Hadoop Architecture

Mr.Prathamesh Ashok Jaisingpure[1], Mr. Hrishikesh Sarode[2], Mr. Hrushikesh Ingole[3], Dr. Makarand Shahade[4]

[1]IIIrd Year BE (CSE Department), JDIET, Yavatmal- jaisingpure88@gmail.com
[2]IIIrd Year BE (EXTC Department), JDIET, Yavatmal- h12sarode@gmail.com
[3]IIIrd Year BE (CIVIL Department), JDIET, Yavatmal- hrushiingole@gmail.com
[4]Professor,RSCOE, Pune- makarandr_shahade@rediffmail.com

Abstract

The Hadoop Distributed File System (HDFS) is specially designed file system used to store very large data sets, and used to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks. By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size.

Keywords -*Big Data, Hadoop, Map Reduce, HDFS, Hadoop Components,* distributed file system

1. Introduction

A. Big Data: Definition

The term Big data refers to data sets or it is a combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. While the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data. Figure No. 1.1 gives Layered Architecture of Big Data System. It can be decomposed into three layers, including Infrastructure Layer, Computing Layer, and Application Layer from top to bottom.

B. SOURCES OF BIG DATA :-

1) SENSORS
2) SOCIAL NETWORKS – FACEBOOK
3) ONLINE SHOPPINGS
4) AIRLINES
5) BANK
6) HOSPITAL DATA

3Vs (volume, variety and velocity) :-

There are three Vs defining properties or dimensions of big data. First Volume is used to refers the amount of data, Variety is used to refers the number of types of data and Velocity is used to refers to the speed of data processing. According to the 3Vs, the big data management challenges result from the expansion of all three properties, rather than just the volume alone the sheer amount of data to be managed.

1)Volume :-

Volume refers to amount of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes.
We currently see the exponential growth in the data storage as the data is now more than text data. We can find data in the format of videos, musics and large images on our social media channels. It is very common to have Terabytes and Petabytes of Data

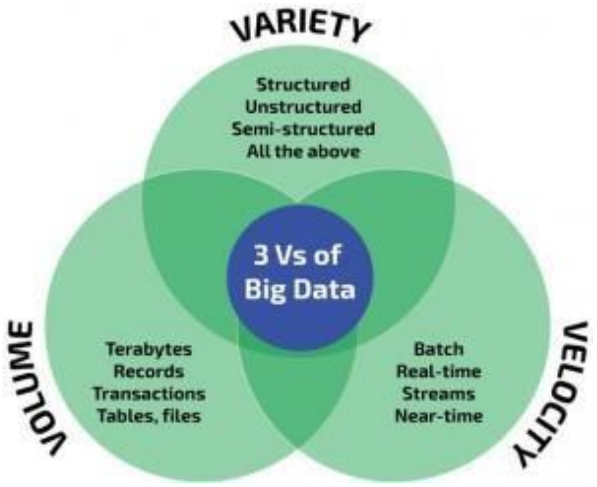
2) Velocity :-

Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. However, news channels and radios have changed how fast we receive the news. Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something interests users.
They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent Big Data

3) Variety :-

Data can be stored in multiple format. For example database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf or something we might have not thought about it. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however it is not the case most of the time. The real world have data in many different formats and that is the challenge we need to overcome with the *Big Data*. This variety of the data represent represent Big Data.



Hadoop :-

Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.

Cluster and Commodity Hardware:-

As You can see in fig. (a) Cluster is a set of machine in a single LAN and Commodity Hardware is a low reliable hardware such as our PCs and LAPTOPs etc.

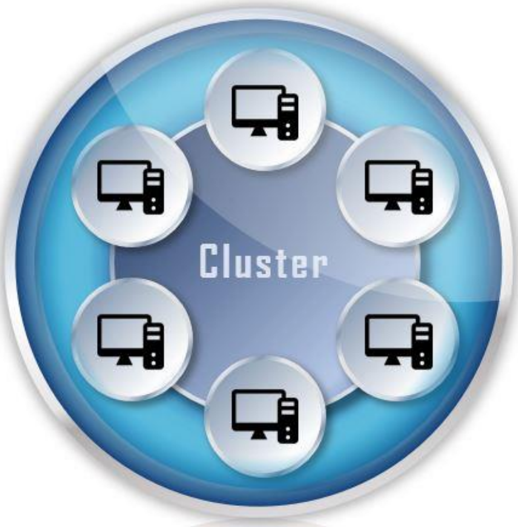


Fig (a) Cluster

HDFS Architechture:-

The Hadoop Distributed File System (HDFS) is a specially design file system to run on commodity hardware. It is same as existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is fault-tolerant and is designed to run on low-cost hardware(commodity hardware). HDFS provides high reliability to access application data which is suitable for the large data sets.

