from the sensor devices would be continuously moving to the database store and this quantity would not be small enough [2].

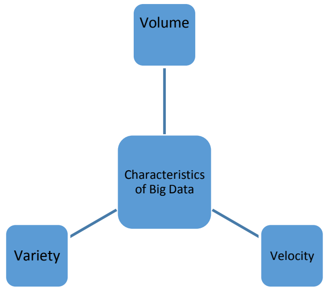

Figure 1: Characteristics of Big Data

Variety: Variety refers to the type of data. Data can be structured (text), unstructured (data generated from social networking sites and sensors) or semi-structured (data from web pages, web logs-mail etc).

Two more characteristics have also been included- Veracity and Value.

Veracity – It means how much the data is related to truth or facts.

Value- It refers to the processing of data and how the data can be combined with other data to extract meaningful information from it.

II. RELATED WORK

Big data is an emerging technology which aims to store, analyze, manage and visualize the fast growing data in various applications like business applications, education, sports, news media, social networking sites, etc. the data can be structured or unstructured. When structured and unstructured data are compared with each other, the unstructured data provides a better reflection of reality for making important decisions [3] [4]. A lot of research has been done in the field of big data. Many researchers are still finding better tools and technologies to implement big data. The only aim is to overcome all the issues and challenges that are creating obstacles in the path of widespread use of big data all over the world. Research work of several researchers is discussed below:

S. Vikram Phaneendra & E. Madhusudhan Reddy et al. Illustrated that in olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as "big data". In this they told that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the Hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems. Hadoop architecture handle large data sets, scalable algorithm does log management application of big data can be found out in financial, retail industry, health-care, mobility, insurance. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - dataprivacy, search analysis, etc [5].
Kiran kumara Reddi & Dnvsl Indira et al. Enhanced us with the knowledge that Big Data is combination of structured, semi-structured ,unstructured homogenous and heterogeneous data .The author suggested to use nice model to handle transfer of huge amount of data over the network .Under this model, these transfers are relegated to low demand periods where there is ample,idle bandwidth available . This bandwidth can then be repurposed for big data transmission without impacting other users in system. The Nice model uses a store –and- forward approach by utilizing staging servers. The model isable to accommodatedifferences in time zones and variations in bandwidth. They suggested that new algorithms are required to transfer big data and to solve issues like security, compression, routing algorithms [6].
Jimmy Lin et al. used Hadoop which is currently the large –scale data analysis " hammer" of choice, but there exists classes of algorithms that aren't " nails" in the sense that they are not particularly amenable to the MapReduce programming model . He focuses on the simple solution to find alternative non-iterative algorithms that solves the same problem. The standard MapReduce is well known and described in many places .Each iteration of the PageRank corresponds to the MapReduce job. The author suggested iterative graph, gradient descent & EM iteration which is typically implemented as Hadoop job with driven set up iteration &Check for convergences. The author suggests that if all you have is a hammer, throw away everything that's not a nail [7].

Wei Fan & Albert Bifet et al. Introduced Big Data Mining as the capability of extracting Useful information from these large datasets or streams of data that due to its Volume, variability and velocity it was not possible before to do it. The author also started that there are certain controversy about Big Data. There certain tools for processes. Big Data as such Hadoop, Strom, apache S4. Specific tools for big graph mining were PEGASUS& Graph. There are certain Challenges that need to death with as such compression, visualization etc [8].
Albert Bifet et al. Stated that streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge, allowing organizations to react quickly when problem appear or detect to improve performance. Huge amount of data is created everyday termed as "big data". The tools used for mining big data are apache Hadoop, apache big, cascading, scribe, storm, apache Hbase, apache mahout, MOA, R, etc. Thus, he instructed that our ability to handle many Exabyte of data mainly dependent on existence of rich variety dataset, technique, software framework [9].

