Big Data Processing with Hadoop: A Review

Likith R[1], Sathisha G[2]
[1] Dept of Computer Science
[2]Asst. Professor, Dept of Computer Science
[1, 2] Atria Institute of Technology, Bangalore, India

Abstract- *We live in an era where data is being generated by everything around us. The rate of data generation is so alarming, that it has engendered a pressing need to implement easy and cost-effective data storage and retrieval mechanisms. Furthermore, big data needs to be analyzed for insights and attribute relationships, which can lead to better decision-making and efficient business strategies. In this paper, we will describe a formal definition of Big Data and look into its industrial applications. Further, we will understand how traditional mechanisms prove inadequate for data processing due to the sheer volume, velocity and variety of big data. We will then look into the Hadoop Architecture and its underlying functionalities. This will include delineations on the HDFS and MapReduce Framework. We will then review the Hadoop Ecosystem, and explain each component in detail.*

Keywords- Big Data, Hadoop, MapReduce, Hadoop Components, HDFS

I. INTRODUCTION

Big Data: Definition

Big data is a collection of large datasets- structured, unstructured or semi-structured that is being generated from multiple sources at an alarming rate. Key enablers for the growth of big data are – increasing storage capacities, increasing processing power and availability of data. It is thus important to develop mechanisms for easy storage and retrieval. Some of the fields that come under the umbrella of big data are - stock exchange data (includes buying and selling decisions), social media data (Facebook and Twitter), power grid data (contains information about the power consumed by each node in a power station) and search engine data (Google). Structured data may include relational databases like MySQL. Unstructured data may include textfiles in .doc, .pdf formats as well as media files.

Benefits of Big Data

Analysis of big data helps in improving business trends, finding innovative solutions, customer profiling and in sentimental analysis. It also helps in identifying the root causes for failures and re-evaluating risk portfolios. In addition, it also personalizes customer and interaction.

1. Valuable Insights

Valuable insights can be derived from big datasets by employing proper tools and methodologies. This dataincludes those stored in the company database, or those obtained from social media and other third party sources. When data is processed and analyzed, one can draw valuable relationships between various attributes that can improve the quality of decision making. Statistics and industrial knowledge can be combined to obtain useful insights

2. New Products and Services

Analyzing big data helps the organization to understand how customers perceive their products and services. This aids in developing new products that are concurrent with customer needs and demands. In addition, it also facilitates re- developing of currently existing products to suit customer requirements.

3. Smart cities

Population increase begets demand. To help cities deal with the consequences of rapid expansion, big data is being used for the benefit of the citizens and the environment. For example, the city of Portland, Oregon adopted a mechanism for optimizing traffic signals in response to high congestion. This not only reduced traffic jams in the city, but was also significant in eliminating 157,000 metric tons of carbon dioxide emissions.

4. Risk Analysis

Risk is defined as the probability of injury or loss. Risk management is a very crucial process which is often over- looked. Frequent analysis of the data will help mitigate potential risks. Predictive analysis aids the organization to keep up to date with recent technologies, services and products. It also identifies the risks involved, and how they can be mitigated.

5. Miscellaneous

Big data also aids Media, Government, Technology, Scientific Research and Healthcare in making crucial decisions and predictions. For example, Google Flu Trends (GFT) provided estimates of influenza activity for more than 25 countries. It made accurate predictions about flu activity.

Challenges of Big Data

1. Volume

Data is being generated at an alarming rate. The sheer volume of data being generated makes the issue of dataprocessing a complicated task. Organizations collect and generate data from a variety of sources and with the help of technologies such as Hadoop, storage and retrieval of data has become easier.

2. Velocity

Velocity refers to the rate at which data is being processed. Sometimes, data may arrive at unprecedented speeds, and thus it must be dealt with in a timely manner. It should be processed in such a speed that is compatible for real time applications.

3. Variety

Data is being generated from various sources , including social media data, stock exchange data and black box data. Furthermore, the data can assume various forms – numerals, text, media files, etc. Thus, big data processing mechanisms must know how to deal with eclectic data.

4. Variability

Data flow can be inconsistent which can be challenging to manage.

