BIG DATA SYSTEM AND HADOOP

Priyanka L.A

Priyankaponnuswami97@gmail.com

Abstract:

We live in an epoch where data is being generated by everything around us. The rate of data generation is so scare, that it has engendered a pressing need to implement cost-effect and easy data storage and retrieval mechanisms. Also, 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 present the basic understanding of BIG DATA is and it is usefulness to an organization from the performance perspective. Big data can be structured, semi-structured or unstructured, resulting in in capability of conventional data management methods. This includes the three V's of big data which are velocity, variety and volume. We will then look into the Hadoop Architecture and its basic functionalities. This will include descriptions on the HDFS and Map Reduce Framework. Hadoop is the basic platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Hadoop is an open source software project that delegate the distributed processing of large volume of data with a very high degree of fault tolerance.

1. Introduction

1.1 Big Data: Definition

Big data is a large volume of datasets- structured, Semi-structured or unstructured that is being generated from multiple sources at an alarming rate. Key enablers for the growth of BID DATA are – increasing processing power, increasing storage capacities and availability of data. It is thus important to develop mechanisms for easy storage and retrieval. This is probably one of the important reasons why the concept of BIG DATA was first enclosed by online Firms like Google, eBay, Facebook, Lingmailkedin etc.

1.2 Benefits of Big Data

Analysis of big data helps in upgrade business trends, finding inventive 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 intuitions

Valuable intuition can be derived from big datasets by employing proper tools and methodologies. This data contain those stored in the company database, or those obtained from other third party sources and social media. When data is analyzed and processed, 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 intuitions.

2. New Products and Services

Examine BIG DATA helps the organization to understand how customers recognize their products and services. This aids in developing new products that are concurrent with customer demands and needs. Furthermore, 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 effect 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 process for optimizing traffic signals in response to high congestion.This not only minimize 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 loss or injury. Risk management is a very critical process which is often over-looked. Frequent analysis of the data will help reduced potential risks. Predictive analysis aids the organization to keep up to date with modern technologies, products and services. It also identifies the risks involved, and how they can be reduced.

5. Miscellaneous

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

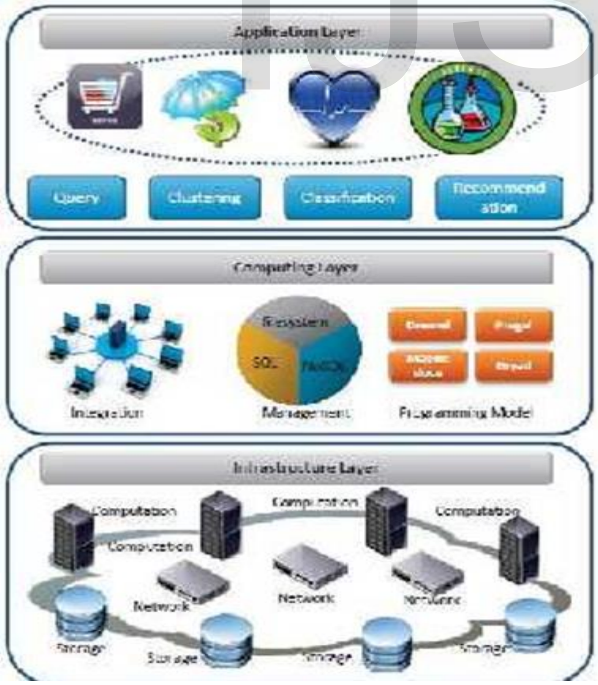

Figure 1: Layered Architecture of Big Data system

1.3 Challenges of Big Data

1. Volume

Data is being generated at an alarming rate. The total volume of data being generated makes the issue of data processing a complicated task. Generate data and Organizations collect 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 in Big Data is a concept which deals with the speed of the dada coming from various sources. This mannerism is not being limited to the speed of incoming data but also speed at which the data flows and aggregated.

3. Variety

Data is being generated from various sources, including stock exchange data, social media data and black box data. Moreover, the data can assume various forms – numerals, text, media files, etc. consequently, 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 data linkages, dataset and hierarchies add to the complexity of data

2. LIMITATIONS OF TRADITIONAL APPROACH

The traditional approach consists of a computer to process and store big data. Data is gather in a Relational Database like MySQL and Oracle. This approach works effectively 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.

3. Hadoop

Apache Hadoop is an open source software framework for processing and storing large clusters of data. It has extensive processing power and it consists of large networks of computer clusters. Hadoop create 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 focus on the former two modules.

