Addressing Big data

Issues and Challenges

Ani Davis K[1]
Asst. Professor
Department of Computer Science,
Vimala College, Thrissur

Divya O M[2]
Asst. Professor
Department of Computer Science,
Vimala College, Thrissur

Thobias Vakayil[3]
Technical Leader, CISCO SYSTEMS,
Bangalore

Abstract -Big data is defined as large volume of data which requires new technologies and architecture to process the data and extract significant result from it. The creation and aggregation of data are increasing and will approach the zetta byte range within a few years. Because of this huge size of data it is very difficult to perform effective analysis using the existing traditional techniques. Volume is one of the main aspect of Big data, others being variety, velocity, value, and veracity. The challenge is, how to operate this volume of data in a proper way. Since Big data is a recent booming technology in the business and scientific environment, it is necessary that various challenges and issues associated with it should be brought into light. This paper gives a brief overview of Big data, its challenges, and also analyses in the perspective ofHadoop , MapReduce and MongoDB.

Keywords - Big data, Hadoop, MapReduce and MongoDB

I. INTRODUCTION

In order to offer efficient products and to keep the position in the market the business environment should mine and analyze their data to get helpful insights. Every year the data is growing exponentially. By the end of 2016,
Cisco estimates that the annual global data traffic will reach 6.6 zetta bytes [1].
The challenge will be how to handle large data requests in optimal time. Big data is a general term used to describe the voluminous amount of unstructured and semi-structured data that a company generates, the data that would take too much time and cost to load into a relational database for analysis. Data can be classified under several categories: Structured, Semi structured and Unstructured.

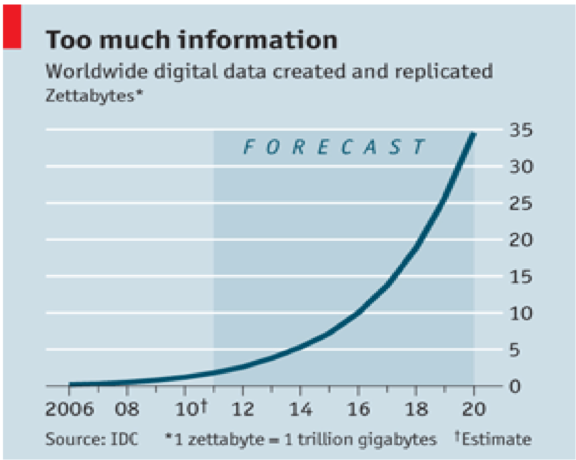

Figure 1: Worldwide Digital data Created and Replicated Zettabytes

Source: "The Leaky Corporation." The Economist. http://www.economist.com/ node/18226961

Structured data are organized into tables, based on defined rules. To work with structured data is easy because data are defined, indexed and easily filtered.
Unstructured data are not organized into tables and cannot be natively used by applications or interpreted by a database. Unstructured data files often include text and multimedia content.
Semi-structured data is structured data, but it is not organized using a conventional data model, like a table or an object-based graph. Many data found on the Web can be delineated as semi-structured. Data integration especially makes use of semi-structured data. In Semi-structured data the similar entities are grouped together but the entities in the same group may not have same attributes.

II. CHARACTERISTICS OF BIG DATA

Big data refers to huge collection of data that expands so quickly that it is difficult to handle with regular database or statistical tools. The request for more complex information is getting higher every year. A set of tools should be needed for processing large volume of data that is extremely complex. In order to extract information from these data, effective tools must be used to navigate and sort it.The methods of sorting data differ from one type of data to another. Bigdata can be characterized by different aspects. The commonly used aspects are

Volume: - "Volume", refers to the quantity of data that is being manipulated and analyzed in order to obtain the desired results. In order to manipulate, analyze and to achieve desired results the huge volume of data should be compressed, analyzed, decompressed. Various OLAP tools help the user to navigate and extract data.

Velocity: - The data movement is now almost real time and the update window has reduced to fractions of the seconds. It is really a challenge because usually data transferring is done at less than the capacity of the systems.

Variety: -Data can be stored in various formats. It may be Structured, Semi-structured, Unstructured. It is heterogeneous in nature.

