Emergence of Big Data with Hadoop : A Review

Prity Vijay[1], Bright Keshwani[2]

[1](Research Scholar, Computer Science,
Suresh Gyan Vihar University, Jaipur,
Email id:prityvijay1@gmail.com)

[2](Prof. & HOD of Computer Dept.
Suresh Gyan Vihar University, Jaipur,
Email id:kbright@rediffmail.com)

***Abstract*: -** We are living in an enormously developed, technical era, where internet is becoming fundamental need of all individual. Today, our social, personal as well as professional life are revolving around world wide web. Thus, giving birth to Big Data at an incredible momentum. Traditional management tools and frameworks are proved, un-fair while dealing with Big Data. This paper emphasize on number of review papers which acquainted us with the milieu of big data as well as emerging technologies helping Big Data. We also put light on the challenges increasing from the use of big data. We try to uncover correct approach to retrieve valuable information from the pile of Big Data.

Keywords: - *Hadoop, Big Data, Hbase, Hive, Pig etc*

I. INTRODUCTION

Computing has became global, number of devices like cell phones, smart phones, laptops, personal sensors are creating infinite stack of information. Before few years back data was recognized, as megabytes and gigabytes. But, today data is counted in terms of terabytes and petabytes. Zettabytes and Yottabytes will arrive soon. Nearly, 500 billion gigabytes of data, originate daily through internet[1].

Gone are the days, when the data was human generated and typically captured in tabular format. Machinery data (call detail logs, web logs, sensor data) arriving from the different sources, in assorted forms, building very complex Big Data. At present , almost each sectors are having minimum 100 terabytes of data with them and this number will be double in every six month[2]. This data blast is altering our world. These data's are dollars only, if handled properly. But, the open question is how to convert these un-structured messy data into precious information ? Many challenges occurs while dealing with Big Data. Big Data needs an application, which suits to unstructured data, provide near to real time analysis and having fault tolerance capacity. Apart from above, it should poses high storage and processing capacity. Diversify and outsized dataset are becoming impractical for traditional data management tools and applications. Therefore Big Data requires new set of tools, applications and frameworks.

II. BIG DATA:DEFINITION AND CHARACTERISTICS

Lot of confusion starts with the word Big data Itself. Big Data is not restricted to a fixed measurement? For some group 1TB can be big, for others 10TB may be big. In simple words, we can say, "Big Data is a circumstances where the volume, velocity and variety of data go beyond an organization's storage or computation capacity for precise and well-timed decision making". Doug Laney[3], characterize Big Data in terms of three V 's.

Volume: Volume refers to the size of data. With the growth of social media, the amount of data is growing very rapidly. Huge amount of data are generated through machines and it surpasses human generated data. This size aspect of data is referred to as *Volume* in the Big Data world.

Velocity: Velocity refers to the speed at which the data is being generated. Different applications have different latency requirements and in today's competitive world, decision makers want the important data information within fraction of second if possible. Generally, in real time or near to real time. A few examples include stock exchange data, tweets on Twitter, status updates/ likes/ shares on Facebook etc. This speed aspect of data generation is referred to as *Velocity* in the Big Data world.

Variety: Variety refers to the dissimilar formats in which the data is being generated. Today,70% of generated data are in unstructured format. Before the evolution of Big Data, the industry was not having any powerful management tools to manage large volume of unstructured data. In this competitive world industry can't depend only on the structured data, they have to guzzle lots of dissimilar data being generated from inside and outside of the enterprise (click stream data, social media, etc.).Apart from the traditional flat files, spreadsheets, relational databases etc., we have a lot of unstructured data stored in the form of images, audio files, video files, web logs, sensor data, and many others. This aspect of varied data formats is referred to as Variety in the Big Data world.

III. RECENT STATISTICS RELATED TO BIG DATA

Google is having 1 billion accounts. Daily, around 100 Terabytes of data is uploaded to Facebook. Facebook manage 30 Petabytes of user generated data, Twitter generates 12 Terabytes of data every day, LinkedIn having 313,000,000 number of total users, mining petabytes of user data for the feature "People You May Know", YouTube users upload 48 hours of new video content every minute of the day. Previously, Decoding of the human genome was taking around 10 years. Now it can be done in 7 days.500+ new websites are created every minute of the day. Instagram user share 40 million photos per day. Hence, Big data is mostly coming from Big Companies like yahoo, Google, Facebook, twitter, LinkedIn, Instagram,IBM and many more. In order to get hidden pattern and many other useful information from the ocean of Big Data there is a need of Big data analytics[1][14].

