# Storing and Processing of Big Data Using Hadoop

K.Jayanthi , Assistant Professor,

S.Samundeeswari, Professor,

COLIN HAMLET ABRAHAM·, UG Student

Department of Computer science and engineering,

PRIST University.

## ABSTRACT

To handle huge data of structured, semi-structured and unstructured data. Big data Hadoop places a vital role to reduce the data by mapping. Big data designates several tools to store, manage, and distribute huge data. Hadoop is an open source software that handles distributed the large set of data acquires several services servers. To scale up from a single server to thousands of servers various techniques are used for fault tolerance. Big data analytics has a concept to manage the life cycle of data as well as privacy and security of data. This paper reviews the data storage domain and map-reduce paradigm for large data sets.

Keywords –Big Data, Hadoop, Components of Hadoop

## 1. INTRODUCTION

Big data is Similar to data, that date was very large or complex for the tradition data processing software. Now they are lived in the data world, they lived in 100% of the data world. 90% of the data are generated in the past two years. The following are the data generator they are sensors, CCTV, social media (facebook, twitter, etc…), online shopping, Airlines, and Hospitals. Big data challenges include capturing data, data storage, data analyses, search, sharing transfer, visualization, querying, updating, information privacy and data source. Let see the Big data in Hadoop. Big Data concept is associated with three key concepts which are Volume, Variety, and Velocity.

V**olume:** The quantity of generated and stored data and its size of data is (GB, TB, PB) to generated. First, the data volume is big that the data are called Big data. In date, the volume is called (GB, TB or PB) that size data will be considered for big data.

**Variety:** Converting large volumes of transactional information into decision has always been a challenge for IT traders. Past years are generated or processed data only for structured data. The more information coming from the channels and emerging technologies mainly from social media, IOT (Internet of Things), mobile sources, documents, XML, email, image, audio video files.

**Velocity:** It is referred to both the speed with which is produced and the processing speed is called velocity. The speed of data generation and processing and analysis process are improved to speed. The velocity of the data processing must be high on big data analysis, velocity is the refer the Data StoredSpeed and processing speed is called velocity.

**Veracity:** It is the extended definition for big data, which refers to the data quality and the data value. Big data must be processed with advanced tools (Analytics and algorithm) to reveal information.

## II. HADOOP:

Hadoop is an open source framework. It is developed by Apache software solution. The founder of Hadoop Doug Cutting. The name may kid gave a stuffed yellow elephant. Shot easy to spell and pronounce, meaningless and not used elsewhere those are my naming criteria. Kids are Good at generating such Google is kid's term. Since 2010 Hadoop is adopted by the organizations for the stored and processing of large volume of data and as a platform of data analysis. The Hadoop is used to following organizations Amazon.com, eBay, Facebook, Google, Twitter, Yahoo, etc., are some companies using Hadoop. The additional software packages can be installed on Hadoop. The Hadoop products are HDFS, Map-reduce, HBase, Hive, Mahout, Oozie, Pig, Sqoop, Whirr, Zookeeper, and flume.

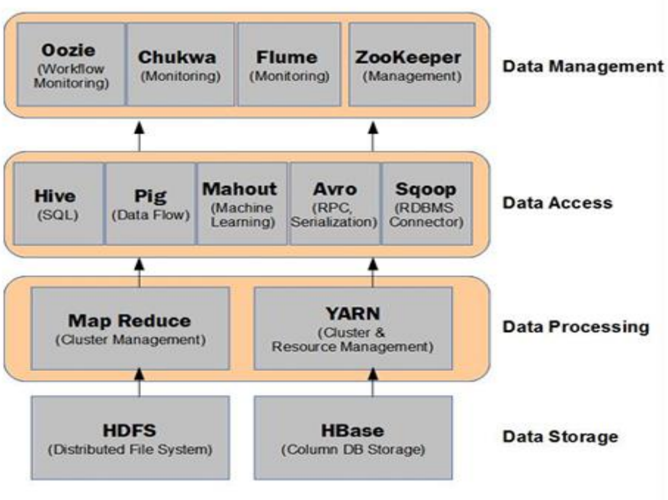
Oozie (Workflow Monitoring)
Chukwa (Monitoring)
Flume (Monitoring)
ZooKeeper (Management)
Data Management
Hive (SQL)
Pig (Data Flow)
Mahout (Machine Learning)
Avro (RPC, Serialization)
Sqoop (RDBMS Connector)
Data Access
Map Reduce (Cluster Management)
YARN (Cluster & Resource Management)
Data Processing
HDFS (Distributed File System)
HBase (Column DB Storage)
Data Storage

Fig.1 : Hadoop Architecture

**FLUME**: Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

**ZOOKEEPER:** Zookeeper is a centralized open-source server for maintaining and managing configuration information, naming conventions and synchronization for a distributed cluster environment. Zookeeper helps the distributed systems to reduce their management complexity by providing low latency and high availability.

