BIG DATA AND HADOOP FRAMEWORK

Kartik Prakash
B.Tech, C.S.E
Amity University Haryana
Kartikmathur1993@gmail.com

Abhinay Thakur
B.Tech, C.S.E
Amity University Haryana
a8h1nay@gmail.com

Vikas Thada
Asst. Professor
Department of Computer Science & Engineering
Amity University Haryana
vthada@ggn.amity.edu

ABSTRACT

In this paper, we describe Big Data and open source framework Hadoop. The statistics are provided, explaining formation of Big Data and the importance of analyzing it. Big Data possess many problems in the real world scenario due to it's vast size, velocity and variety. Several techniques are developed to process Big Data. Hadoop framework is one such technique and is described along with the pseudo code for its mapper and reducer function. The file system of Hadoop - HDFS is explained as well with the help of appropriate diagram.

Keywords

Big Data, Hadoop, Mapper, Reducer, HDFS

1. INTRODUCTION

Big Data can be considered as a vast amount of Data. In technical terms it is such a big amount of data which escapes beyond the processing capacity of any database system. The amount of data is so huge it becomes a problem to store or process it in a traditional database system. Whole amount of data together forms a very complex structure, it becomes difficult to iterate and gain useful information through it.

Hadoop is an open source framework developed in Java by Doug Cutting and Mike Cafarella[1] in 2005 while working at yahoo. He named it after his son's toy elephant[2]. It is a widely used technology to process big data efficiently. It involves mapping technique, Reduce technique and Distributed File System called HDFS (Hadoop Distributed File System).

2. FACTORS DEFINING BIG DATA

Some factors are required to define Big Data appropriately and this section contains three major factors giving us an idea about how to define it[3].

2.1 Volume

Huge chunks or large amount of data in terms of sizing ranging from few terabytes to petabytes or even zeta bytes forces problem in the real life scenario of data processing.

2.2 Velocity

It refers to the rate in the sense of volume per second getting stored in the data warehouse of any enterprise or organization.

2.3 Variety

Data is available in different forms. It can be a structured, unstructured or semi-structured form of heterogeneous or homogeneous data.

Table 1. An example of Structured Data

Student	Course	Marks	Batch
Kartik	B.tech	98	2012
Abhinay	B.tech	97	2012
Rahul	BSC	67	2014
Pallavi	BJMC	90	2011
Oshi	MBA	80	2011

3. STATISTICS ABOUT BIG DATA FORMATION

According to [4], in 2008 the amount of data produced globally was 14.7 exabytes (nearly 1million terabytes) which is 3 times of data produced in 2003. The McKinsey Global Institute approximated that the data volume is increasing 40% per year and will grow 44 times in the duration of 2009-2020[5].

The major part of big data on World Wide Web is user generated which includes mainly images, text, videos and other media contents. The rate of increase in data is exponentially multiplying with each year. The nature of this data is mainly heterogeneous and due to its vast size and high velocity it leads to high complexity of processing of data.

In aviation industry, approximately 10 terabytes of data is generated from a thirty minute flight. There are almost 25,000 flights flying per day thus producing data in petabytes[6].

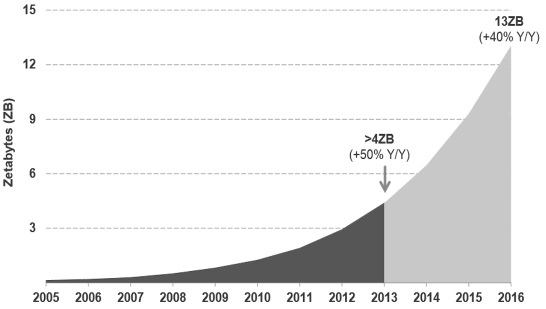

Fig.1. Growth Rate of Big Data[7]

4. IMPORTANCE OF BIG DATA ANALYTICS

4.1 Data Mining

An unstructured Big Data is of least value until a proper technique is applied to retrieve useful information and thus gaining knowledge from it. According to[8], large databases contains a hidden predictive information which is needed to be extracted for useful gains of an organization.

4.2 Business Importance

One of the major reasons of formation of big data in an enterprise is from the information gathered through enquiry of customers' behavior and habit patterns. Processing of this data is essential to establish better customer's relations and to create targeted ad campaigns. Extraction of this knowledge helps the organization to grow financially.

4.3 Genomic and Big Data

A single human Genome takes around 140 Gigabytes of data. The genetic sequencing data at European Bioinformatics Institute doubles in less than a year[9]. Proper processing of this Big Data is required to provide solution to analyze, store and generate useful information.

4.4 Cosmology

Cosmology is the study of universe and its properties as a whole by using scientific methodologies. These studies produce large amount of data sets generally in terms of petabytes. Proper analysis and accurate results are required to be retrieved from the big data to conduct further experiments and study.

5. TECHNOLOGY TO WORK WITH BIG DATA - HADOOP

Hadoop is a framework developed in Java which is used to extract useful data from Big Data efficiently. The framework uses Map Function, Reduce Function and Distributed file system. The concept of distributed file system is what makes it efficient as it distributes data over various nodes in a cluster in order to achieve parallel processing of data. In a multi-node cluster implementing Hadoop framework consist of two types of nodes – master and slave/worker.