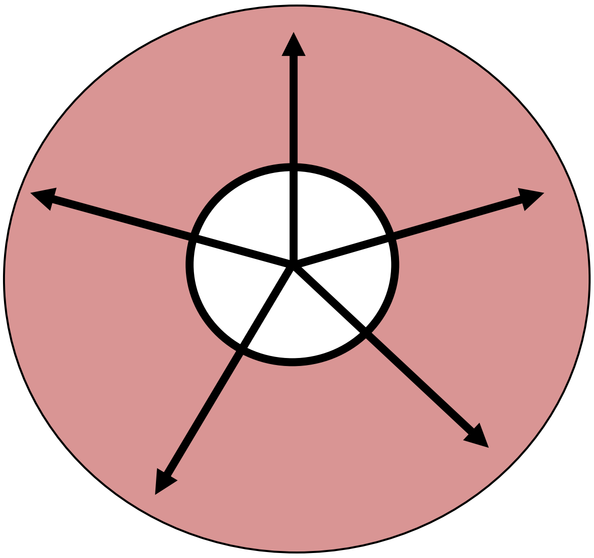

Figure 2: Characteristics of Big Data

Value: - The fourth "V" is "Value" and is all about the quality of data that is stored and the further use of it. Various statistical tools can be used to transform raw data into information that has value.

Veracity: - The possible consistency, credibility, validity of data is good enough for Big data. If a data which is send doesn't reach the destination properly, it results in its wear and tear.

III. CHALLENGES IN BIG DATA ANALYSIS

The organizations dealing with Big data are facing difficulties to create, manipulate, and manage data. The on-hand database management tools or traditional data processing applications are not appropriate to analyze the Big data. Some problems are:

1. Privacy: Trust on the agency or organization is very much related to the privacy of information. Handling of this voluminous data is difficult more over the constraints and accessing personal and confidential is more complicated. Most of the Big data analysis is happened by dividing the whole data into different sectors so the privacy of the data is questioned here.
2. Storage: The traditional methods and current tools are unable to store the massive data. The relational database such as SQL database is not suitable to keep the semi structured and unstructured data.
3. Retrieval of information: Manipulate the huge amount of data and retrieve the hidden pattern or other useful information from this diverse data is not easy.
4. Time/Speed: The time variable is also an important factor. In the fast moving world the organizations requires the relevant data to perform analysis and make decisions much more rapidly. Retrieve the reliable data within a time interval is crucial matter.
5. Heterogeneity: Transition between structured and unstructured data required for analysis will affect end to end processing of data. Invention of new non-relational technologies will provide some flexibility in data representation and processing [3].
6. Scale: In every moment the size of data is exponentially growing. Managing the large and rapidly increasing data requires a lot of hardware resources. The infrastructure must be developed to get high performance that needs high cost also.

IV. BIG DATA ANALYTICS

Big data Analytics is really about two things, Big data and Analytics and how the two have teamed up to create one of the profound trends in business intelligence[4]. Big data analytics enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of valuable business information and insights. It aid organizations for a better understanding of the information contained within the data and will also help identifying the data which is most important to the business and future business decisions. Big data analysts basically want the *knowledge* that comes from analyzing the data [5].

Big data analytics use to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. Big data analytics can be used for predictive analysis. Enterprises are increasingly looking to find actionable insights into their data.

V. APACHE HADOOP

The good news for the organization to handle the Big data is –Hadoop. Hadoop is open source software which supports the distributed analytic system. Hadoop's origin comes from Google File System GFS [6] and MapReduce [7] which become Apache HDFS and Apache MapReduce respectively. Nowadays Hadoop is widely used to analyze Big data. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model [8].

Hadoop has two main components MapReduce and file system, HDFS. MapReduce is the processing part of Hadoop and controls the jobs. HDFS refers to Hadoop Distributed File System stores all the data redundantly required for computations.

