Survey Paper on Big Data and Hadoop

Varsha B.Bobade

Department of Computer Engineering,
JSPM's Imperial College of Engineering & Research, Wagholi, Pune,India

---***---

Abstract - *The term 'Big Data', refers to data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to capture, manage, process or analyzed. To analyze this enormous amount of data Hadoop can be used. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, Apache Hive, No SQL and HPCC. These technologies handle massive amount of data in MB, PB, YB, ZB, KB and TB*

Key Words: *Bigdata , Hadoop Framework, HDFS, Mapreduce , Hadoop Component*

I. INTRODUCTION

Big Data is as a collection of large dataset that can not be processed using traditional computing techniques .Big Data is not merely a data rather it has become a complete subject which involve various tools, techniques and framework. The need of big data generated from the large companies like facebook, yahoo, Google, YouTube etc for the purpose of analysis of enormous amount of data also Google contains the large amount of information

Big Data is a term that refers to dataset whose volume (size), complexity and rate of growth (velocity) make them to difficult to captured ,managed ,processed or analyzed by conventional technology and tools such as relational databases. There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data

The data in it will be of three types.

Structured data: Relational data.

Semi Structured data: XML data.

Unstructured data: Word, PDF, Text, Media Logs.

1.1 The challenges of big data

1. Volume:

Volume refers to amount of data. volume represent the size of the data how the data is large. The size of the data is represented in terabytes and petabytes.

2. Variety:

Variety makes the data too big. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more.

3. Velocity:

Velocity refers to the speed of data processing. The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive.

4. Value:

The potential value of Big data is huge.Value is main source for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.

5. Veracity: Veracity refers to noise ,biases and adnormality When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data.

2. Hadoop: Solution for Big Data Processing

Hadoop is an Apache open source framework written in Java that allows distributed processing of large dataset across cluster of computers using simple programming model Hadoop creates cluster of machines and coordinates work among them . It is designed to scale up from single servers to thousands of machines, each offering local computation and storage Hadoop consists of two component Hadoop Distributed File System(HDFS) and MapReduce Framework.

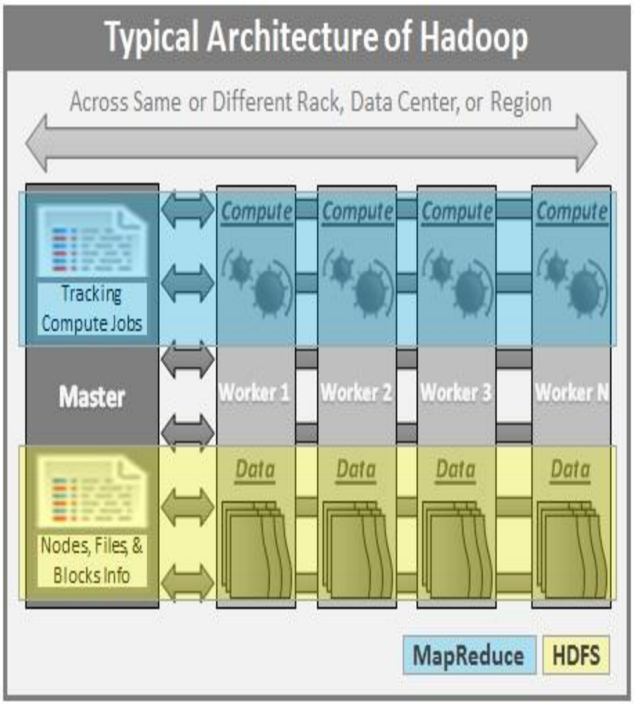


Fig -1: Architecture of Hadoop

a) HDFS (Hadoop Distributed File System)

HDFS is a file system designed for storing very large files with streaming data access pattern, running clusters on comodity hardware. HDFS manages storage on the cluster by breaking incoming files into pieces called 'blocks' and stroing each blocks redundantly across the pool of the server. HDFS stores three copies of each file by copying each piece to three different servers. Size of each block 64MB. HDFS architecture is broadly divided into following three nodes which are Name Node, Data Node, HDFS Clients/Edge Node

1. Name Node

It is centrally placed node, which contains information about Hadoop file system . The main task of name node is that it records all the metadata & attributes and specific locations of files & data blocks in the data nodes. Name node acts as the master node as it stores all the information about the system .and provides information which is newly added, modified and removed from data nodes.

2. Data Node

It works as slave node. Hadoop environment may contain more than one data nodes based on capacity and performance . A data node performs two main tasks storing a block in HDFS and acts as the platform for running jobs.

