Hadoop Role Displaying Population Statistics

Salem Abdulali
Department of Computer Science,
Institute of Materials and Engineering,
Kastamonu University/Turkey

Weiam S.Elsaghair
Graduate School of Electrical and Computer Engineering Altınbaş university
Istanbul/ Turkey

Abstract
Hadoop is the big data processing platform for businesses. Hadoop technology for the analysis of massive files. Hadoop is an open source, Java-based programming system that allows extremely large data sets to be processed and stored in the distributed computing environment. It facilitates big data analysis by solving the challenges in managing big data. As small elements can efficiently and rapidly be analyzed, Hadoop can break down large computer problems into smaller tasks. Hadoop is a software stack open source for data storage and running applications on commodity hardware clusters. It provides massive storage, enormous processing capacity and the ability to handle almost unlimited concomitant tasks, for different kinds of data. All these pieces are parallelly analyzed and the analysis results are grouped in order to generate the final output.. This article aims at presenting a comprehensive description of the entire Big Data program that consists of many phases and main elements for each level of big data processing.

Keywords. *Big Data, Hadoop Architecture , Mapreduce, Hadoop Ecosystem, Apache Hadoop.*

1. Introduction

Big data is an unstructured large set of data which is insufficient to handle traditional data processing application software. Big data issues are captured, stored, recorded, analyzed, data-searched, shared, transferred, viewed, queried and updated, and privacy. The three big data dimensions known as length, variety and speed. [1]The provision of more accurate analyses by major data technologies may result in more practical decision-making leading to increased operational efficiency , cost savings and reduced risks for the company. Hadoop is an open source framework that stores and processes large data in an environment distributed[2]. It is designed to scale up thousands of machines from single servers, each of which provides local calculations and storage.

Born to improve usage and solve major Big Data problems. Apache Hadoop. The Web media produces immense information every day, and details on about one billion pages of contents have become very difficult to handle. Apache Hadoop is a software platform for distributed storage and distribution of unstructured data sets on commodity hardware built computer clusters. Data storage, data processing, access to data , data administration, security and operations are provided by hadoop services. Scalability is Hadoop's greatest strength. It improves without any inconvenience from working with a single node to thousands. Data management is provided by various domains of Big Data including videos, texts, transactional information, sensor data, statistical data, conversation with social media, search engine queries, ecommerce data , financial data, weather data, news updates, forum discussions, executive reports and so forth[3].

2. Material and Method

2.1 HADOOP ECOSYSTEM

The large volume of data generated by billions in online activities and transactions needs to be constantly improved and big data evolved. The Hadoop environment is a place for different types of diverse and emerging instruments and techniques. Mapreduce and the distribute file system (Hadoop Distributed File System, HDFS) are two components of Hadoop 's broader data management ecosystems. All of these elements allow users in real time to process large data sets and tools for supporting various Hadoop projects, planning and management of cluster resources [4].