To achieve parallel execution, Hadoop implements a programming model named MapReduce. [9], [10]. MapReduce takes the data and split into different small chunks. Then process the small chunks and passes the partial output to master node. The master node collects the output from the other node and evaluates the original output [8]. The master node is known as Namenode and other nodes are Datanodes. The master Name Node as a coordinator of HDFS manages the file system namespace by keeping index of data location and regulates access to files by clients. Files and directories are represented on Name Node and it executes operations like opening, closing and renaming files and directories. Data Nodes are responsible for storing the blocks of file as determined by the Name Node. The Hadoop framework minimizes network congestion and increases bandwidth across the clusters as it schedules the computation closer to where data (or its replica) is present rather than migrating the large data sets to where application is running. This increases the overall throughput of the system[11]. Hadoop file system is reliable to process and store the large amount of data but not helpful to process the real-time data. So Hadoop is an OnLine Analytical Processing tool not an OnLine Transaction Processing.

VI. MONGODB

MongoDB is an open-source document database, and leading NoSQL database. MongoDB (from "humongous") is a cross-platform document-oriented database. MongoDB and Hadoop are fundamentally different systems. MongoDB is a database while Hadoop is a data processing and analysis framework. MongoDB focuses on storage and efficient retrieval of data while Hadoop focuses on data processing using MapReduce[12]. A document-oriented data base replaces the concept of a “row” with a more flexible model, the “document”. By allowing embedded documents and arrays, the document-oriented approach makes it possible to represent complex hierarchical relationships with a single record. There is no predefined schema: a document’s keys and values are not of fixed types or sizes [13]; so manipulating the data is an easy task.

The key concept behind the design of MongoDB is that there should always be more than one copy of the database. If a single database should fail, then it can simply be restored from the other servers. Because MonogoDB aims to be as fast as possible, it takes some shortcuts that make it more difficult to recover from a crash [14].

VII. EXPERIMENTAL FINDINGS

The transactional databases use structured data and its relationships whereas, analytical databases use both structured and unstructured data. Big data is analytical database.

Using the traditional database Oracle, to explore a pattern from one billion records, it took 20 seconds and when applied it in MongoDB, we get the result within 2 seconds.

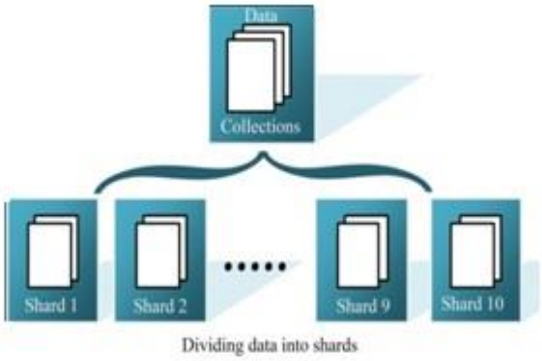

Figure 3: Dividing Data into Shards

In MongoDB, when it uses its own MapReduce framework, the master node divides the data into different shards. Each shard is considered as separate database because sharding removes the relationship between the tables.

One billion records were divided into 10 shards i.e. each shard had one million records. After processing, Master Node collected the results from each shard and combined the results according to the key.

The speed of processing depends on how the Database in MongoDB is sharded. The time and the no of shards are inversely proportional. The following graph represents the same.

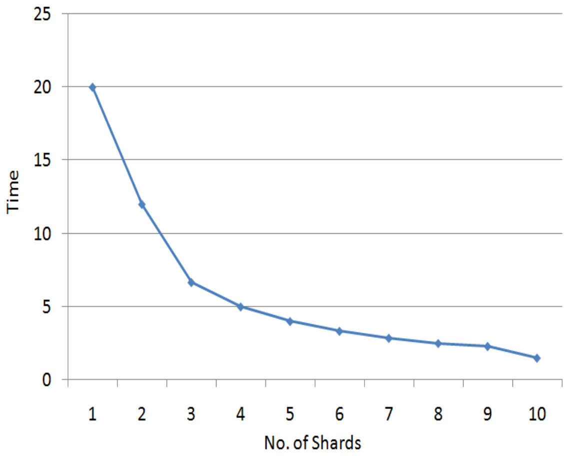

Figure 4: Graph shoeing Time and the no of shards are Inversely Proportional.