HDFS is open source framework which is a subproject of Apache software foundation.

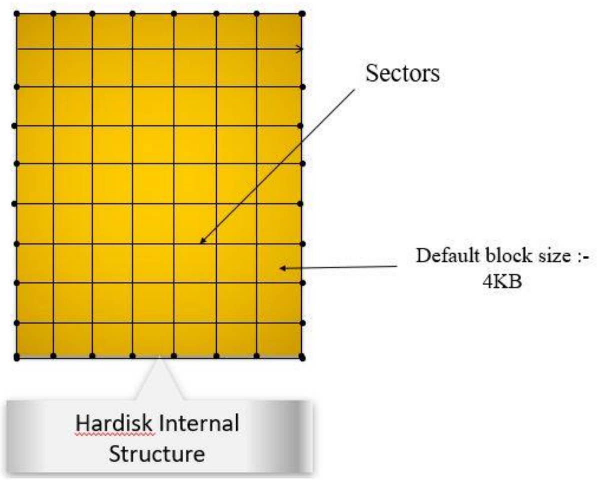
Example1:

As shown in fig. b(1), Suppose this is Hardisk Internal Structure whose size is 500GB, all lines shown are what sectors.

In this 500GB of Hardisk, there are serval blocks whose default block size is 4KB.

If your are storing some data in this hardisk and after storing if there is some remaing space so that space will not be used for some other files means one sector will not be used again for some other work.

Here we have to maintain more no. of hardisk or systems and one sector will not be used again for some other work.



This is what our Normal File System(NFS) is.

Fig. b(1) NFS - Hardisk Internal Structure

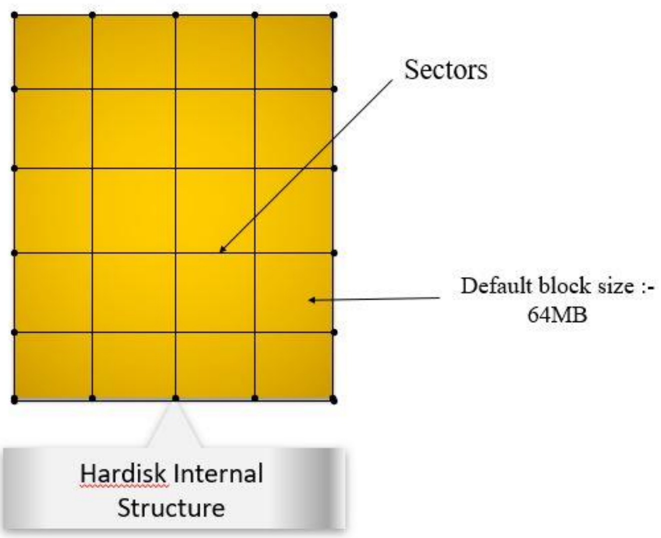


Fig. b(2) HDFS - Hardisk Internal Structure

Example2: As shown in fig. b(2), Suppose this is Hardisk Internal Structure whose size is 500GB, all lines shown are what sectors.

In this 500GB of Hardisk, there are serval blocks whose default block size is 64MB

If your are storing some data in this hardisk and after storing if there is some remaing space so that space will be used for some other files means one sector will be used again for some other work.

This is what our Hadoop Distributed File System(HDFS) is.

MapReduce:-

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce.

The Mapreduce is one of the main component of the Hadoop Ecosystem. MapReduce is designed to process a large amount of data parallel by dividing work into some smaller and independent task.

The whole job is taken from the user and divided into smaller tasks,and assign them to the worker nodes

MapReduce programs take input as a list and convert to the output as a list also

Hadoop:- For storing and processing here we having two core concepts in Hadoop.

