BIG DATA STORE AND SECURE PROCESS IN HADOOP DISTRIBUTED FILE SYSTEM IN HADOOP FRAMEWORK

Jaswinder Singh*, Ashwani Sethi
*Research Scholar, Department of Computer Applications, Guru Kashi University, Talwandi Sabo, Bathinda, Punjab, India
Professor, Department of Computer Sciences & Engineering, Guru Kashi University, Talwandi Sabo, Bathinda, Punjab, India

Abstract

Traditional Database cannot store Big Data in an appropriate way because big data large data and a variety of data come from different Internet communication devices. Daily more data collect at web server from different social sites by client's viewers and transactions. Due to that data not in a structure all data together in unstructured form. Unstructured data not possible store in row and column wise or in tabular form. The Purpose of this paper how to Big data store and secure process in Hadoop Framework with Hadoop distributes file system.

Keyword: Hadoop, HDFS, Map Reduce

Introduction

In these days online network communication has been built, due to product shopping, online, blog and web post has been started with the internet. Daily more people send information to other people via internet through social web sites as like Facebook, tweeter and billions video upload and download on YouTube and some people purchase and sale goods with online web shopping sites. That's why structured and unstructured data together on the internet. The traditional database system has not ability to manage millions bytes of data. [1]

Big data is the drawback of managing a large quantity of "unstructured" data. The quality of massive knowledge involves a brand new kind of software package agglomeration and hardware organization. At the start of this century, studies reportable huge growth of data that exceeded Moore's Law. [2]

Hadoop Distributed File System provide a partition of data file and computation through several of web hosts, and execution application computations in parallel systematic way.[3] When we process Big data in the Hadoop Framework in Map Reduce algorithm then the first data stored in different file in HDFS and again it process in the Map Reduce Framework.

Big Data

"Big Data" which needs a computation ability of process large data labelled in pet bytes [4] today massive data sources are present throughout the world. Data used for processing could also be

obtained from measurement devices, radio frequency identifiers, social network message flows, earth science data, remote sensing, and location data streams of mobile subscribers and devices and audio and video recordings.[5] Big data may be a data, available at heterogeneous,autonomous sources, in extreme great amount, that get updated in fractions of seconds [6]. Declared that huge data [7] may be a heterogeneous mix of each structured and unstructured that needed massive space for storing and effective ways to manage. Marko Grobelnik [8], [9]

declared that there would be most growth of digital knowledge to thirty-five trillion gigabytes in 2020. Huge data have become the new line of today's information management by generating and intense great amount of system data. The challenge offered by this classification is expounded to data management. Management challenges typically check with secured data storage.

Map Reduce

The Map Reduce user specifies 2 functions known as Map and reduces, that operate on data arranged in key/value pairs. The primary step is data splitting that is completed by the Map Reduce frameworks. The splits are then processed by Map functions that are ordinarily applied to every line of each split. [10] The Map Reduce framework work as like a machine learning and batch learning. When as data scan in map reduce its work as a complete learning model. The algorithm set of key-value pairs of HDFS data is processed a typical batch-oriented workflow. This method is illustrated in Figure1

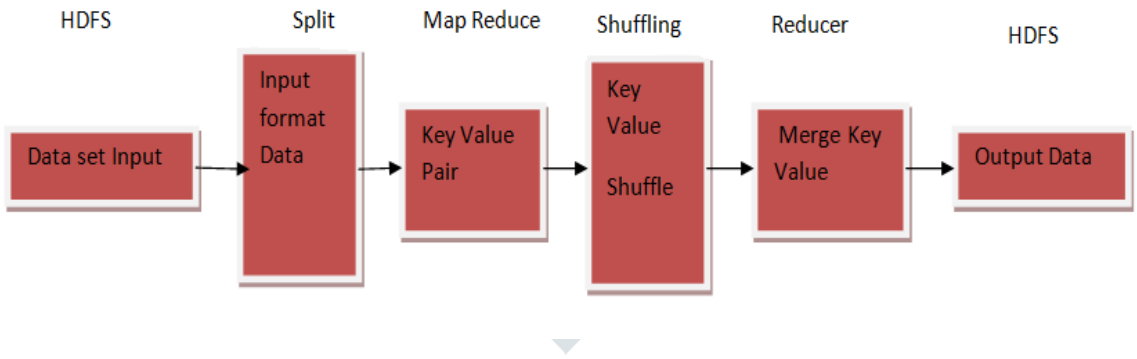

Figure: 1 Map Reduce Machine Learning Process

HDFS

HDFS refer to a Hadoop distribute file system. It is a different type of file system comparable to the traditional database file system. It cans ability to store data with distributed way on different computers commodity hardware. HDFS designed master slave structure, where on Name Node work as like a master and Data Node work as like a slave, Name Node retrieve information any time from Data Node. Data file no chance to lose or damage data. [11]