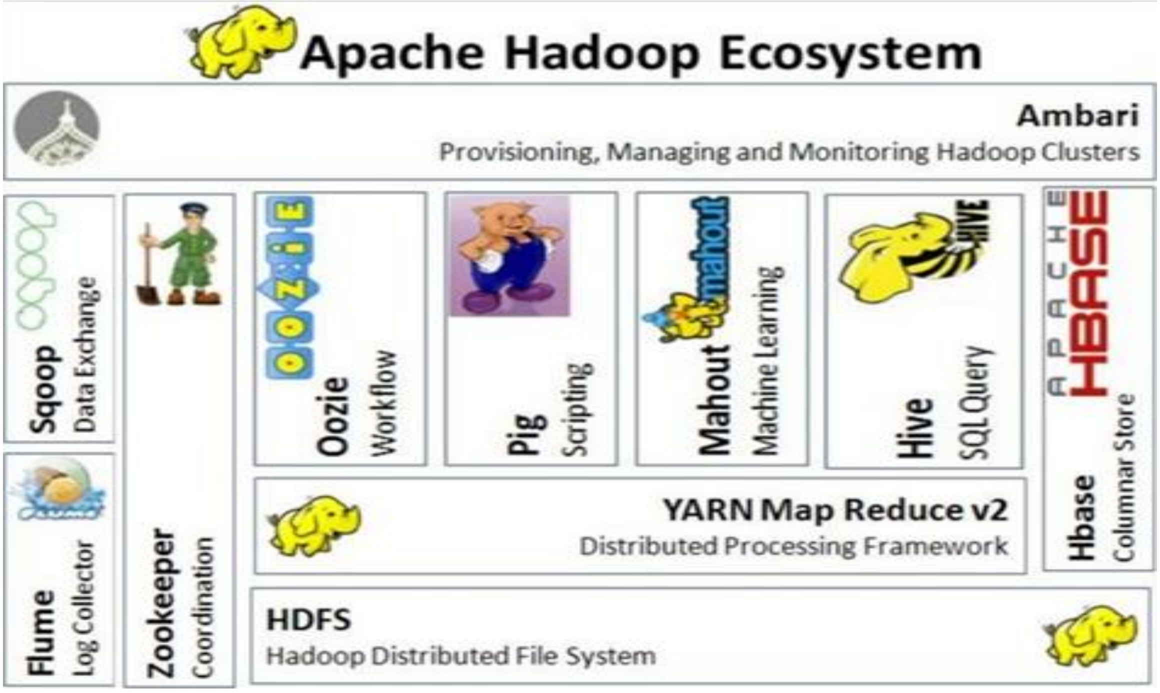

Fig 1: Apache Hadoop Ecosystem

As shown in Figure1, Hadoop frame contains four modules [5]:

a) Hadoop Common: These are Hadoop modules' Java libraries and utilities. These libraries contain abstractions of the file system and of the OS level containing the required Java files and scripts for starting Hadoop.

b) Hadoop YARN: The system for organizing research and control of cluster capital.

c) Hadoop Distributed File System (HDFS): a distributed file system that provides access to application information with high throughput.

d) Hadoop Map Reduce: YARN-based parallel processing framework for large sets of data.

Hadoop requires a thorough understanding of the diverse components of the Hadoop architecture in each layer of the Hadoop Stack and the design of a Hadoop cluster that is in charge of data processing.

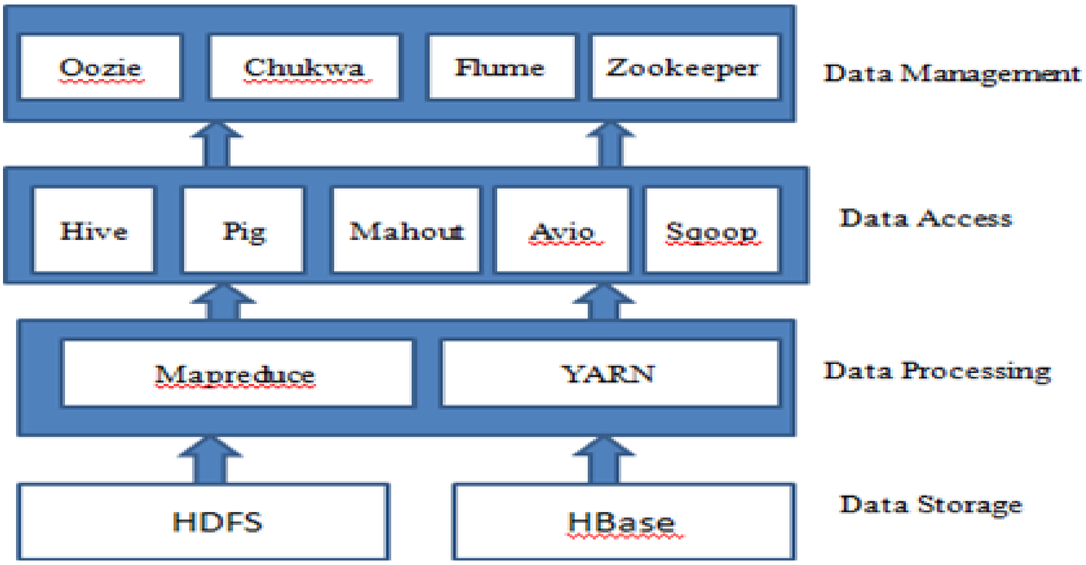

Fig 2: Hadoop Ecosystem Elements at various stage of Data Processing

The Hadoop module includes other projects that include:

A) Apache Ambari: it is the instrument for Hadoop cluster management , monitoring and provision. The HDFS and Map Reduce programs can be supported.

b) Cassandra: a distributing system that manages an enormous amount of data stored on a number of commodity servers.

c) HBase: it is a distributed, non-relation database management system, operating efficiently and highly scalable in sparse data sets.

d) Apache Spark: The big data program is highly agile, scalable and secure, and operates in a wide variety of applications including real time, machine learning and ETL is flexible.

e) Hive: it is a data store tool that is used primarily in analysis, searching and summarization on top of the Hadoop framework, analyzed data concepts.

f) Pig: Pig is a high-level framework that enables us to work with Apache Spark or Map Reduce for data analysis.

g) Sqoop: The frame for transferring data from relation databases to Hadoop is used. It is a command-line interface based program.

h) Oozie: The workflow management system is planned to execute workflow routes on Hadoop to complete the task successfully.

I) zookeeper: is a centralized Open Source service used to coordinate distributed Hadoop applications.

3. HADOOP FRAMEWORK

3.1 Hadoop Distributed File System

File systems such as HDFS are designed to tackle big data challenges. Hadoop, MapReduce and HDFS are a core component that ensures high stability. Hadoop stores data petabytes with HDFS and manages them. With HDFS, the hardware or network of commodities often called nodes can be linked. These nodes are connected to a cluster that distributes the data files to. With the capacity of the entire HDFS cluster and the nodes, data storage and processing using the MapReduce method can be reached easily[6]. The distributed file system (Hadop Distributed File System -HDFS) is based on the Google File System (GFS) and offers a distributed file system designed to run efficiently and defect-tolerantly on large clusters of small computers. HDFS is a file system designed for the removal of large documents with streaming access to information.

HDFS uses a master / slave architecture in which master consists of one NameNode that maintains metadata for the file system and one or more DataNodes, which saves actual information. One file is divided into several blocks and the blocks are stored in a number of DataNodes in an HDFS name space. In NameNode the block mapping to DataNodes is determined. The Filesystem DataNodes recognize read-and-write operation. Block creation, removal and replication is managed based on a NameNode instruction. A shell like other file systems and a list of available commands to interact with the file system is provided by HDFS.

3.2 Hadoop Working

Hadoop is based on the Master Slave Architecture as shown in Fig2 with HDFS and Map Reduce data storage and distribution. Hadoop HDFS is the main node to store data and Task Tracker[7] is the primary node for storing data parallel with Hadoop Map Reduced. Hadoop architecture's slave nodes are other Hadoop cluster machines which store data and make difficult calculations. Growing slave node has a task tracker and data node that sync the task and name node processes. The master or slave structures can be set up in the cloud in the Hadoop architectural implementation[8].