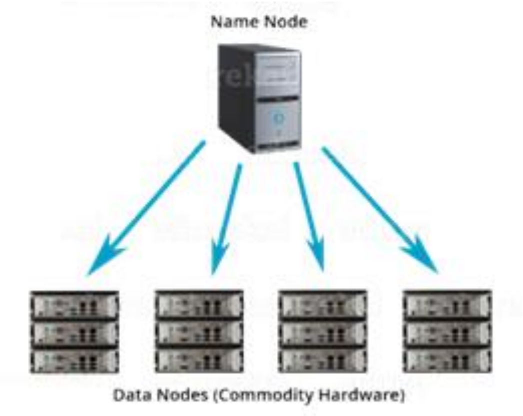
i) HDFS

ii) MapReduce

Hadoop is used for storing and processing huge data sets, so we can say that Hadoop is a combination of HDFS and MapReduce, as HDFS is used for storing data and MapReduce is used for processing a data.

5 Services of HDFS:-

NameNode: It is the master daemon that maintains and manages the data block present in the DataNodes.



Secondary NameNode: The Secondary NameNode works concurrently with the primary NameNode as a helper daemon. It performs checkpointing.

Job Tracker: Job Tracker is the master daemon for both Job resource management and scheduling/monitoring of jobs. It acts as a liaison between Hadoop and your application.

DataNode: DataNodes are the slave nodes in HDFS. Unlike NameNode, DataNode is a commodity hardware, that is responsible of storing the data as blocks.

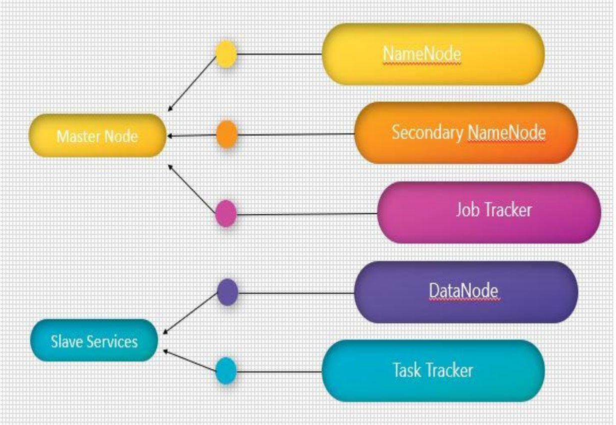


Fig. 3(a) 5 Services of Hadoop

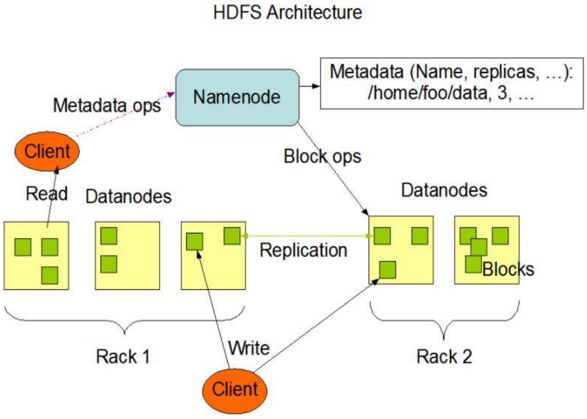


Fig. 4(a) HDFS Architecture

Conclusion:-

We have entered an era of Big Data and Hadoop. The paper describes the concept of Big Data along with 3Vs (Volume, Velocity and Variety of Big Data) and Hadoop describes the concept of HDFS and MAPREDUCE.
The paper also focuses on Hadoop storing and processing technique. Here in this paper we have discuss HDFS and MAPREDUCE with data storing example.
The paper describes Hadoop which is an open source software used for processing of Big Data.

TaskTracker : A TaskTracker is a node in the cluster that accepts tasks - Map, Reduce and Shuffle operations - from a JobTracker. TaskTracker runs on DataNode, Mostly on all DataNodes

REFERENCES

[1]. Kiran kumara Reddi & Dnvsl Indira **"Different Technique to Transfer Big Data : survey"** IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}
[2]. Jimmy Lin "**MapReduce Is Good Enough**?" The control project. IEEE Computer 32 (2013).
[3] Shipa, Manjit kaur, "Big Data and Methodology", 10 Oct, 2013

[4] Pareedpa, A.; Dr.Antony Selvadoss, "Significant Trends of Big Data", 8 Aug, 2013
[5] Gurpeet Singh Bedi, Ashima, "Big Data Analysis with Dataset Scaling in Yet another Resource Negotiator
(YARN)", 5April, 2013
[6] Hadoop-The Definitive Guide, Tom White, Edition-3, 27Jan, 2012
[7] Mrigank Mridul, Akashdeep Khajuria,Snehasish Dutta,Kumar N, "Analysis of Big data using Apache Hadoop and MapReduce",Volume 4, May 2014
The Hadoop Distributed File System: Architecture