Big Data Analysis Using Hadoop Framework

Ashwini A. Pandagale & Anil R. Surve
Shivaji University,
Department of Computer Science and Engineering,
Walchand College of Engineering,
Miraj Road, Vishrambaug, Sangli – 416415.

Received Feb 26, 2015 Accepted March 10, 2016

ABSTRACT Hadoop is an open source framework overseen by Apache Software Foundation.Earlier yahoo,google developed their file system (GFS) Google File System.Before 5-7 years data generated in a year is generating in a day.Traditional analysis system can't solve the timing issue and storage capacity of large data set which is in Gigabytes,Terabytes.Hadoop framework is the solution for this problem.It is the parallel distributive system running on hundred or thousand number of nodes.As this system is flexible in nature,uses low cost hardware,scalable and most important is the fault tolerance.This paper takes review of Hadoop Framework which gives details about how data get stored and processed.It also gives idea of various services,terminologies and required components.

***Key words*:** Hadoop Framewok, Big data problem, HDFS, MapReduce.

Introduction

Big Data concerned with large-volume, complex, growing data sets generated from an autonomous sources stated by Xindong Wu, Fellow and S. M. Wei Ding. Extremely large datasets can not be processed using centralized system.As Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences facing the problems of time required for processing,analyzing the data moved to distributed system.Massively parallel systems such as Hadoop can be run on hundreds or thousands of nodes is the answer for Big data challenges.Most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. We have to do an implementation of Hadoop cluster, HDFS storage and Map Reduce Framework for processing,large data sets by considering prototype of Big Data application scenarios .B. Aditya B. Patel and U. Nair explains about different methods for implementation addressing problems arised during the work . This paper explains the experimental work on Big Data challenges to extract meaningful information and its optimal solution using the Hadoop Framework.

Hadoop Framework

Hadoop

Hadoop is an open source framework which is overseen by Apache Software Foundation,so called as Apache Hadoop given as deails on http://hadoop.apache.org.. As company invest much prefers open source software.Hadoop provides storage and processing stored data using map reduce.

A) Hadoop Distributed File System(Hdfs) Architecture

Dipayan Dev; Ripon Patgiri explained HDFS,Hadoop Distributed File System is a specially designed file system which stores

data for processing with cluster of commodity hardware and with streaming access pattern . It provides high throughput access to large data sets. HDFS is a master/slave architecture. Namenode is a Master node and Datanodes are slave nodes.

(a)Commodity Hardware

Commodity hardware such as personal computers or laptops used in day to day life costing less and beneficial to use economically.

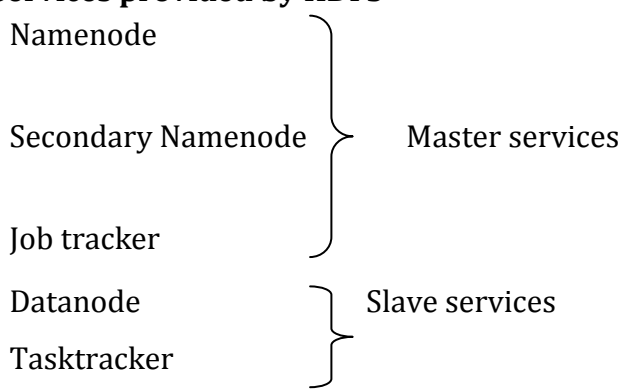
b) Cluster

The set of nodes(Personal Computers) arranged in single LAN(Local Area Network).

c) Streaming access pattern

HDFS is built around the idea that the most efficient data processing pattern is a write-once, read-many-times without changing the contents of file stored on HDFS.

Services provided by HDFS



Master services can communicate to Slave services and vice versa.Also Namenode can communicate Datanode and vice versa.And Jobtracker can communicate with Tasktracker and vice versa.