5. Complexity

The relationships between various attributes in a dataset, hierarchies and data linkages add to the complexity ofdata

II. LIMITATIONS OF TRADITIONALAPPROACH

The traditional approach consists of a computer to store and process big data. Data is stored in a Relational Database like MySQL and Oracle. This approach works well when the volume of data is less. However, when dealing with larger volumes of data, it becomes tedious to process it through a database server. Hence, this calls for a more sophisticated approach. We will now look into Hadoop – its modules, framework and ecosystem.

III. HADOOP

Apache Hadoop is an open source software framework for storing and processing large clusters of data. It has extensive processing power and it consists of large networks of computer clusters. Hadoop makes it possible to handle thousands of terabytes of data. Hardware failures are automatically handled by the framework.

Apache Hadoop consists of 4 modules:

a. Hadoop Distributed File System(HDFS)
b. Hadoop MapReduce
c. Hadoop YARN
d. Hadoop Common

This paper will primarily concentrate on the former two modules.

Hadoop Distributed File System (HDFS)

Apache Hadoop uses the Hadoop Distributed File System. It is highly fault tolerant and uses minimal cost hardware. It consists of a cluster of machines, and files are stored across them. It also provides file permissions and authentication, and streaming access to system data.

The following figure depicts the general architecture of HDFS

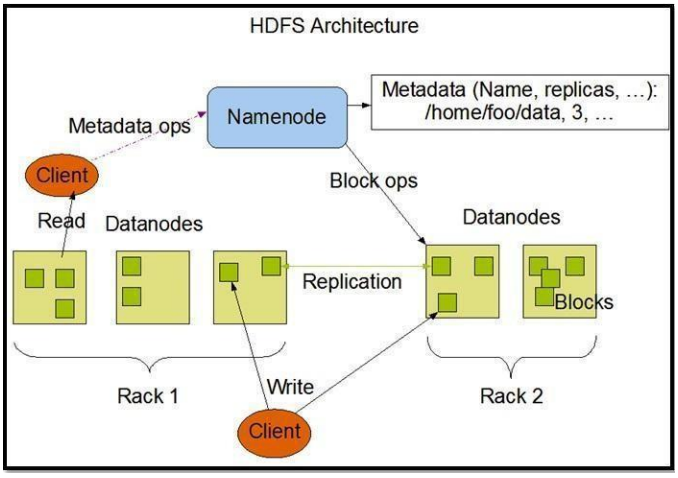

Figure 1: **HDFS Architecture**

HDFS follows the Master- Slave Architecture. It has the following components.

1. Name node

The HDFS consists of a single name node, which acts as the master node. It controls and manages the file system namespace. A file system namespace consists of a hierarchy of files and directories, where users can create, remove or move files based on their privilege. A file is split into one or more blocks and each block is stored in a Data node. HDFS consists of more than one Data Nodes.

The roles of the name node are as follows:

a. Mapping blocks to their data nodes.
b. Managing of file system namespace
c. Executing file system operations- opening, closing and renaming of files.

2. Data node

The HDFS consists of more than one data node. The data nodes store the file blocks that are mapped onto it by the Name node. The data nodes are responsible for performing read and write operations from file systems as per client

request. They also perform block creation and replication. The minimum amount of data that the system can write or read is called a block. This value however is not fixed, and it can be increased.

Hadoop MapReduce Framework

Hadoop uses the MapReduce framework for distributed computing applications to process large amounts of data. It is a distributed programming model based on the Java Programming language. The data processing frameworks are called mappers and reducers. The MapReduce framework is attractive due to its scalability.

It consists of two important tasks : Map and Reduce

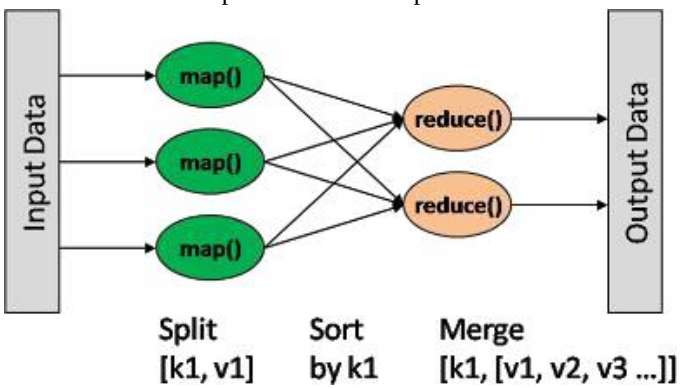

Figure 2: **MapReduce Framework**

1. Map stage

The map function takes in a set of data as the input, and returns a key-value pair as the output. The input may be in the form of a file or directory. The output of the map stage serves as input to the reduce stage.