IV. PROBLEM MANAGING BIG DATA WITH TRADITIONAL APPROACH

Big data is all about- mammoth volume, high velocity and assorted dataset.RDBMS leader of IT world since last 30 years doesn't fit for Big Data. Because of the fact that it cannot handle outsized data of different variety. Moreover, RDBMS is very strict towards schema, having lots of constraints for the data passing to it, which suits to homogenous dataset. But, Big Data cannot be deal by such a rigid management system having lots of if's and buts'. As we are aware of the fact that in Big Data nature of data is not always known. Some time it may possess schema other time not. Another drawback of RDMBS is that data is analyzed based on relationships. In Big Data maintaining relationship between unstructured data (images, videos, Mobile generated information, RFID etc) is next to impossible. Apart from above , Big Data Analytics should posses very fast processing speed like real time or near to real time, which RDBMS doesn't guarantee. Therefore, NoSql with distributed file system can be a better approach for analyzing Big Data[4]. Last but not the least altering, maintaining and integrating Big Data is very costly solution when we deal Big Data with traditional approaches. All the above mentioned reason collectively created, a very severe need of new approaches for Big Data analytics[5].

V. HADOOP : BIG DATA NEED

Storage problem of Big Data is only part of the game[6].To cope up with, it incredible techniques are required. Although, for the management of Big Data many approaches are available . In spite , Hadoop along with its sub projects(Hive,Pig,Hbase,Zookepear,Flume,Oozie,Sqoop,etc)could be great option for the Big Data analytics[7]. In the last six years, Hadoop for dealing Big Data, become standard of many organization[8]. As several technologies associated to the Big Data are open source, vendors are showing their interest in integrating these tools with their own products[18].

The Apache Hadoop, an open source framework designed to support distributed parallel processing of large amount of data sets across clusters of computers using simple programming models. it is written in java. it can run on commodity hardware, scaling up from single nodes to thousands of computer, thus forming cluster. Each node in the cluster offers local computation and storage. Hadoop promise high availability to end user. Instead of depending on hardware to deliver high-availability, its framework itself can identify and figure out single point of failures at the application layer, thus providing a high availability service to its user[8].It has a number of commercially supported distributions from companies such as Map Technologies, Hortonworks, Cloudera, etc[9].

Doug Cutting, master mind of Hadoop[10],share journey of Hadoop in an interview. Hadoop get its name from the toy elephant of creator son[11].Hadoop was started by two Yahoo employee, Doug Cutting and Mike Cafarella, in 2006.it was initially created to support Nutch[12],an open source web crawler. Hadoop was motivated by Google MapReduce and Google File System,launched by Google in 2003 to handle billions of data[13]. In order to share it Google released white papers explaining GFS and MapReduce in 2004.After a year, Yahoo started to use Hadoop and then in 2008, it was taken over by Apache, thus, presently known as Apache Hadoop. At present Hadoop is ten years old and has publicized a incredible escalation in these years.

As predicted by IDC, Hadoop alone will be worth $813 million in 2016. Similarly, Forrester Predictions, 2015 says, “Hadoop is a essential for large enterprises, making the keystone for flexible future data platform needed in the age of the customer."In the last six years, Hadoop has become one of the most powerful data handling and management frameworks for distributed applications[13].

Hadoop is an open source Apache framework, is mainly divided into two parts[15] – HDFS : Hadoop Distributed File System for storage and processing and MapReduce, programming language, in JAVA to look after all the programming task. Hadoop works on master and slave architecture consisting one Name node (Master) and various Data node (Slave).HDFS is cost effective, highly fault tolerance, reliable, data processing system which is designed to run on cheap commodity hardware and to store terabytes (TB) and petabytes (PB) of distributed unstructured data very easily. Map Reduce, act as software for processing large datasets. It has basically two main functions Map and Reduce[16]. Map split the data into <key, value> pair and generate intermediate value, reduce conclude final output, out of intermediate value produced by map function. Work flow of Map, Reduce consists of mapping, sharing, shuffling and reducing.Jene and Dittrich[17] ,concluded that Hadoop, become standard for many organization various techniques come to boost up the performance of Hadoop. Main problem of Hadoop is its physical data organization including data layout and indexes. Therefore many effective data layout and number of indexing techniques have been proposed recently. HAIL: indexing technique, improve 70% of Hadoop performance

In Map Reduce replicas[19] (duplicate copy) are managed in HDFS for better reliability and availability. Over all, replica management system is taken care by Name node. So performance of HDFS depends mainly on name node, if name node fails HDFS suffer from architectural drawback. Metadata management[20] is critical to distributed file system. In HDFS architecture, a single master server manages all metadata, while a number of data servers store file data. This architecture is not appropriate for storage demand in cloud computing, as the single master server may become a performance bottleneck. Hadoop and its subsets are unsteady and immature, which leads to permanent modification of this framework that imposes costs of continuous training in organization, lack of expert man power, lack of ability of real time data processing. This can be solving by combination of Hadoop with Nosql Database. Moreover, storm and samza can be used for real time processing of high volume of data.