(1) Some Terminologies:

i. Namenode

Namenode is the focal point of an HDFS file system. It maintains the metadata I. e. the size, type, namespace information and block information.It stores data in main memory and also stored on disk for persistence storage. Drawback of Namenode is if it gets crashed then the whole system will get down .

ii. Secondary Namenode

To overcome the crashing problem of Namenode ,Secondary Namenode comes into existence.The purpose behind secondary Namenode have a Checkpoint in HDFS. So also called as a Checkpoint node .

iii. Datanode

Datanode is used to store data in the Hadoop File system. When Namenode request, it responds to File system operations .

iv. Job tracker

Job tracker keeps control on Tasktracker. It is the process which performs resource management (Managing the task trackers), tracks the resource

availability and task life cycle managemen(Tracking its progress, fault tolerance, etc.) .

v. Tasktracker

Tasktracker follows the order of Jobtracker and updating with progress status periodically.

(2) How large dataset stored in HDFS?

Assumptions:

Consider a file of 200 MB and file name as file.txt. Actual file size is in GB/TB/PB/ZB.For understanding purpose consider small sized file.Block size=64 MB. As block size is 64 MB, Number of blocks=4.Out of 4,3 blocks are of 64 MB and 4th one will be about 8 MB.Let S1,S2,S3,S4 are the 4 blocks .

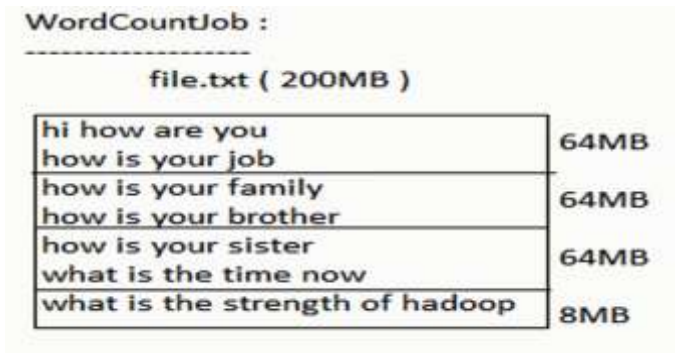


Figure 1(a): Sample Input File

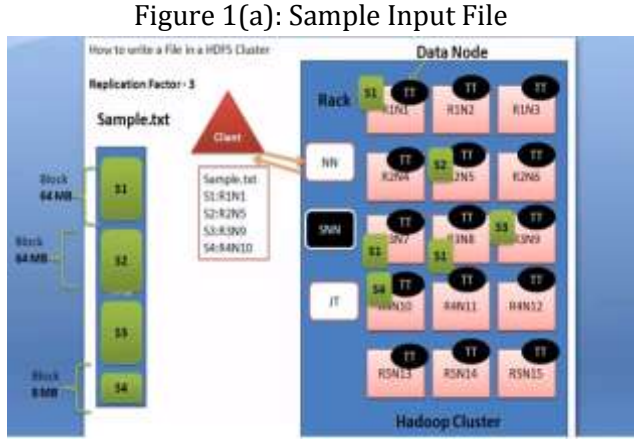


Figure 1(b): Writing a File in a HDFS Cluster

Steps:

1. The client sends request Namenode which is manager to store large dataset, say 200 MB in a cluster.As the client is unaware of free space available to store the data at datanode.

2. Namenode is having metadata. It checks metadata (for available free space)and responds to client with free space is available at datanode say R1N1,R2N5,R3N7,R3N8,R3N9,R4N10 etc.

3. As replication is used for backup and replication factor=3.S1 block will be stored on R1N1 ,R3N7,R3N8 datanode.Similarly S2,S3,S4 block will be stored on available datanodes .

B)MapReduce Flowchart

To process the data, two functions are used:map and reduce .

Following are the steps to process the data.

1. Job tracker has a program to run the required job.But it is unaware of file.txt information on datanode as namenode and datanode can not communicate with each other.So,Tasktracker requests Namenode to get information about file. Namenode provides the required information through metadata it has.

2. Now,Jobtracker has required information and it orders to Tasktracker to process the data by applying the program ,it has.Datanode follows the orders and process the dataThis whole process of applying program on the file called as map process .

3.Before processing an input file divided into blocks.Number of blocks depend upon size of block.Number of blocks equal to the number of input splits.

