International Journal on Recent and Innovation Trends in Computing and Communication
Volume: 5 Issue: 7

ISSN: 2321-8169
749 – 752

Elementary Concepts of Big Data and Hadoop

Hasmukh B. Domadiya
Assistant Professor,
National Computer College,
Jamnagar.

Dr. Girish C. Bhimani
Head of Department,
Department of Statistics,
Saurashtra University, Rajkot

Abstract: This paper is an effort to present the basic importance of Big Data and also its importance in an organization from its performance point of view. The term Big data, refers the data sets, whose volume, complexity and also rate of growth make them more difficult to capture, manage, process and also analyzed. For such type of data –intensive applications, the Apache Hadoop Framework has newly concerned a lot of attention. Hadoop is the core platform for structuring Big data, and solves the problem of making it helpful for analytics idea. Hadoop is an open source software project that enables the distributed processing of enormous data and framework for the analysis and transformation of very large data sets using the MapReduce paradigm. This paper deals with the architecture of Hadoop with its various components.

Keywords: *Big Data, Hadoop, MapReduce, HBase, Avro, Pig, Hive, YARN, Sqoop, ZooKeeper, Mahout*

*****

I. Introduction:

Big data is a collection of enormous dataset that not only processed by any traditional computer techniques. Big data is not only a term , but it has various tools, techniques and also framework. The term Big data, refers the data sets, whose volume, complexity and also rate of growth make them more difficult to capture, manage, process and also analyzed. Generally, in Big data the data in it will be of three types: Structured Data –Relational data, Semi-structured data-XML data and also Unstructured data-Word, PDF, Text, Media Logs (1).

II. Big Data:

Current age is in the data age. According to Tom, It's not easy to measure the total volume of data stored electronically, but an IDC estimate put the size of the digital universe at 4.4 zettabytes in 2013 and is forecasting a tenfold growth by 2020 to 44 zettabytes. A zettabyte is 10^21 bytes, or equivalent one thousand exabytes, one million petabytes, or one billion terabytes. That is more than one disk drive for every person in the world (2).

The data in big data is a big no matter how we serving it. Current era is totally a society of sharers, but in just sixty seconds of different sharing and searching activities on internet. In the below figure, we can see that millions of items are accessed, searched, transferred and also shared. It's a shocking amount of data to think about. (3)

Challenges of Big Data:

1. **Volume:**
   Volume refers to amount or sizes of data. The storage data size may be represented in Terabytes or zeta bytes.
2. **Variety:**
   Variety refers different types of data and also different sources of data. Data may be in structured, unstructured and also sometimes it is in semi-structured.
3. **Velocity:**
   Velocity refers the speed of data processing. The data comes at very high speed.
4. **Veracity:**
   Veracity refers to noise, biases and abnormality when we dealing with high volume, velocity and variety of data, the all of data are not going to 100% correct, there will be dirty data (4).

 ISSN: 2321-8169
749 – 752

Hadoop: The solution of Big data:

Hadoop is open-source software for reliable, scalable and also distributed computing. The Apache Hadoop library is a framework that allows distributed processing of enormous data sets across clusters of computers using a simple programming model.

Hadoop was derived from Google's Map Reduce and Google File System (GFS) (5). Currently, Hadoop consists of Hadoop kernel, MapReduce, HDFS and different various components like Apache Hive, Base etc.

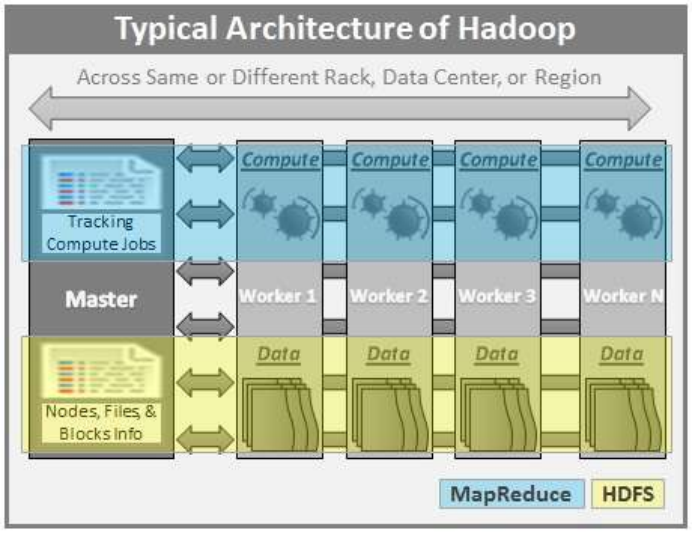

Fig: Architecture of Hadoop (6)

A. HDFS (Hadoop Distributed File System)

A HDFS is designed to hold a large amount of information like terabytes or petabytes and also provide access to this data to many clients distributed across a network. HDFS has master/slave architecture. Trough HDFS, data will be written once and then read several times.