VI. SUB-PROJECT SUPPORTING HADOOP:A LITERATURE

Around 2009, Hadoop had been proved to be good approach for Big Data over traditional data ware house. However, writing Hadoop program was very much difficult. People started missing simplicity of query language. Therefore, Facebook[21] come up with the solution known as Hive. Nearly, in the same year Yahoo[22] developed Pig. Intention of both Hive and Pig was to bring simplicity to the complex code of MapReduce. Both Hive and Pig is an open-source solution built on top of Hadoop. Hive is a data ware house that supports queries like SQL known as HiveQL, which are compiled into map reduce jobs are executed using Hadoop. Pig is a data flow language, generates the code in Pig Latin. Unlike Pig, in Hive a schema is mandatory.

YSmart aims to provide a generic framework to translate SQL queries into optimized MapReduce jobs, and executing them efficiently on large scale distributed cluster systems. YSmart[23] can merged in Hive, to make Hive perform better, and can also be an independent SQL-to-MapReduce translator.

SCOPE[24] is heavily influenced by SQL.It is designed for easy and efficient processing of massive amounts of data stored in distributed, sequential files. It provides efficient query processing functionality.

Performance of Hive degraded in map reduce because of cost of saving intermediate results. Therefore network cost becomes higher and multiple replicas are usually kept. AQUA (Automatic Query Analyzer)[25], a query optimization method designed for MapReduce based data ware house systems. Given an SQL-like query, AQUA generates a sequence of MapReduce jobs, which minimizes the cost query processing

Objective was study and analyze various scheduling techniques, which are important to increase performance in Hadoop. Author introduced Quincy Scheduler[26], which is better in terms of distributing job into a distributed data node.

Lucene[27], suitable for application which requires full text indexing and searching capability. Distributed Lucene is based on two Apache open source projects, Lucene and Hadoop and may be said as the first of its kind which focuses on text based indexing. Lucene, a free open source information retrieval software library, originally created in Java

Hadoop image processing interface (HIPI) library[28], which hides the highly technical details of hadoop's. The goal of HIPI is to create a tool that will make development of large scale image processing and provide student and researchers to create large application easily. In HIPI culling is stage before mapping stage and finally provide image encoder and decoders that run behind.

In some cases similar multiple queries, common tables, and join tasks arrive simultaneously, arising many opportunities for computation sharing common jobs. Executing common tasks only once can remarkably reduce the total execution time of a batch of queries. Therefore, Shared Hive[29](Multiple Query Optimization framework), to improve the overall performance of Hadoop.SharedHive transforms a set of correlated HiveQL queries into a new set of *insert queries* that will produce outputs within a shorter execution time.

VII. CONCLUSION

Currently we are passing through Big Data phase. This paper focused on concept of Big Data along with 3 Vs.it also present problems and challenges while processing Big Data. In order to gain from Big Data these challenges must be addressed. The paper describes various pros and cons of Hadoop as a Big Data management tool. Although, Hadoop with its ecosystem is a powerful solution for handing Big Data. But still, Hadoop doesn't sounds good for frequently changing data. In other words we can say that at present there is no transaction support in Hadoop. Hadoop's only used for OLAP. Machine learning algorithm for Big Data need to be more robust and easier to use. Therefore, still betterment is needed in Big Data solution.

REFERENCES

[1]. Shiplap and M. Kaur,"*Big Data and Methodology-A review*", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.3(10), Oct 2013, pp. 991-995

[2]. SAS White Paper,(2012),"*Big Data Meets Big Data Analytics*".Accessed:http://docplayer.net/765971-Big-data-meets-big-data-analytics.html

[3]. D Laney," *3d data management: Controlling data volume, velocity and variety*", META Group Inc,6 Feb 2001

[4]. V. Shukla, P. K. Dubey," *Big Data: Moving Forward with Emerging Technology and Challenges* ", International Journal of Advance Research in Computer Science and Management Studies, Vol.2(9), Sept 2014, pp. 187-193