Hadoop Framework

Hadoop consists of 2 components, namely, HDFS and Map Reduce engine, of parallel processing functions. The cluster consists of various numbers of daemon/servers. Data Node stores the data in HDFS, provides the situation of data, and connects to Name Node, that doesn't keep the data, however just stores the data which will be mapped from file to block and site of block. A secondary Name Node assists in observing the state of the cluster just in case the latter fails. Integration happens between HDFS and Map Reduce that is performed through master and slave nodes. Typically, a master node consists of Task Tracker and Job Tracker from the Map Reduce layer, whereas slave nodes comprise only the Task Tracker from this layer (Figure 2). Job Tracker is liable for programming Map to Reduce jobs and assigns the tasks to the Task Trackers that are used for execution in every slave that's used for execution in each slave node. [12]

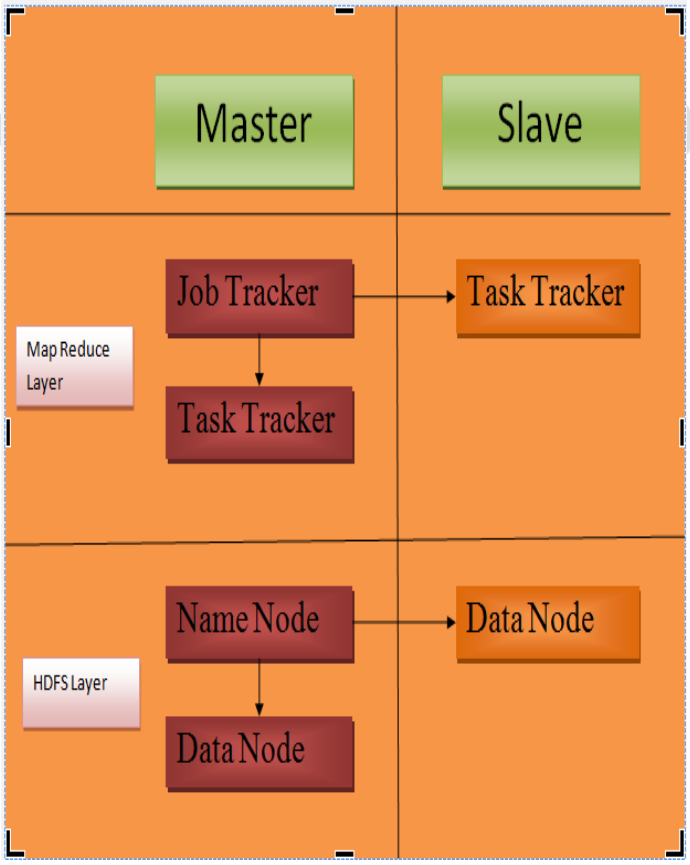

Figure: 2 Hadoop distributed file system and Map Reduce layer in Hadoop.

Experimental Step: HDFS store and secure process big data dataset .First of all I installed Oracle VM (Virtual Box Machine) at PC. And one takes a log file (200Mb) of the kind a Big data. Further Process at VM machine in the Hadoop Framework for a same result how to data process in Job Tracker and how make name node and a slave as a node in as cluster. After processing the file in Hadoop machine result show that one name node Active and further 30 Under-Replicated Block(slave node) where data process . All data processed on slave node fully secure. If one slave node fail data, not loss. Data can recover from other nodes. Name Node show in Blow Figure 3 and Figure 4.