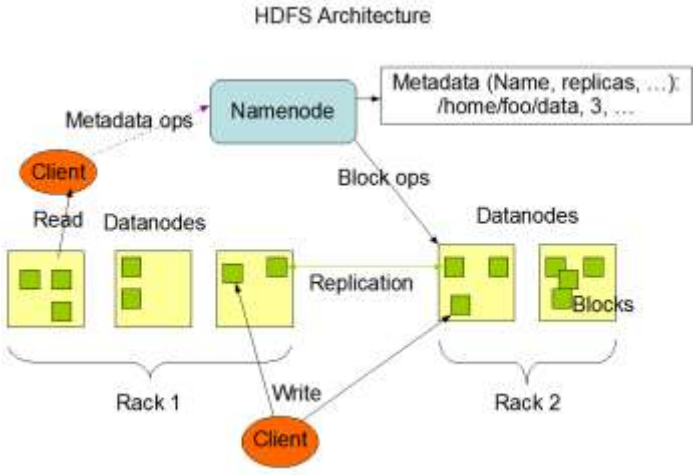

Fig: Architecture of HDFS (7)

HDFS is a block-structured file system. Individual files are divided into the chunks of a fixed size. These all small chunks are stored across a group of one or more machines with data storage capacity. HDFS stores three copies of each file by copying each piece to three different servers. Size of each block is 64 MB. HDFS architecture is divided into three different nodes, which are Name node, Data Node and also HDFS client node (8).

Name Node:

Name node is centrally placed node, which contains whole information about Hadoop file system. The main work of this node is it records all the metadata and attributes and also specific location of files and data blocks in the data nodes. This node is also called as a master node because in this node stores all the details about the system and also provides newly added, modified and also removed details from any data nodes.

Data Node:

Data node manages the data storage (9) of their system. Data nodes perform read and write operations on the file systems, as per client request. This node perform operations like chunk/block create, delete and also reproduction according to the instructions of the name node.

HDFS Clients:

An HDFS client is also known as Edge node. It is basically a link node between name node and data node.

B. MapReduce Architecture

Solanke poona et al. MapReduce is a programming model and related execution for processing and also generating enormous data with a parallel processing. This architecture poised of a two different functions. Map function used for filtering and sorting data and reduce function that performs a summary operation (10).

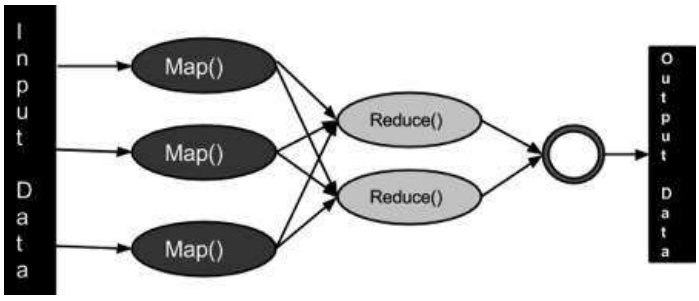

Fig: MapReduce Algorithm (11)

Map stage:

This stage is used to process the input data. Generally the inputted data is in the form of file or directory and is also stored in the Hadoop Distributed File System (HDFS). The inputted file is passing to the (11) map function through line by line. The map processes that data and also creates multiple small chunks of data.

Reduce stage:

This stage is combination of the shuffle stage and (11) also the Reduce stage. The main job of this stage is to process the data that comes from the map stage. After processing, it generates a new set of output, which will be stored in HDFS.

III. Literature Review:

S. Vikram Phaneendra et al., mentioned that in previous days the data was very few and anyone can handle it very easily through the RDBMS but currently it is too difficult to handle enormous data through that tools. This major problem to handle huge data is solved through the Big Data. In this paper, they explain the different dimensions of Big Data like Volume, Velocity, Variety, Value and also complexity (Veracity). In this paper, they also illustrated that big data can be found at anywhere like in Finance, Retail Industry, Healthcare, Insurance etc (12).

According to Albert Bifet et al., discussed that streaming examination in current era is now becoming the fastest and most proficient way to attain meaningful and useful knowledge, allowing any industry to respond quickly when any type of issues can appear or detect to improve the performance. In this paper, they also discussed different tools used for mining big data are Apache Hadoop, mahout etc (13).

Bernice Purcell, discusses that in Big Data includes structured data, semi-structured and also unstructured data. In this paper, they also discussed that Hadoop architecture is used to process different unstructured and semi-structured using MapReduce to locate all related data then select only the data directly answering the query (14).

Manyika et al. (2011), describes the major contributions of big data can make to businesses: Transparency creation, performance improvement, population segmentation, decision making support and also innovative business models, products and services (15).