3. HDFS Clients/Edge node

HDFS Clients sometimes also know as Edge node . It acts as linker between name node and data nodes. Hadoop cluster there is only one client but there are also many depending upon performance needs .

b) MapReduce Framwork

MapReduce is defined as a programming model for processing and generating large sets of data. There are two phases in MapReduce, the "Map" phase and the "Reduce" phase. The system splits the input data into multiple chunks, each of which is assigned a map task that can process the data in parallel. Each map task reads the input as a set of (key, value) pairs and produces a transformed set of (key, value) pairs as the output. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (key, value) pairs to reduce task, which groups them into final results.

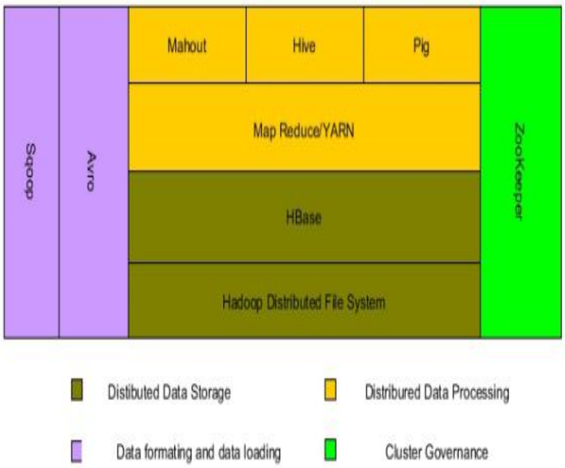
3. The Hadoop Ecosystem

1) HBases

Hbase is distributed column oriented database where as HDFS is file system. But it is built on top of HDFS system. HBase is a management system that is open-source, versioned, and distributed based on the BigTable of Google. It is Non-relational, distributed database system written in Java. It runs on the top of HDFS. It can serve as the input and output for the MapReduce. For example, read and write operations involve all rows but only a small subset of all columns.

2) Avro:

Avro is data serialization format which brings data interoperability among mutlple components of apache hadoop. Most of the components in hadoop started supporting Avro data format. It works with basic premise of data produced by component should be readily



Fig

Fig - 2: Hadoop Ecosystems

consumed by other component Avro has following features Rich data types, Fast and compact serialization, Support many programming langguages like java, Python.

3) Pig:

Pig is platform for big data analysis and processing. Pig adds one more level abstraction in data processing and it makes writing and maintaining data processing jobs very easy. Pig. can process tera bytes of data with half dozen lines of code.

4) Hive:

Hive is a dataware housing framework on top of Hadoop. Hive allows to write SQL like queries to process and analyze the big data stored in HDFS. It is Data warehousing application that provides the SQL interface and relational model. Hive infrastructure is built on the top of Hadoop that help in providing summarization, query and analysis.

5) Sqoop:

Sqoop is tool which can be used to transfer the data from relational database environments like oracle, mysql and postgresql into hadoop environment Sqoop is a command-line interface platform that is used for transferring data between relational databases and Hadoop.

6) Zookeeper:

Zookeeper is a distributed coordination and governing service for hadoop cluster It is a centralized service that provides distributed synchronization and providing group services and maintains the configuration information etc. In hadoop this will be useful to track if particular node is down and plan necessary communication protocol around node failure.

7) Mahout:

Mahout is a library for machine-learning and data mining. It is divided into four main groups: collective filtering, categorization, clustering, and mining of parallel frequent patterns. The Mahout library belongs to the subset that can be executed in a distributed mode and can be executed by MapReduce.

3. CONCLUSIONS

I have entered an era of Big Data. The paper describes the concept of Big Data along with Operational vs. Analytical Systems of Big Data. The paper also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data. In this paper we have tried to cover all detail of Hadoop and Hadoop component and future scope.

REFERENCES

[1] Andrew Pavlo, "A Comparison of Approaches to Large-Scale Data Analysis", SIGMOD, 2009.

[2] Apache Hadoop: http://Hadoop.apache.org

[3] Dean, J. and Ghemawat, S., "MapReduce: a flexible data processing tool", ACM 2010.

[4] DeWitt & Stonebraker, "MapReduce: A major step backwards", 2008.

[5]Hadoop Distributed File System, http://hadoop.apache.org/hdfs

[6] HadoopTutorial: http://developer.yahoo.com/hadoop/tutorial/module1.html

[7] J. Dean and S. Ghemawat, "Data Processing on Large Cluster", OSDI '04, pages 137–150, 2004

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

[9] Jean-Pierre Dijcks, "Oracle: Big Data for the Enterprise", 2013.

[10] Greenplum Analytics Workbench ,visit us at www.greenplum.com

[11] shilpa Manjit Kaur," BIG Data and Methodology- A review" ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.