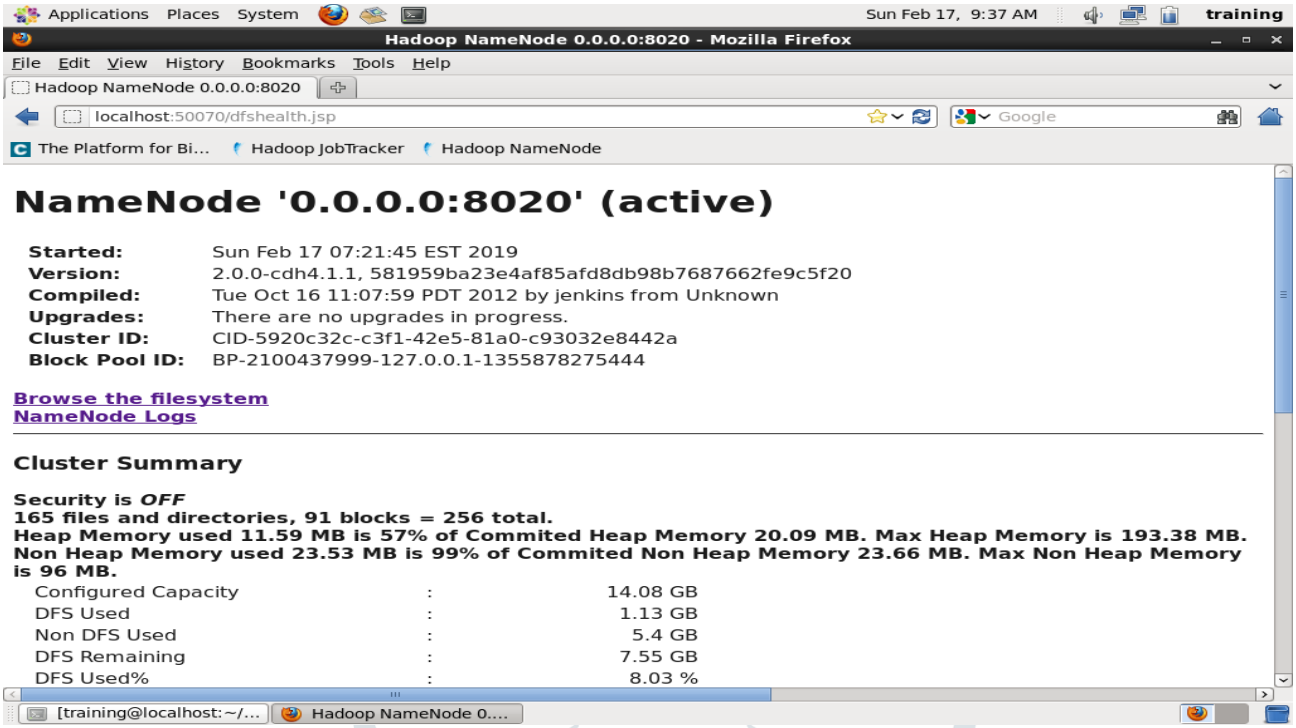

Figure: 3 Hadoop Name Node

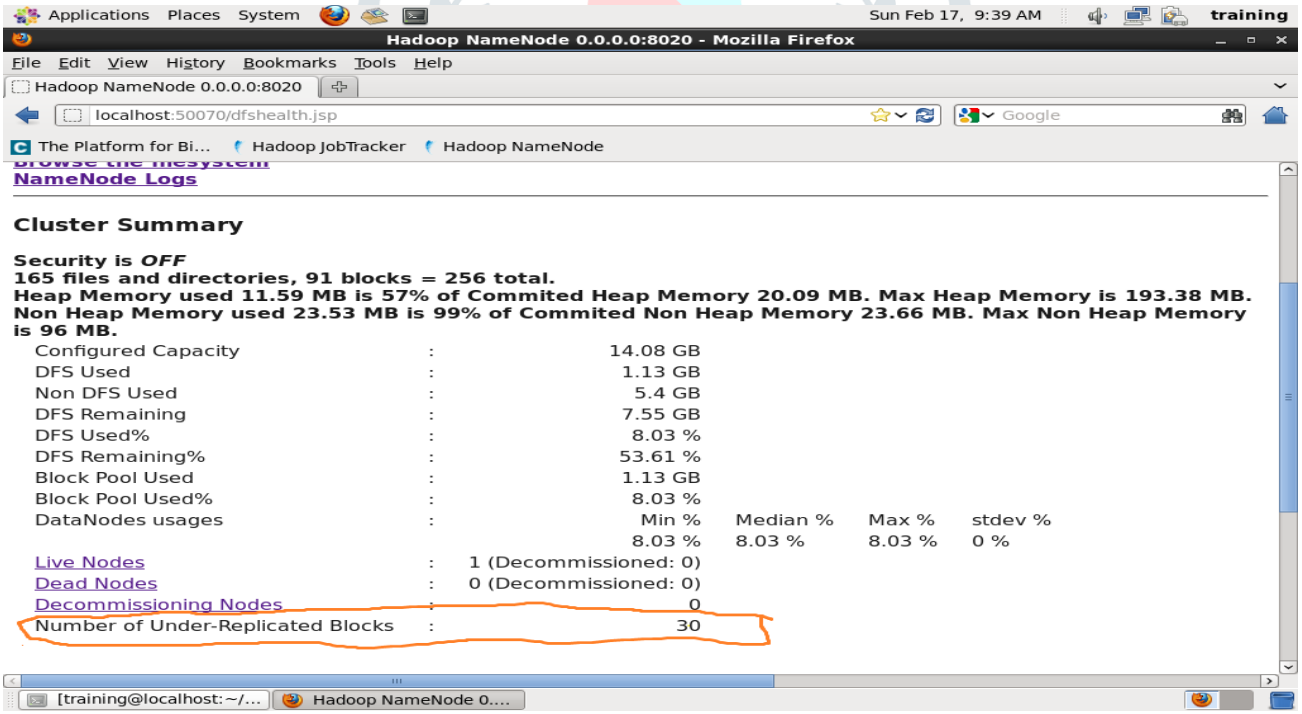

Figure: 4 Hadoop Name Node (Replicated Blocks in Circle)

Conclusion & Results

As shown figure 3 and figure 4 the proposed work experiment take one 200Mb text Log file in Hadoop Framework. HDFS (Hadoop distributed file system) data file divide in more clusters. As a Show cluster

Summary 165 files and 91 directories, total 256, And Hadoop Name node one Name Node, create 30 replicated blocks for processing a file data.

Here we see that all data process in Replicated blocks no chance of loss important data. Unfortunately, if any one file corrupt. We can recover data file from another block. So, always our data securely store and process in Hadoop distributed file system.

Reference:

[1] Apoorva Gupta," Big Data Analysis Using Computational Intelligence and Hadoop: A Study", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)978-9-3805-4416-8/15/IEEE.

[2] Usamah Algemili,"Investigation of Reconfigurable FPGA Design",2016 IEEE 2nd International Conference on Big Data Security on Cloud, DOI 10.1109/Big Data Security –HPSC-IDS.2016.75

[3] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler," The Hadoop Distributed File System", 978-1-4244-7153-9/10$26.00@2016 IEEE

[4] Hadeel T. E Kassabi,IKbal Taleb,Mohamed Adel Serhani,Rachunda Dssouli"Policy- based Qos enforcement for adaptive Big Data Distribution on the Cloud", 2016 IEEE Second International Conference on Big Data Computing Service and Applications.