**OOZIE:** Apache Oozie is a Java Web application used to schedule Apache Hadoop jobs. Oozie combines multiple jobs sequentially into one logical unit of work. It is integrated with the Hadoop stack, with YARN as its architectural center, and supports Hadoop jobs for Apache MapReduce, Apache Pig, Apache Hive, and Apache Sqoop

**ARROW:** Arrow Apache is an in-memory data structure specification for use by engineers building data systems. It has several key benefits: A columnar memory-layout permitting O(1) random access. ... Developers can create very fast algorithms which process Arrow data structures.

**HBASE:** HBase is non SQL oriented distributed database based on the Google Big table module and it is used to HDFS storage media.

**HIVE:** Hive used to a SQL query language and its names it HiveQL. Hive is a data storage platform by using a query.

## III. HDFS (HADOOP DISTRIBUTED FILE SYSTEM)

It is made for storing a data special file system streaming access pattern. It is a specially designed file system for storing huge data set with the cluster. HDFS default block size is 64MB and it is increased to up to 128MB. HDFS gives five services to store a data namely Name Node (NN), Secondary name node, Job Tracker, Data node, Task tracker.

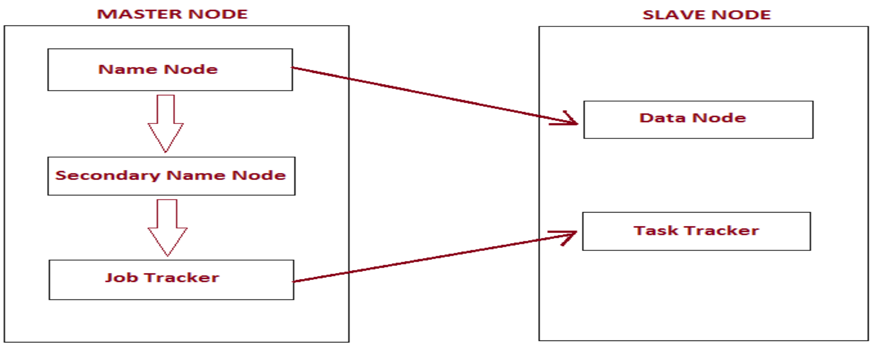

**Fig.2:HDFS Architecture**

### HDFS Master / Slave Architecture

- Single Name Node (Master Node)
- One or more data node (Slaves)

Name mode manages five systems name space :

- All clients interactions start with Name Node
- Data Node stores for as a block
- Data Node send a heartbeat or block report to Name Node
- File is broken into Blocks and stored on another data node
- Name Node maintains file to block mapping, location order of blocks and other meta data.
- Default block size is 64MB to 128 MB
- Client directly interacts with Data Node for reading / writing blocks

Name Node Data Node can be isolated on single Node cluster for learning

**THE REPLICATIONS :** The commonly HDFS provides the 3 replication if you want to increase the replication that all as can in HDFS.

## IV. MAP REDUCE

Map Reduce is a program model for processing the large set of data in distributed algorithm on a cluster. The HDFS Map reduce system which is commonly referred to as Hadoop. The basic unit of information in Map Reduce is a **(key, values)** pair. All type of data transferred to this basic unit. The Map Reduce Module is a 3 stages Map Stage, Shuffle Stage and Reduce Stage.

**MAP STAGE :** Map Stage is a stage which all other single worker nodes uses the map function to the local data and written the data output to a temporary storage. In this process master node ensure the single copy of an redandent input data which can be processed.

**SHUFFLE STAGE :** In this stage worker nodes redistribute the databse on the output keys that all the dates belong to a one key which located on the same worker node.