5.1 MapReduce

MapReduce is a programming model which was initially developed by Google[10]. Hadoop is open source implementation of the same concept. The large input is divided into appropriate sizes and then provided to the map function in the first stage of processing. In a multi-node cluster the map function produces a <key, value> pair at master node and passes it to the slave or worker nodes for further processing. This step can be repeated by worker nodes thus forming a multi-level tree structure. The reduce function takes the output data from worker nodes and combines the data to form the required output at the master node.

5.1.1 Mapper Function

Following is the pseudo code for mapper function[11]:

```
void Map (key, value){
        for each word x in value:
                output.collect(x, 1); }
```

5.1.2 Reducer Function

Following is the pseudo code for reducer function[12]:

```
void Reduce (keyword, <list of value>){
        for each x in <list of value>:
                sum+=x;
        final_output.collect(keyword, sum);
}
```

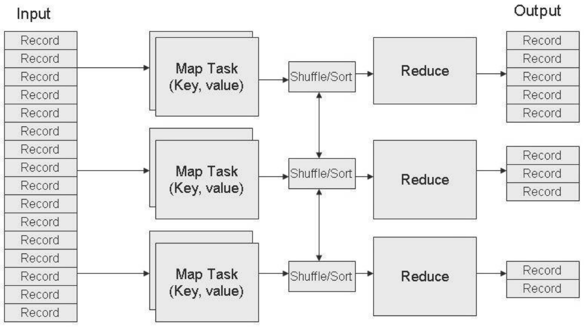

Fig.2. MapReduce Diagram.

5.2 HDFS

HDFS is the file system component of Hadoop based on distributed file system to store huge chunks of data sets. Distributed File System means storing data at different nodes in a cluster to increase efficiency of data processing and reduce effective cost. Generally, enterprises host applications and store data at same servers, using distributed file system brings reliability to this stored data and high bandwidth for the flow of data. HDFS consist of name node which manages the meta-data of file system and data node which stores the actual application data [13].

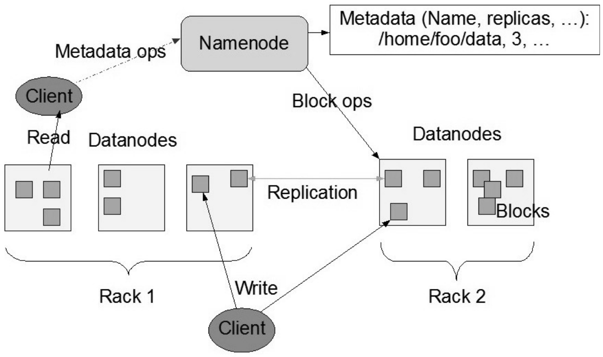

Fig. 3. HDFS Architecture

6. CONCLUSION

Processing of Big Data is an essential part of today's world. From this paper we have learned about what Big Data is. For example, consider the three V's used to define the Big Data. Formation of Big Data and the Statistics related to it are depicted in the form of a graph. From the graph it can be concluded that the data is growing exponentially both in volume and complexity.

Importance of processing and retrieving any useful information is crucial for various types of organizations as seen in the section 4 of this paper. Hadoop is an open source framework used to process Big Data. Hadoop consists of three major parts – Mapper, Reducer and HDFS (Hadoop Distributed File System).Mapper and Reducer together are called as MapReduce programming model.

HDFS is an important part of Hadoop as it enables the data to get distributed between different nodes of a cluster and hence resulting in greater efficiency both in terms of storage and processing.

7. ACKNOWLEDGEMENTS

We would like to acknowledge the research work done by others in the same field from which we were able to learn about Big Data and Hadoop framework. We would also like to thanks our faculty for providing us an endless support and guidance on this topic.

8. REFERENCES

[1] Cafarella, M. J. Web.eecs.umich.edu

[2] V, Ashlee. *"Hadoop, a Free Software Program, Finds Uses Beyond Search"*. The New York Times. 2009.

[3] Lancy, D. "*3D Management: Controlling Data Volume, Velocity and Variety"*. Gartner Inc. February, 2001.

[4] Bounie, D., and Gille, L. "*International Production and Dissemination of Information: Results, Methodological Issues, and Statistical Perspectives"*. International Journal of Communication. Vol. 6. 2012.

[5] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H. *"Big Data: The next frontier for innovation, competition, and productivity"*. McKinsey Global Institute. May 2011.

[6] *"Oracle: Big Data for the Enterprise"*. An oracle white paper. June 2013.

[7] KPCB, IDC Digital Universe

[8] Thearling, K. "*An introduction to Data Mining, Discovering hidden value in your data warehouse"*. www.thearling.com

[9] Marx, V. *"The Big Challenges of Big Data. Nature"*, Nature 498 (2013), 255-260.

[10] Dean, J. and Ghemawat, S. *"MapReduce: Simplified Data Processing on Large Clusters"*. OSDI 2004.

[11] salsahpc.indiana.edu, *"Pseudo code Mapper"*.

[12] salsahpc.indiana.edu, *"Pseudo code Reducer"*.

[13] Shafer, J., Rixner, S., and Cox, A. L. "*The Hadoop Distributed Filesystem: Balancing Portability and Performance"*. 2010.