[5] Ekaterina Olshannikova, Aleksandr Ometov ,Yevgeni Koucheryavy, Thomas Olsson,"Visualizing Big Data With augmented and Virtual reality: challenges and research agenda", OlshanniKova et al. Journal of Big Data(2015)2:22 DOI 10.1186/s40537-015-0031

[6] J. Ram Singh, V.Bhuvaneswari," Data Analytic on Diabetic awareness with Hadoop Streaming using Map Reduce in Python", 2016 IEEE International Conference on Advances in Computer Applications

[7] Vibhavari Chavan , "survey paper on Big Data" (IJCSIT) International Journal of Computer Science And Information Technologies, vol. 5 (6) , 2014, 7932-7939

[8] Puneet Singh Duggal, "Big Data Analysis: Challenges and Solutions", November, 2013.

[9] Nawsher Khan "Big Data: Survey, Technologies,Opportunities, and Challenges", Hindawi Publishing Corporationthe Scientific World Journal Volume 2014, Article ID 712826

[10] Tomasic, A.Rashkovska and M. Depolli,"Using Hadoop Map Reduce in a Multicluster Environment", MIPRO 2013, May20-24;2013,Opatija Croatia.

[11] Sara Land set, Taghi M. Khoshgoftaar, Aaron N. Richter and Tawfiq Hasanin," A survey of open source tools for machine learning with big data in the Hadoop ecosystem", Land set et al. Journal of Big Data (2015) 2:24,DOI 10.1186/s40537-015-0032-1

[12] Amin Mohebi, Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan and Ram Yahyapour, "Iterative big data clustering algorithms: a review" 7 July 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.2341.