Study of Big Data and Hadoop the Future of Information Technology System

[1]Mr. Akash Nigam, [2]Mr. Sonu Kumar, [3]Prof. Amrata Soni
[1]B.E. Student, [2]B.E. Student, [3]Assistant Professor
[1]Department of Computer Science,
[1]Mittal Institute of Technology, Bhopal, India

***Abstract* - Today Technology is rapidly changes and it is common that every one depends upon the new technologies and uses the server for storing and managing the database. Over the network there are too large set of database and now a day it is challengeable for manage those data base. In the age of Big-Data, Hadoop has evolved for handling those dataset. Hadoop is an open source project based on distributed computing having HDFS file system (Hadoop Distributed File System). Hadoop have many advantages that make them highly useful it has fault-tolerance capability & it can be deployed on low cost machines. Hadoop is useful for high volume of data set and it also provides the high speed access to the data set. Hadoop architecture is based on clustering which have the collection of nodes. Map-Reduce program in Hadoop used to collect data according to the query and this program is written in java.**

***Index Terms* -Big Data, Hadoop, Map Reduce, HDFS, Hadoop , Clustering**

I. INTRODUCTION

Big data is a collection of large set of structured and unstructured data. It is difficult to process by traditional database and software techniques. Big Data (Data Set) has processed a data of high, volume, velocity, variety to gain proposed data value and get promising authenticity of the considered data and extract information that make influence on cost, decision making and process control. When we discuss about Big Data, than this discussion is incomplete without "HADOOP" a well know product in the market of big data. Hadoop is Linux based product used by Google, Yahoo, and Twitter etc.

II. CHARACTERISTICS OF BIG DATA

There are three characteristics of Big Data. It is also known as 3Vs which are Volume, Velocity and Variety (As shown in the diagram 1.1).

(a) Velocity (Data in motion) - It shows the speed of data processing it takes milliseconds to minutes for the processing of data. Ex. a) 400 million tweets are done on twitter daily. b) 1 million transaction handles by Walmart in an hour. Velocity major by the how much time taken to capture, Storing and analyzing any data.

(b) Volume (Size of Data):- When data is in the gigabytes it is probably not Big Data, but at the terabyte and petabyte level and beyond it may very well be. Volume is a main issue to the problem of why traditional relational database management systems (RDBMS, data warehouses as we know them today) fail to handle Big Data. Underlying that failure are more complex issues of cost, reliability, long query times, and their inability to handle new sources of unstructured or semi-structured data like text.[2]

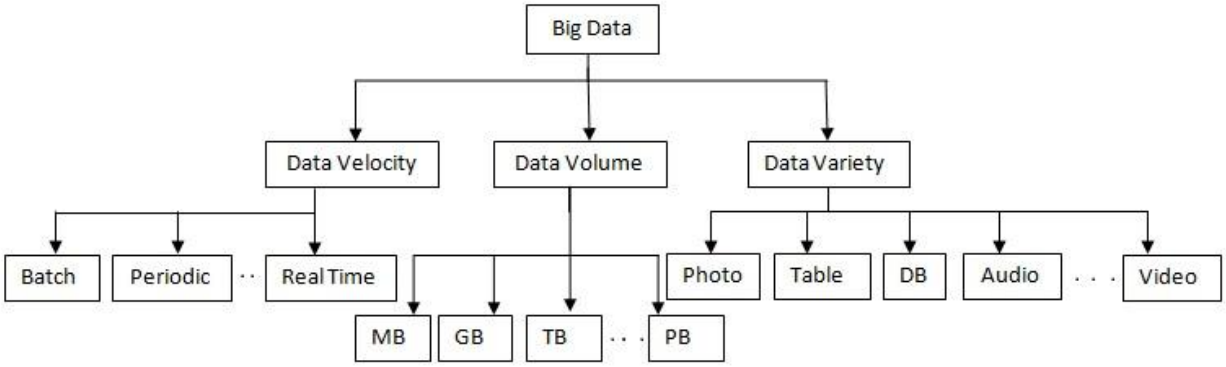

Figure 1.1: Characteristics of Big Data

(c) Variety of data: There are so many Different types of data and sources of data present over the network. Data variety exploded from structured data, unstructured, semi structured, images, audio, video, XML etc.