2. Reduce stage

The reduce function will combine the data tuples into a smaller set. The map task always precedes the reduce task. The output of reduce stage is stored in the HDFS.

Hadoop Ecosystem

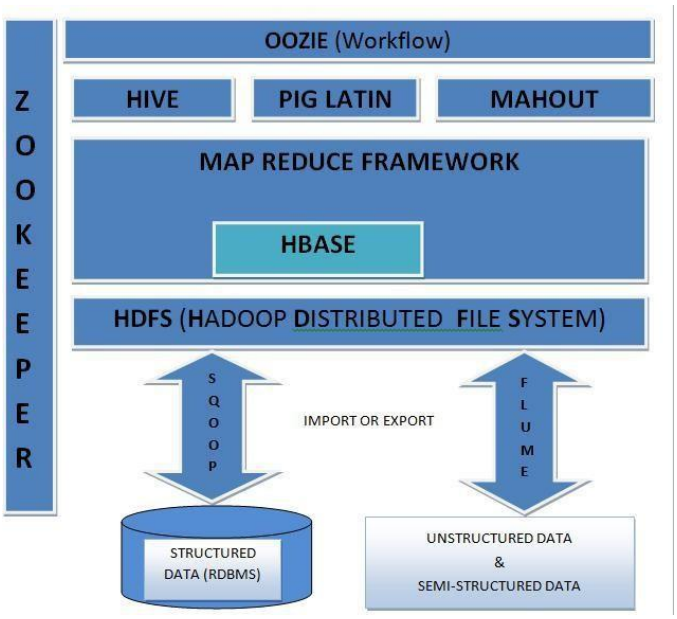

Figure 3: **Hadoop Ecosystem**

1. HBase

HBASE or Hadoop Database is a NoSQL Database, i.e., it is non-relational. It is built on top of the HDFS System written in Java. It is the underlying technology of social media websites like Facebook.

2. Hive

Hive is a structured Query Language. It uses the Hive Query Language (HQL), and it deals with structured data. It runs MapReduce Algorithm as its backend, and it is a data warehousing framework.

3. Pig

Pig also deals with structured data, and it uses the Pig Latin Language. It consists of a series of operations applied to input data, and it uses MapReduce in the back-end. It adds a level of abstraction to data processing.

4. Mahout

It is an open source Apache Machine Learning library in Java. It has modules for clustering, categorization, collective filtering and mining of frequent patterns.

IV. CONCLUSION

This paper starts off by giving a formal definition to Big Data. Then, the challenges of handling big data are examined, followed by the limitations of using the traditional big data processing approach. We then delve into the details of Hadoop and its components, and its MapReduce framework.

REFERENCES

[1] Dean, J. and Ghemawat, S., "MapReduce: a flexible data processing tool" Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
[2] Varsha B.Bobade, "Survey Paper on Big Data and Hadoop", IRJET, Volume 3, Issue 1, January 2016
[3] Bijesh Dhyani,Anurag Barthwal, "Big Data Analytics using Hadoop", International Journal of Computing Applications, Volume 108, No.12, December 2014
[4] Ms. Gurpreet Kaur,Ms. Manpreet Kaur, "Review Paper on Big Data using Hadoop", International Journal of Computing Engineering and Technology, Volume 6, Issue 12, Dec 2015, pp. 65-71
[5] Harshwardhan S. Bhosale et al, "Review paper on Big Data using Hadoop", International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014
[6] Poonam S. Patil et al. "Survey Paper on Big Data Processing and Hadoop Components", International Journal of Science and Research, Volume 3, Issue 10, October 2014.
[7] Apache HBase. Available at http://hbase.apache.org
[8] Apache Hive. Available athttp://hive.apache.org
[9] Abhishek S, "Big Data and Hadoop", White Paper
[10] Konstantin Shvachko et.al, "The HadoopDistributedFile System"