III. PROPOSED WORK

Big data is a very essential requirement in today's world. Several issues and challenges are also associated with big data such as fault tolerance, scalability, heterogeneity, privacy, security, etc. that need to be overcome while dealing with big data. Several tools and techniques are available to overcome these issues and challenges. Big data provides a new method to traditional data analysis, which has a variety of technologies, including Hadoop, MapReduce, Grid Computing, Cloud Computing, Big Table, and so on[8]. This paper focusses on the following technologies:

A. Hadoop – Hadoop is a programming framework developed by Google's MapReduce which is used to process large datasets by breaking down them into various parts. The current apache Hadoop ecosystem consists of the Hadoop kernel, MapReduce, HDFS, and many more. HDFS and MapReduce are explained below[10]:

a) HDFS architecture –HDFS stands for Hadoop Distributed File System. It is an essential component of Hadoop which is used to store huge datasets. The main task of HDFS is to distribute the data to various clusters of computers (machines) and then processing of this data is done. The advantage of using HDFS is that it coordinates the work among machines and if any one of them fails, Hadoop continues to operate by shifting the work from one machine to another without losing data or interrupting work [11].

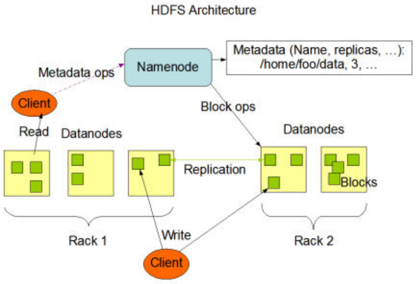

Figure 2: Architecture of HDFS

b) MapReduce – MapReduce is a parallel programming framework that allows operations to be applied over large datasets. The main task of MapReduce is to divide the problem into smaller parts and then run those subparts in a parallel fashion. MapReduce consists of two functions: Map and Reduce.
Map: This function generates a key/value pair and performs sorting and filtering of data.
Reduce: This function combines all the intermediate values and gives the output.

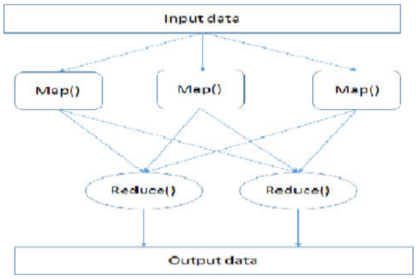

Figure 3: Architecture of MapReduce

B. Grid Computing- A grid is a system in which a number of servers are connected to each other through a high speed internet. It is a distributed computing model in which the servers are geographically apart from each other and the users can access the data transparently from any location. Although Grid is beneficial as it provides hardware for storage of data but it has a drawback that current Grid infrastructure is not capable enough to handle Big Data. Thus research is still going on to find a solution to this problem so that it can deal with large volume of data.

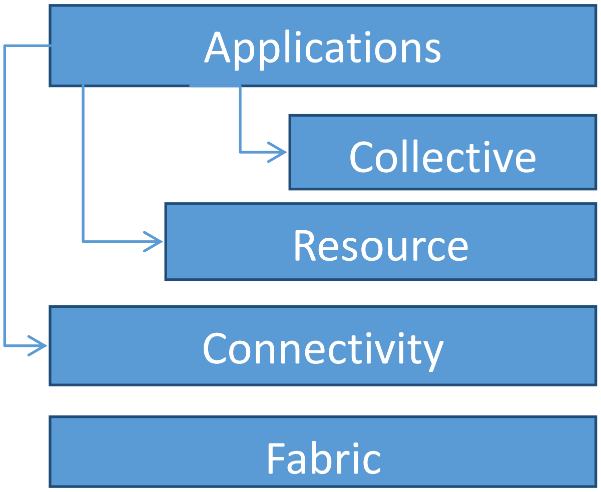

Figure 4: Architecture of Grid Computing

Challenges with grid computing [12]:

- Data movement
- Data replication
- Resource management
- Job submission

C. Bigtable – Bigtable is a distributed storage system developed by Google for managing very large size data. It can handle data up to petabytes. It is a distributed and sparse map which is indexed by a row key, column key, and a timestamp; each value in the map is an uninterpreted array of bytes.
(row:string, column: string, time:int64)→string

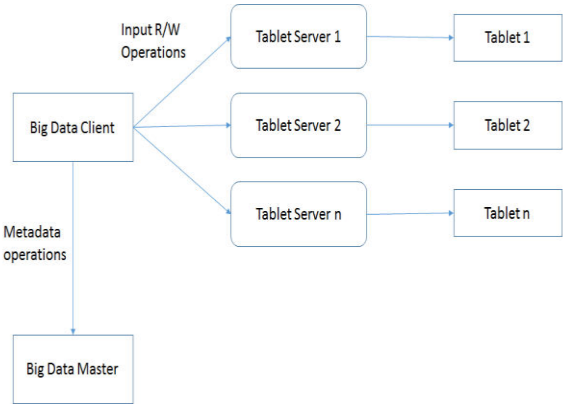

Figure 5: Architecture of Bigtable

Challenges with Bigtable implementation are:

- Reliability storage: durability and replication
- Sequential storage
- Structured storage

Table 1: Comparison between HDFS, MapReduce, Grid Computing and Bigtable

Characteristics	HDFS	MapReduce	Grid Computing	Bigtable
Processing model	Distributed processing	Parallel processing; batch processing	Distributed processing	Distributed processing
Storage space	Distributed storage	Clusters of computers	Distributed storage	Distributed storage
Scalability	Highly scalable as more nodes can be added easily	Scalable	Less scalable than Hadoop	Scalable; number of rows and columns can be added
Type of data	Heterogeneous i.e. structured, unstructured and semi-structured data	Unstructured data		Structured data
Cost	Hadoop is not very expensive as it runs on cluster of commodity hardware [13].	Inexpensive	Inexpensive access to data	Join operations are less costly because of the denormalization - Replication/distribution of data is less costly because of data independence [14]
Flexibility	Flexible i.e. it can store both structured and unstructured data	Flexible	Highly flexible	Highly flexible
Processing speed	The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing.	High processing speed than grid computing	High processing speed	High processing speed
Fault tolerance	Highly fault tolerant	Highly fault tolerant		

IV. CONCLUSION AND FUTURE SCOPE

Due to large increase in the size of data in various fields, it has been noticed that information overload problems prevail everywhere. Many other technical challenges mentioned in this paper must be taken into consideration. The challenges include not only just the scalability, but also heterogeneity, privacy timeliness and security at all levels of data interpretation. Big data is presently implemented using Hadoop. Hadoop is an open source software used for processing of big data. However the increasing amount of data is making Hadoop insufficient to handle data. Big data is a technology that has a great scope and important role in future. To take out the best benefit from Hadoop, extensive research needs to be carried out and revolutionary tools and techniques needs to be developed to carefully comprehend and correctly respond to various challenges.

REFERENCES

[1]http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/

[2] Lawal Muhammad Aminu “Implementing Big Data Management on Grid Computing Environment” International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 3 Issue 9, September 2014 Page No. 8455-8459

[3] Agrawal et al., 2011; Baer et al., 2011 Agrawal, D., Das, S., &Abbadi, A. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM EDBT Conference, March 22–24, 2011, Uppsala, Sweden. http://dx.doi.org/10.1145/1951365.1951432

[4]Baer, T. (2011). 2012 Trends to Watch: Big Data. Ovum Report, OI00140-041. Baer, T., Sheina, M., and Mukherjee, S. (2011). What is big data? The big architecture. Ovum Report, OI00140-033.

[5] S.Vikram Phaneendra & E.Madhusudhan Reddy “Big Data- solutions for RDBMS problems- A survey” In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[6]Kiran kumara Reddi & Dnvsl Indira “Different Technique to Transfer Big Data: survey” IEEE Transactions on 52(8) (Aug.2013) 2348 {2355}

[7]Jimmy Lin “MapReduce Is Good Enough?” The control project. IEEE Computer 32 (2013).

[8]Umasri.M.L, Shyamalagowri.D ,Suresh Kumar.S “Mining Big Data:- Current status and forecast to the future” Volume 4, Issue 1, January 2014 ISSN: 2277 128X

[9]Albert Bifet “Mining Big Data in Real Time” Informatica 37 (2013) 15–20 DEC 2012

[10] Zan Mo, Yanfei Li Research of Big Data Based on the Views of Technology and Application American Journal of Industrial and Business Management, 2015, 5, 192-197 Published Online April 2015 in SciRes. http://www.scirp.org/journal/ajibmhttp://dx.doi.org/10.4236/ajibm.2015.54021

[11] Harshawardhan S. Bhosale1, Prof. Devendra P. Gadekar2 “A Review Paper on Big Data and Hadoop” International Journal ofScientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153

[12]https://www.quora.com/What-are-the-main-features-of-Hadoop

[13]https://www.slideshare.net/sandpoonia/1-grid-computing

[14]http://stackoverflow.com/questions/782913/googles-bigtable-vs-a-relational-database