In order to perform the Big data experiments, the setup of Hadoop data cluster and Hadoop Distributed File System (HDFS) for storage, was used. Here we explain the number of concurrent Map and Reduce tasks that are allowed to run on each node. We configured our cluster to run ten concurrent tasks per server. Each Map/Reduce program that is run is partitioned into M map tasks and R reduce tasks. One node was organized as Master node and other nodes were chosen as slave nodes. The master node runs the “master” daemons means the NameNode is used for the HDFS storage layer and the JobTracker for the MapReduce processing layer. The slave nodes run the “slave” daemons denote that the DataNode is used for the HDFS layer and the TaskTracker for MapReduce processing layer. The master node was also used as slave node to increase the processing nodes. The software used for master and slave nodes was Sun Java 1.6, Ubuntu Linux 10.04 LTS and hadoop 1.0.3.

In our experiment we took 10 lakhsrecords which contain a total volume of 200GB of data. We created 10 chunks of the data and each chunk was having ~20GB data. By using the Hadoop file system, searching happens simultaneously in each chunk of 20GB data. Here we are able to search the data in a very quick manner.

For example: In order to search a pattern, like the reason for the dropout of students from the University, was taking 1 minute from the Oracle database. But by using the HDFS and Elastic search, we were able to achieve this by 5 seconds.

In short, the study supports the finding that the use of Hadoop Distributed File System (HDFS) and MongoDB saves much time compared to traditional databases. Also the Hadoop file system goes through only the previous data, not with the real time transaction.

VIII. CONCLUSION

In this paper we have conveyed about the characteristic of Big data and the challenges faced. Analysis of the Big data is time consuming task and required a lot of human resources and computer resources. Here, constructing a viable solution for large and complex data is a challenge. Hadoop file system, which uses Mapreduce helps the organizations to correlate the required patterns from the Big data. Hadoop as well as MongoDb works on the historical data. But it does not take into account the transactional data. It is a big challenge to analyse the Big data on real time. Analysis of Big data on real time is a new domain in the research.

REFERENCES

[1] Global data center traffic – Cisco Forecast Overview- http://www.cisco.com/en/US /solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index _White_Paper.html

[2] Stephen K, Frank A, J. Alberto E, William M, Big Data: Issues and Challenges Moving Forward”, IEEE, 46th Hawaii International Conference on System Sciences, 2013.

[3] DunrenChe, MejdlSafran, and ZhiyongPeng, From Bigdata to Bigdata Mining: Challenges, Issues, and Opportunities, DASFAA Workshops 2013, LNCS 7827, pp.1-15, 2013.

[4] http://www.webopedia.com/TERM/B/big_data_analytics.html

[5] S.Ghemawat, H. Gobioff, S. Lenug. “The Google file system”. In Proc. Of ACM Symposium on Operating System Principles, LakeGeorge NY, Oct 2003, pp29-43.

[6] J. Dean, S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” In Proc. of the 6th Symposium on Operating Systems Design and Implementation, San Francisco CA, Dec. 2004.

[7] Aditya B. Patel, Manashvi Birla, Ushma Nair, “Addressing Big Data Problem Using Hadoop and Map Reduce”, Nirma University International Conference On Engineering, 2012

[8] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, “Dryad: distributed data-parallel programs from sequential building blocks,”SIGOPS Oper. Syst. Rev., vol. 41, no. 3, pp. 59–72, Mar. 2007.

[9] P. Mundkur, V. Tuulos, and J. Flatow, “Disco: a computing platform for large-scale data analytics,” in Erlang ’11: Proc. of the 10th ACM SIGPLAN workshop on Erlang. New York, NY, USA: ACM, 2011, pp.84–89

[10] Kamalpreet Singh, Ravinder Kaur, “Hadoop: Addressing Challenges of Big Data”, IEEE International Advance Computing Conference (IACC), 2014

[11] Deep Mistry, “MongoDB vs Hadoop: Big Solutions for Big Problems”, White paper.

[12] Chodorow Kristina, “MongoDB: The Definitive Guide”, O’REILLY

[13] Tim Hawkins, Eelco Plugge, Peter Membrey, David Hows,The Definitive Guide to MongoDB: A complete guide to dealing with Big Data using Mongo DB” , Apress, 2013