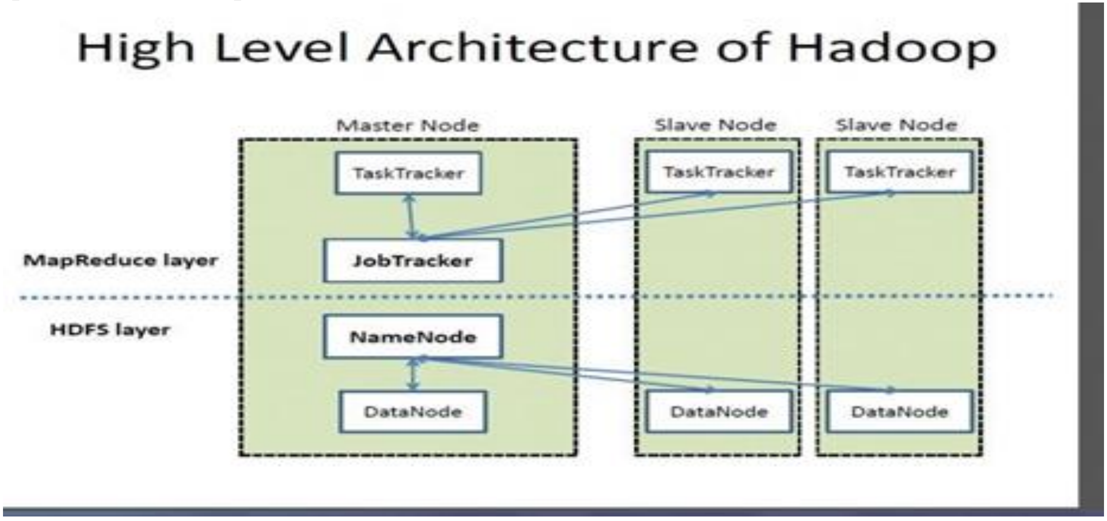

Fig3:High Level Architecture of Hadoop.

3.3 Role of Distributed Storage

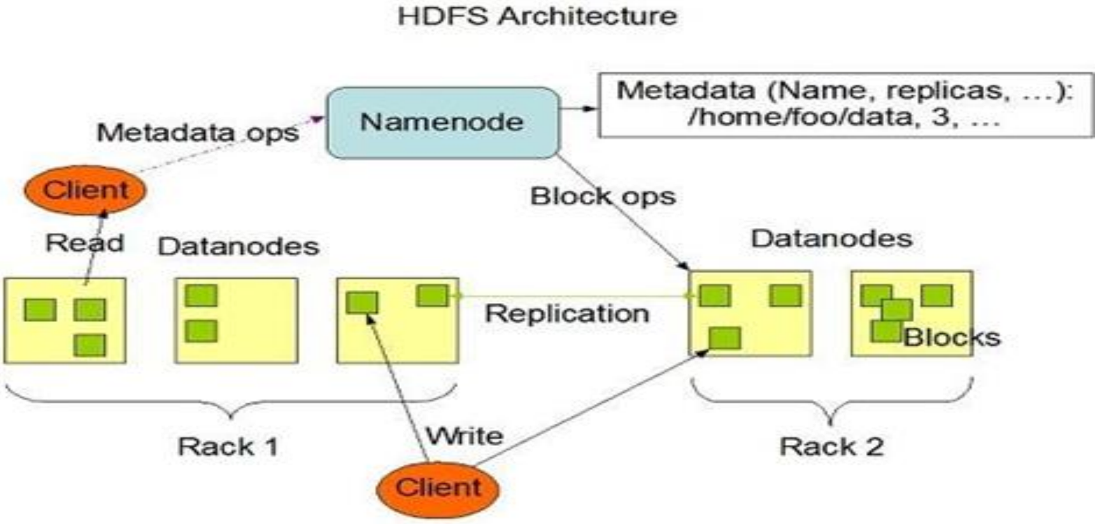

Fig 4:HDFS Distributed Storage

The hadoop cluster replicates a file on the HDFS into several bocks. A HDFS block is a 64 MB default storage block within the underlying file system. The block size can be increased to 256 MB according to different specifications. Program data and file system metadata are stored independently on the dedicated servers by the distributed file system (HDFS) [8]. The two key components of the Hadoop HDFS architecture are NameNode and DataNode. Application data shall be stored on servers known as DataNodes. Application data shall be stored on servers known as NameNode. In order to assure data reliability, HDFS replicates the file contents on several DataNodes based on the replication factor. HDFS must fulfill certain pre-requisites in order for the Hadoop architecture to be functional.

- A high performance is expected of all hard disks.
- High speed network for data transfer control and replication blocking.