Y. Lee et al., describes that Most of the MapReduce applications on Hadoop are developed to analyze large texts, web contents or log files. In this paper, the author specially highlighted on that first packet processing method of Hadoop that always analyzes any packet trace files in sequential manner through the reading all packets across from multiple HDFS chunks (16).

IV. Hadoop Components:

1). HBase:

HBase is a non-relational column oriented distributed database designed to run always on top of the Hadoop Distributed File System (HDFS). HBase is an open-source and distributed based on the Google's BigTable. This is an example of NoSQL data store and it's written in Java (17).

2). Avro:

Avro is data serialization format which brings data interoperability among multiple components of Apache Hadoop.

3). Pig:

Pig was initially developed at Yahoo Research around 2006 but it moved into the Apache software foundation list in 2007(18). It is a platform for analyzing enormous data sets which is consists of high level query language like SQL for expressing data analysis and also processing. Pig can easily process tera bytes of data with very little lines of code and through this platform writing and maintaining data processing jobs very easy (19).

4). Hive:

Hive is Data Warehousing application that provides the SQL interface and also relational model. It is also structures data into the well-known database concepts like tables, columns, rows and also partitions. It supports primitive as well as complex data types like maps, lists and structs (18).

5). YARN:

Yet Another Resource Negotiator (YARN), is included in the latest Hadoop release and its goal is to allow the system to serve as a general data processing framework. It supports programming models other than MapReduce, while also improving scalability and resource utilization (20).

6). Sqoop:

Sqoop is tool which can be used to transfer the data from relational database environments like Oracle, mysql into Hadoop environment. This is a command line interface platform that is used to transferring data between relational databases and Hadoop (4).

7). ZooKeeper:

Zookeeper is a distributed coordination and also governing service for Hadoop cluster. It is a centralized service that provides distributed synchronization and also maintaining the configuration information(4).

8). Mahout:

Mahout is a simple and extensible programming environment and framework for building scalable algorithms very quickly(21).

V. Conclusion:

In this paper, the author provided an overview of the Big Data and Hadoop. As an author, this paper considering the five main characteristics of Big Data and

discussing different components of Hadoop. I also believe that the information proposed here can help to discuss technologies already existing and also newly emerging in the opportunity, classify them and efficient mechanism and also guide them.

References

[1] http://www.tutotialspoint.com/hadoop/hadoop_big_data_overview.htm

[2] Tom White ,*"Hadoop: the Definitive Guide"* 4th edition, O'Reilly publication

[3] https://ricoheuropebusinessdriver.com/category/big-data

[4] Varsha B. Bobade (Jan-2016), *"Survey paper on Big Data and Hadoop"*, International Research Journal of Engineering and Technology (IRJET), Vol-3, Issue-1, pp. 861- 863, e-ISSN:- 2395-0056, p-ISSN:-2395-0072.

[5] Apache Hadoop, http://hadoop.apache.org

[6] https://www.mssqltips.com/sqlservertip/3140/big-data-basics--part-3--overview-of-hadoop

[7] Hadoop Architecture: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[8] Hadoop Tutorial: http://developer.yahoo.com/hadoop/tutotial/module1.html

[9] HDFSTutorial :https://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm

[10] Solanke Poonam G, B.M.Patil (2-2016), *"An Efficient Approach for Processing Big Data with Inceremental MapReduce"*, International Journal for Scientific Research & Development, Vol-4, Issue-2, pp. 53-57, ISSN (online) 2321-0613.

[11] MapReduce Tutorial: https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm

[12] S.Vikram Phaneendra & E.Madhusudhan Reddy (Apr 19-23, 2013) *"Big Data- solutions for RDBMS problems- A survey"* In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan).

[13] Albert Bifet, *"Mining Big Data in Real Time"*, Informatica 37 (2013) 15-20 Dec-2012.

[14] Bernice Purcell, *"The emergence of "big data" technology and analytics"*, Journal of Technology Research.

15). Manyika, J., Chui,. M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011, June). *"Big data: The next frontier for innovation, competition, and productivity"*. McKinsey Global Institute. Retrieved from http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

[16] Y. Lee, W.Kang, and Y.Lee (April-2011), *"A Hadoop based packet trace processing Tool"*, International Workshop on Traffic Monitoring and Analysis (TMA 2011).

[17] Apache HBase, https://hbase.apache.org/

[18] Ashish Thusoo, Joydeep Sen Sama, Namit Jain, Zhend Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy, *"Hive- A Petabyte Scale Data Warehouse Using Hadoop"*, By Facebook Data Infrastructure Team.

[19] Apache Pig, https://pig.apache.org

[20] Poonam S. Patil and Rajesh N. Phursule (Oct-2014), *"Survey paper on Big Data Processing and Hadoop Components"*, International Journal of Science and Research (IJSR), Vol-3, Issue-10, pp. 585-590, ISSN (Online): 2319-7064.

[21] Apache mahaout, http://mahout.apache.org/