III. ARCHITECTURE DIAGRAM OF HADOOP

Hadoop architecture based on Master/Slave model for communication (As shown in fig. 1.2). Master is known as Name node and Slave are Data node. Name node control the access of data from the client and storage of data is managed by Data node. Hadoop divides the file into a number of blocks and these blocks are stored in Data node. Replicated blocks are also available to provide the high availability of the Hadoop system.

Components of Hadoop: The major components of Hadoop are:
Name Node: - Name Node is heart of Hadoop system. It manages the file system it contain the information of data block. This information is stored permanently on to local disk in the form of namespace image and edit log file. The Name node also knows the location of the data blocks.
Data Node: - It stores the blocks of data and retrieves them. The Data Nodes also reports the blocks information to the Name Node periodically.
APACHE HIVE: - It is data warehouse software which works on huge data which are present in distributed Storage. Hive use language which is similar to SQL known as HIVEQL. We can also use map/reduce program to plug-in when custom mappers and reducers is inconvenient or inefficient to express this logic in HiveQL.
HBASE: - It is Similar to Google big table designed to provide easy and fast access to large amount of structured data. It has one benefit of fault tolerance by a Hadoop File System. It is columns-oriented database which are filter by row.
HDFS: - Hadoop work on Distributed file system design. It run on cluster of computers. HDFS is fault tolerances which use low cost computers. HDFS holds tremendous amount of data and provide easy access. To store that amount of data, files are stored over multiple machine, so it seem like data is store in redundant fashion but in case of data loss ,we can recover it easily. HDFS also supports parallel processing.
Map Reduced: - It is a program model for distributed computing supports on java. it is a processing technique. MapReduced contain two tasks. First, is Map task which take data and convert it into individual elements broken into key/value pairs (Tuple). Secondly, reduced task take that data which is the outcome of Map task and break that data tuple into smaller set key/value pairs (Tuples).

IV.CASE STUDY OF ADVERTISEMENT TARGETING IN HADOOP

Hadoop is most preferable technology to select the best ads to show any given customer. Advertisement targeting is a unique kind of advisor engine, which select those ads where customer shows interest. This targeting system use customer data and make hypothesis on the basis of their preferences and behavior , then show most preferable that ads to customer.
Hadoop collect huge amount of data from the servers, and process that using over cluster of computer. Then Business analyst see reports and take important action to improve revenues, after observing customer behavior refined model is used to select best ads for customer in real time.

V.LITERATURE REVIEW

S. Vikram Phaneendra & E. Madhusudhan Reddy *et.al*. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Today, Data is generated from various different sources and can arrive in the system at various rates. To process these large amounts of data is a big issue today. In this paper we discussed Hadoop tool for Big data in detail. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. We also discussed some hadoop components which are used to support the processing of large data sets in distributed computing environments. In future we can use some clustering techniques and check the performance by implementing it in hadoop.[1]

Kiran kumara Reddi & Dnvsl Indira *et.al*. Is proposed an Online Aggregation approach for efficient big data processing. The aggregated results will be shown to the user and if user is agree with the obtained results then the further processing of MapReduce can be stopped. The early result will be shown to user by taking snapshots of the intermediate results obtained. As a local aggregate the combiners are implemented alongside with mapper and reducers for the aggregation. The combiners processes intermediate data produced by the mappers. And the result is then forwarded to reducers for further processing. The processing time of Hadoop is reduced at some instance. [3]

Albert Bifet *et.al*. Is discussed the challenges that in our opinion, mining evolving data streams will have to deal during the next years. We have outlined new areas for research. These include structured classification and associated application areas as social networks. Our ability to handle many exabytes of data across many application areas in the future will be crucially dependent on the existence of a rich variety of datasets, techniques and software frameworks. There is no doubt that data stream mining offers many challenges and equally many opportunites as the quantity of data generated in real time is going to continue growing.[5]