[5]. Lavastorm Analytics,(2014),*"Why Most Big Data ProjectsFail"*, Accessed: http://www.lavastorm.com/assets/Why-Most-Big-Data-Projects-Fail-White-Paper.pdf

[6]. B. Purcell," *The emergence of "big data" technology and analytics*", Journal of Technology Research,pp.1-6

[7]. V. S. Patil and P. D. Soni,*"HADOOP SKELETON & FAULT TOLERANCE IN HADOOP CLUSTERS "*,International Journal of Application or Innovation in Engineering & Management (IJAIEM) Vol . 2(2), Feb 2013,pp. 247-250

[8]. Hadoop Wiki,*"Apache Hadoop"*, Accessed : http://wiki.apache.org/hadoop

[9]. Juniper Networks,(2012),*"Introduction to Big Data:Infrastructure and Networking Considerations"*, JuniperNetworks.Accessed:http://www.juniper.net/us/en/local/pdf/whitepapers/2000488-en.pdf

[10]. D. Harris,(2013),*"The history of Hadoop:From 4 nodes to the future of data"*.Accessed: https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/

[11]. Vance, Ashlee (2009-03-17). "Hadoop, a Free Software Program, Finds Uses Beyond Search". The New York Times. Archived from the original on 11 February 2010. Retrieved 2010-01-20.

[12]. Michael J. Cafarella, Web.eecs.umich.edu. Retrieved 2013-04-05.

[13]. Intellipaat. "Hadoop Creator goes to Cloudera". *Intellipaat Blog*. Retrieved 2 February2016.

[14]. Wibibon blog,"A Comprehensive List of Big Data Statistics",Acessed : http://wikibon.org/blog/big-data-statistics/

[15]. Vidyasagar S. D," A study of Hadoop in information Technology Era", S.D,Global Research- Analysis, Vol 2(2),Feb 2013

[16]. Diana and Maclean,"A very Brief Introduction mapreduce", Acessed: http://hci.stanford.edu/courses/cs448g/a2/files/map_reduce_tutorial.pdf,2011.

[17]. Jene Dittrich and Jorge," Efficient big data processing in Hadoop Mapreduce",Acessed at: http://vldb.org/pvldb/vol5/p2014_jensdittrich_vldb2012.pdf ,2014.

[18]. Sheikh Ikhlaq and Dr. Bright Keswani," Computation of Big Data in Hadoop and Cloud Environment ",IOSR Journal of Engineering (IOSRJEN), Vol. 06, Issue 01 (January. 2016)

[19]. Chandra S and Sathyan S ,"Study On Replica Management And High Availability In Hadoop Distributed File system", Journal of Science , Vol 2(2),2012.

[20]. Varade M and Jethani V," Distributed Metadata Management Scheme in HDFS", Intern- ational Journal of Scientific and Research Publications, Vol 3(5),2013.

[21]. Facebook Data Infrastructure Team(2010),"Hive–A Petabyte Scale Data Warehouse Using Hadoop", Facebook Data Infrastructure Team, http://infolab.stanford.edu/ ~ragho /hive-icde.pdf,2010

[22]. Yahoo Team,"The Pig Experience", VLDB,2009

[23]. Facebook Data Infrastructure Team,"YSmart:Yet Another SQL-to-MapReduce Translat-or,"International Conference on Distributed Systems",2011

[24]. Chaiken R,Jenkins B,Larson P,Ramsey B,Shakib D, Weaver S, Zhou J,"SCOPE: Easy And Efficient Parallel Processing of Massive Data Sets", research.microsoft.com/en-us/um/people/jrzhou/ pub/Scope.pdf,2011.

[25]. Wu S, Li F ,Mehrotra S,Chin Ooi,"Optimization for Massively Parallel Data Processing " ,2011.

[26]. Arora S and Goel M," Survey Paper on Scheduling in Hadoop, International Journal of Advanced Research in Computer Science and Software Engineering", Vol 4(5),2014

[27]. Butler M and Rutherford J," Distributed Lucene: A distributed free text index for Hadoop ", HP Laboratories,2008

[28]. Sweeney C, Liu L, Arietta S, Lawrence J,"HIPI: A Hadoop Image Processing Interfa- ce for Image-based mapreduce tasks",cs.ucsb.edu/~cmsweeney/papers/undergrad_thesis .pdf,2011.

[29]. Tansel Dokeroglu, Serkan Ozal1, Murat Ali Bayir, Muhammet Serkan Cinar and Ahmet Cosar (2014), "Improving the performance of Hadoop Hive by sharing scan and computation tasks", Accessed at: http://link.springer.com/article/10.1186%2Fs13677-014-0012-6#page-1