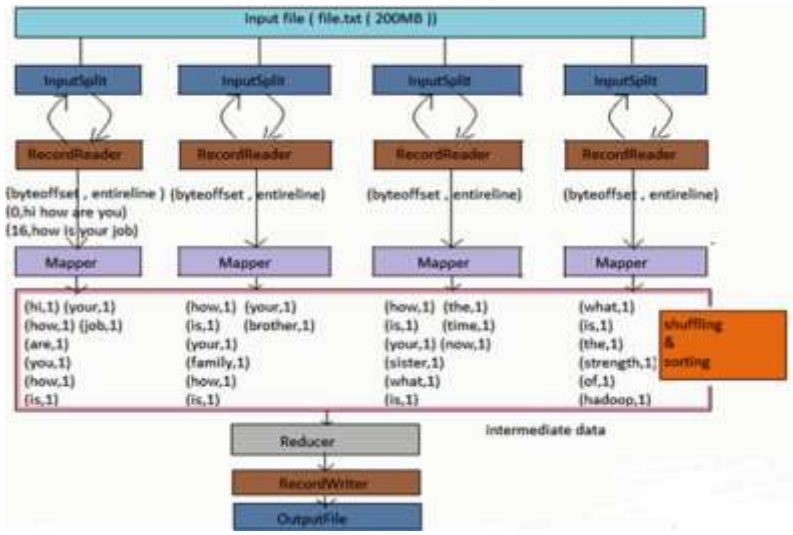


Figure 2: Map Reduce Flowchart

a) Input split

Logical representation of the data stored in file blocks known as input splits.

b) RecordReader

This is an interface between input split and mapper.As mapper understands only <key,value> pattern.To convert input text into <key,value>,record reader is used.

Number of input splits equal to number of RecordReader and number of mappers as shown in Fig. 2.

How data is processed in Mapreduce?

Steps:

1) As input file is divided into number of input splits.

2) RecordReader takes input from Input split and keeps the entire statement in the form of (byteoffset ,entire line). Byteoffset is the address for the record.The lines of text taken to convert into <key,value> format are known as a Record.

3)File format decides the basis of conversion of record into <key,value> pair.

4)Mapper takes input from RecordReader and convert the record into <key,value> pair.It generates intermediate data.On this intermediate data,shuffling and sorting has to be done.

5)In shuffle phase ,number of words which are redundant get combined .For example,in input record "how" appears five times.It will output as (how ,1,1,1,1,1).In this manner all words will be shuffled.

6)In sort phase,output of shuffle phase taken as input and arranged in alphabetical order

or according to the count of the value field in <key,value> pair.

7)At the last phase of Reducer,the count of words will be given as final output .In above example, how 5 will be output .

Conclusion

Hadoop Framework has gained so much popularity as it is a remedial solution for large data sets.As a traditional analyzing method was storing restricted and time consuming. In this paper, a brief idea about Hadoop Framework and its core components are explained.Storage of data and processing stored data gives brief idea about how both phases run in reality.

References

1. B. Aditya B. Patel and U. Nair, \Addressing big data problem using hadoop and map reduce,,"NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING, NUiCONE-K.
2. Characteristics of Big Data - http://www.datatechnocrats.com/tag/big-data/
3. Dipayan Dev & Ripon Patgiri, "Performance evaluation of HDFS in big data management",High Performance Computing and Applications (ICHPCA), 2014 International Conference on Year: 2014Pages: 1 - 7, DOI: 10.1109/ICHPCA.2014.7045330IEEE Conference Publications X. Z. S. M. G.-Q.
4. Hadoop distributed _file system,http://hadoop.apache.org.
5. Hemant Hingave & Rasika Ingle, "An approach for MapReduce based Log analysis using Hadoop", IEEE SPONSORED 2'ND INTERNATIONAL CONFERENCE ON ELECTRONICS AND COMMUNICATION SYSTEMS (ICECS '2015)
6. Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.
7. Vasiliki Kalavri & Vladimir Vlassov(2013) "MapReduce: Limitations, Optimizations and Open Issues", 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications.
8. W. Xindong Wu, Fellow and S. M. Wei Ding, \Data mining with big data,," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.,,, vol. 26, no. 1, JANUARY 2014M.
9. Yahoo hadoop tutorial"http://developer.yahoo.com/hadoop/tutorial/"

Pleasure in the job puts perfection in the work.

~ Aristotle