3.1 Hadoop Distributed File System (HDFS)

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

The following figure illustrate the general architecture of HDFS

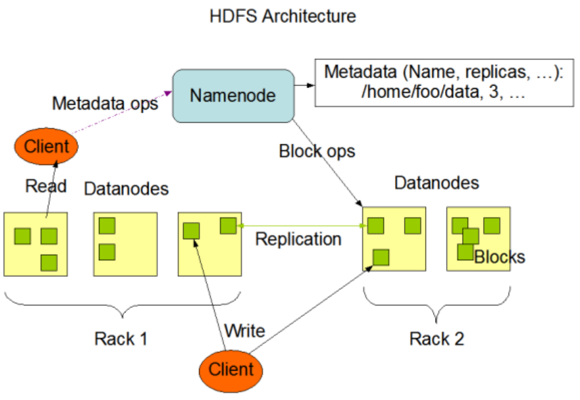

Figure 2: HDFS Architecture

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

1. Name node

The HDFS involve 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 directories and hierarchy of files, where users can create, move or remove files based on their privilege. A file is lop into one or more blocks and each block is stored in a Data node. HDFS consists of more than one Data Nodes.

The part 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 involve of more than one data node. The data nodes cache 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 read or write is called a block. it can be increased, and This value however is not fixed.

3.2 Hadoop MapReduce Framework

Hadoop uses the MapReduce framework for dispense 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 reducers and mappers. The MapReduce framework is delightful due to its scalability. It include of two important tasks : Map and Reduce

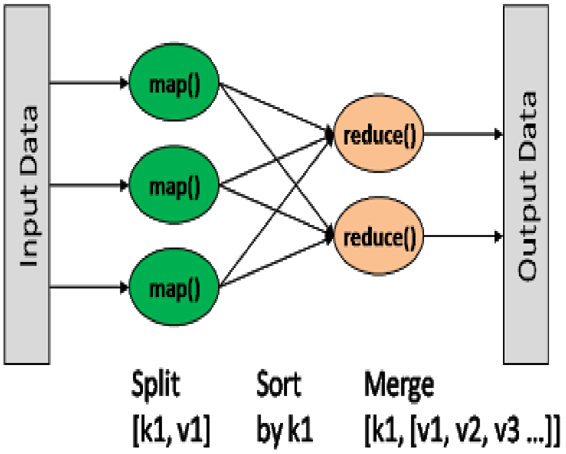

Figure 3: MapReduce Framework

1. Map stage

The map function takes in a volume 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 obey as input to the reduce stage.

2. Reduce stage

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

3.3 Hadoop Ecosystem

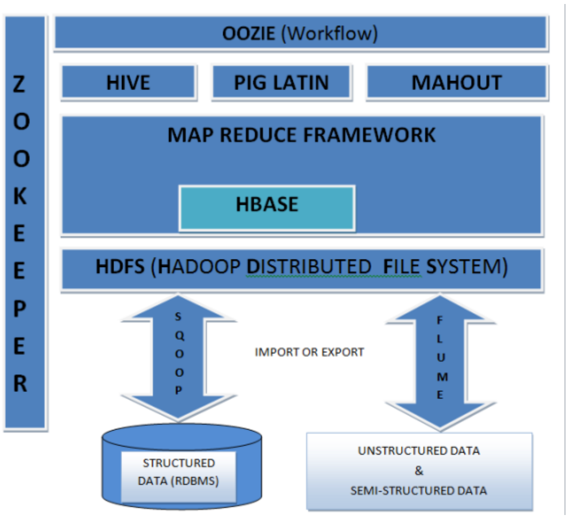

Figure 4: Hadoop Ecosystem

1. HBase

Hadoop Database or HBase is a NoSQL Database, i.e., it is non-relational. It is create 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 deals with structured data, and it uses the Hive Query Language (HQL). It dash MapReduce Algorithm as its backend, and it is a data warehousing framework.

3. Pig

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

4. Mahout

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

4. Conclusion

We have entered an era of Big Data. The paper represent the concept of Big Data along with 3 Vs, Volume, Velocity and variety of Big Data. The paper also focus on Big Data processing problems. These technical challenges must be addressed for fast and efficient processing of 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 scour into the details of Hadoop and its components, and its MapReduce framework. The paper represent Hadoop which is an open source software used for processing of Big Data.

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 at http://hive.apache.org

[9] Abhishek S, "Big Data and Hadoop", White Paper

[10] Konstantin Shvachko et.al, "The Hadoop Distributed File System"