**REDUCE :** In this stage they process each group of output data

## V. ILLUSTRATION OF WORD COUNT PROBLEM IN MAP REDUCE

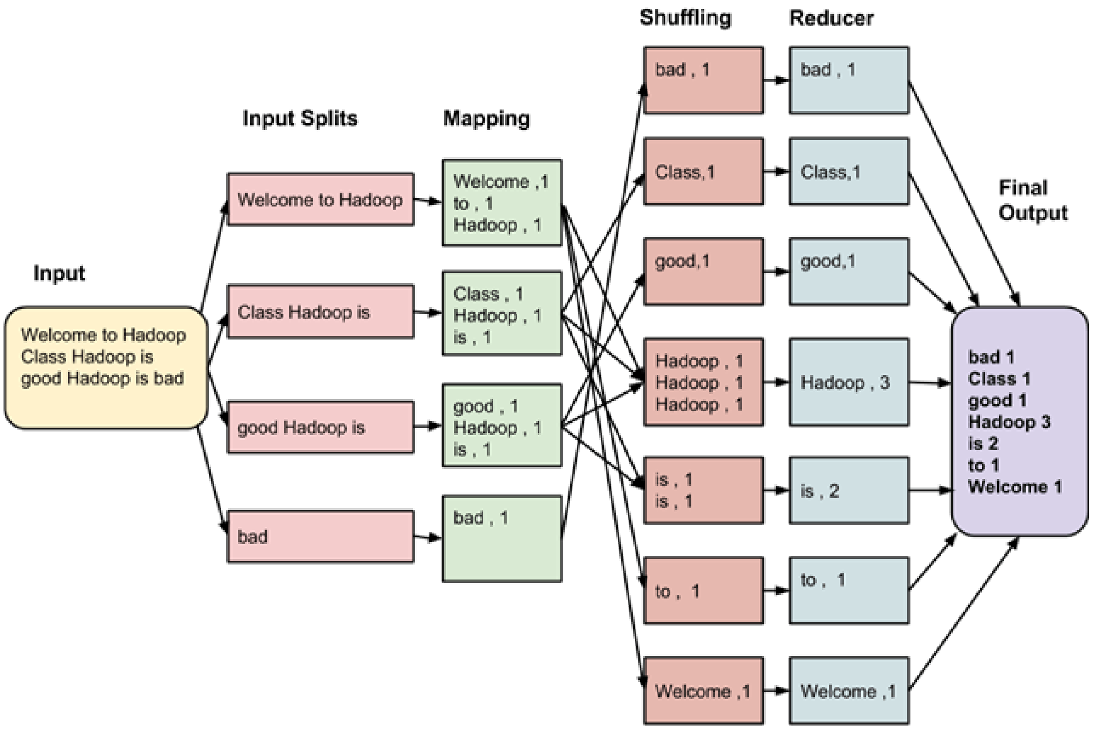
Input
Welcome to Hadoop Class Hadoop is good Hadoop is bad
Input Splits
Welcome to Hadoop
Class Hadoop is
good Hadoop is
bad
Mapping
Welcome ,1 to , 1 Hadoop , 1
Class , 1 Hadoop , 1 is , 1
good , 1 Hadoop , 1 is , 1
bad , 1
Shuffling
bad , 1
Class,1
good,1
Hadoop , 1 Hadoop , 1 Hadoop , 1
is , 1 is , 1
to , 1
Welcome ,1
Reducer
bad , 1
Class,1
good,1
Hadoop , 3
is , 2
to , 1
Welcome ,1
Final Output
bad 1 Class 1 good 1 Hadoop 3 is 2 to 1 Welcome 1

Fig3: WORD COUNT PROBLEM

The original data will be split to the input stage. The mapper is only read for key and value. Record reader is convert to the text in the form of key and value. Then the mapper is started to work on shuffling and sorting. The shuffle data will be send to map reducer then the data will be controlled send to record writer then show the output data.

## VI.CONCLUSION

Handling of huge of structures, semi-structure, unstructured data is a major issue of organization to maintain and retrieve data various big data analytics method has developed. In this paper has discussed how to handle huge data using various Big data analytics concept and illustration word count problem using map reduce technology of Hadoop.

## VII .Reference

[1] H. S. Bhosale and P. D. P. Gadekar, “A Review Paper on Big Data and Hadoop,” vol. 4, no. 10, pp. 1–7, 2014.
[2] Y. Demchenko, C. Ngo, and P. Membrey, “Architecture Framework and Components for the Big Data Ecosystem,”2013.
[3] W. Fan and A. Bifet, “Mining Big Data : Current Status , and Forecast to the Future,” vol. 14, no. 2, pp. 1–5.