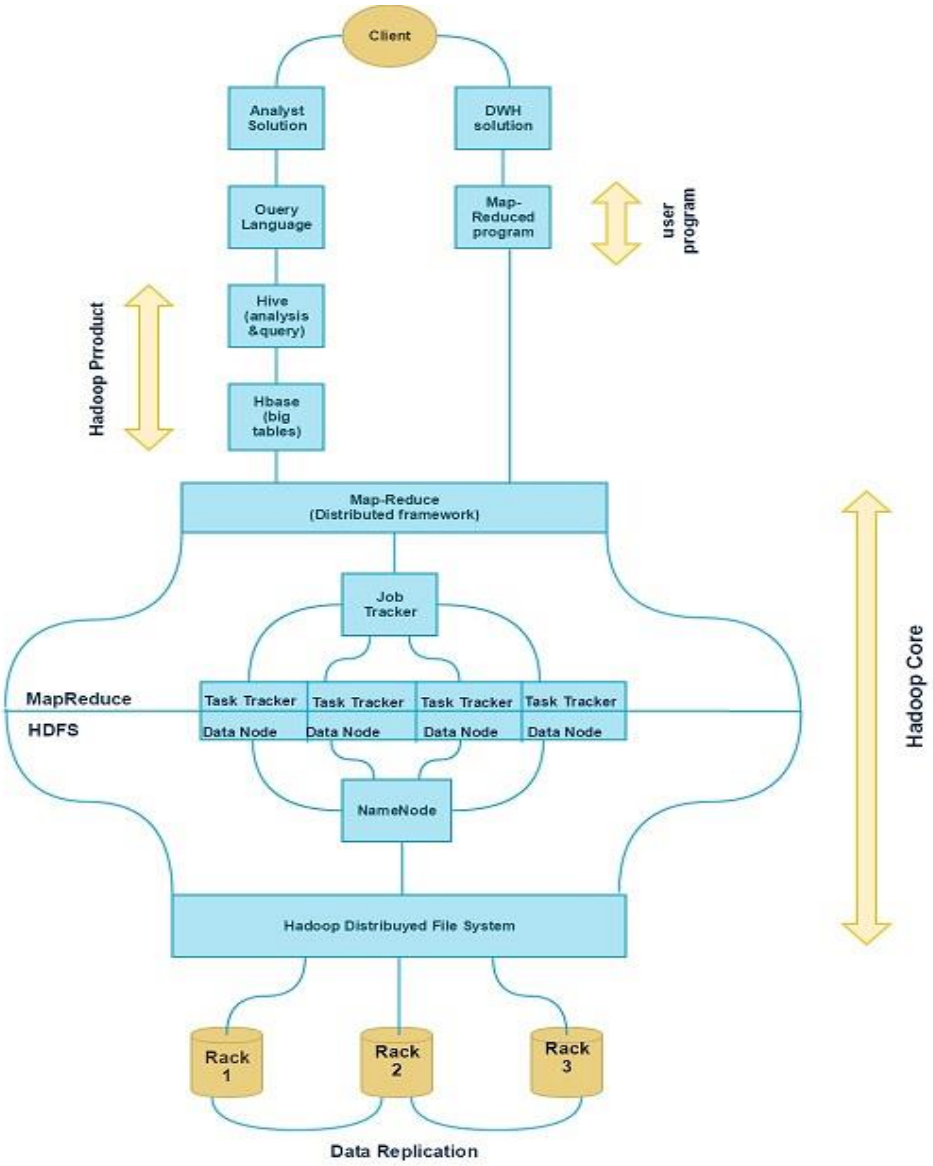

Figure 1.2: Architecture Diagram of Hadoop

VI. CONCLUSION

Now a day Big Data (Hadoop) is in huge demand in the market. As over the network there is huge amount of data but there is problem to How to manage? How to access the relevant dataset? Hadoop is the new era to handle those dataset. It have so many advantages that make it more useful in the market. Like it deploy on the low cast hardware machine and used by large set of audiences on huge amount of dataset. . In this paper we have studied Big Data, Characteristics of Big Data (3Vs Volume, Velocity, and Variety) and Architecture diagram of Big Data and study of previous research papers.

VII. ACKNOWLEDGMENT

We would like to express our gratitude to all those who gave us the possibility to complete this paper. We want to thank the Computer Science Department of the Mittal Institute of Technology, Bhopal for giving us permission to commence this paper in the first instance, to do the necessary research work and to use departmental data. We are deeply indebted to Prof. Amrata Soni from the CS Department Mital Institute of Technology, Bhopal for her guidance, stimulating suggestions and encouragement.

VIII. REFERENCES

1) 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).
2) http://data-magnum.com/how-many-vs-in-big-data-the-characteristics-that-define-big-data/
3) Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data: survey" IEEE Transactions on 52(8) (Aug.2013) 2348 {2355} Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013).
4) 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
5) Albert Bifet "Mining Big Data In Real Time" Informatica 37 (2013) 15–20 DEC 2012.
6) Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013.

7) Sameer Agarwal†, Barzan MozafariX, Aurojit Panda†, Henry Milner†, Samuel MaddenX, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data" Copyright © 2013ì ACM 978-1-4503-1994 2/13/04
8) Yingyi Bu _ Bill Howe _ Magdalena Balazinska _ Michael D. Ernst "The HaLoop Approach to Large-Scale Iterative Data Analysis" VLDB 2010 paper "HaLoop: Efficient Iterative Data Processing on Large Clusters.
9) Shadi Ibrahim⋆ _ Hai Jin _ Lu Lu "Handling Partitioning Skew in MapReduce using *LEEN*" ACM 51 (2008) 107–113
10) Kenn Slagter · Ching-Hsien Hsu "An improved partitioning mechanism for optimizing massive data analysis using MapReduce" Published online: 11 April 2013 © Springer Science+Business Media New York 2013
11) Ahmed Eldawy, Mohamed F. Mokbel "A Demonstration of SpatialHadoop:An Efficient MapReduce Framework for Spatial Data" *Proceedings of the VLDB Endowment, Vol. 6, No. 12 Copyright 2013 VLDB Endowment 21508097/13/10.*
12) Jeffrey Dean and Sanjay Ghemawat "MapReduce: Simplified Data Processing on Large Clusters" OSDI 2010.
13) Niketan Pansare1, Vinayak Borkar2, Chris Jermaine1, Tyson Condie "Online Aggregation for Large MapReduce Jobs" August 29September 3, 2011, Seattle, WA Copyright 2011 VLDB Endowment, ACM.
14) Tyson Condie, Neil Conway, Peter Alvaro, Joseph M. Hellerstein "Online Aggregation and Continuous Query support in MapReduce" *SIGMOD'10,* June 6–11, 2010, Indianapolis, Indiana, USA. Copyright 2010 ACM 978-1-4503-0032-2/10/06.
15) Jonathan Paul Olmsted "Scaling at Scale: Ideal Point Estimation with 'Big-Data" Princeton Institute for Computational Science and Engineering 2014.
16) Jonathan Stuart Ward and Adam Barker "Undefined By Data: A Survey of Big Data Definitions" Stamford, CT: Gartner, 2012.
17) Balaji Palanisamy, Member, IEEE, Aameek Singh, Member, IEEE Ling Liu, Senior Member, IEEE" Cost-effective Resource Provisioning for MapReduce in a Cloud"gartner report 2010, 25.
18) Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" Volume 4, Issue 5, May 2014" 27.
19) Kyong-Ha Lee Hyunsik Choi "Parallel Data Processing with MapReduce: A Survey" SIGMOD Record, December 2011 (Vol. 40, No. 4).
20) Chen He Ying Lu David Swanson "Matchmaking: A New MapReduce Scheduling" in 10th IEEE International Conference on Computer and Information Technology (CIT'10), pp. 2736–2743, 2010.