3.4 NameNode

Inodes that contain various attributes, including permissions, modification timestamp, disk space quota, namespace quota and access times, are represented on NameNode on all HDFS files and directories. Memory NameNode maps the entire layout of the file system. The Inodes and the List of Blocks that describe the metadata include two images files and edits for persistence during the resetarts. The edits file contains changes that were made to the fsimage file contents. When NameNode is starting, fsimage file is loaded and the edits file contents are applied to retrieve the latest file system state [2].

If you don't restart the hadoop cluster together for months, you will have a huge downtime, as the file editions are increasing in size. In this case secondary NameNode is saved. Secondary NameNode receives fsimage and edits logs on a regular basis from primary NameNode, loading the fsimage and editing log files to the main memory by applying each file operation from log to fsimage editing. Secondary NameNode copies a new fsimage file to a NameNode primary and will also update the fsimage file's modified time to fstime file when it updates the fsimage file[3].

3.5 DataNote

A DatenNode can perform CPU-intense jobs. Job work, such as semantic and langage analytics, stats, machine learning tasks and intensive I / O tasks, including clusters, data imports , exports of data, searching, decompression and indexing. DataNode manages the status of the HDFS Node and interaction with blocks. For data processing and transfer[7] a DataNode requires a lot of I / O.

When every DataNode is started, it connects to NameNode and makes a handshake which checks the name space identificator and the DataNode software version. If one of them does not suit, the DataNode automatically shuts down. A DataNode checks block replicas by sending a block report to the NameNode. The first block report is submitted as soon as the DataNode registers. Each 3 seconds, DatoNode sends the DataNode heart beat to validate the functionality of the DataNode and the block replicas it hosts[8].

4. Results and Discussion

4.1 Working of Hadoop Mapreduce Architecture

Hadoop MapReduce is a software platform for applications that efficiently, fault-tolerantly process large quantities of data in parallel on large commodity hardware clusters[4]. The word Map Reduce refers to two separate tasks performed by Hadoop programs:

a) The Map task: it is the first task to take input data and to transform them into data sets where each element is split up into tuples.

b) The task reduction: The output of a map task is used as input and combines the data tuples in a smaller collection of tuples. After the map task, the reduction task is always done.

Input as well as output can be saved in a file system. The framework monitors, supervises and re-executes failing tasks for the planning tasks. A single master job tracker and one slave task tracker per cluster node[3] are included in the map reduction framework. The Master is responsible for the management of property, the monitoring of resources availability and the preparation, monitoring and reworking of the failed tasks of the job aspect on slaves. The Task Tracker Slaves perform the tasks as directed by the master and regularly provide the master with task details. For the Hadoop Map Reduction service, the job tracker is a single fault point, meaning that if the job tracker goes down, all work is stopped[6].

4.2 HADOOP EcoSystem Advancement

1. It gives easy access to the user to write and test the systems distributed quickly and distributes data and functions automatically over the machines and uses the primary CPU core parallelism.

2. In order to look for and handle the application layer errors, Hadoop library is developed.

3. Servers can be dynamically added or deleted at any point in time from the cluster.

The most popular and efficient big data platform provides the most reliable storage layer in the world: HDFS, the MapReduce batch processing motor and the YARN Resource Management Stack. Apache Hadoop offers the highest reliability for your storage system. Open Source – The open source software is Apache Hadoop. The data storage is distributed in HDFS throughout the cluster, data is processed simultaneously in the node clusters. The code can be modified to meet business requirements. By default all three block replicas are stored in Hadoop over the cluster and only changed if needed. In such cases, when any node falls the information on that node can be easily retrieved from other nodes, the fault tolerant of Hadoop can be examined. Data can be reliable, which is stored on the cluster despite machine failures, due to the replication of data in the cluster. There is often a hardware failure because of several copies of the data available and usable. Hadoop is very scalable and hardware can be used in a unique way

To the nodes be easily added. Hadoop is not very costly as it runs on the hardware cluster. No specialized machine is required for this. Hadoop offers enormous cost savings as more nodes are easy to add up. There are also growing nodes without any downtime and without a lot of pre-planning if the requirements increase.

5. CONCLUSION

This article deals with the Hadoop ecosystem and has researched its key components and Hadoop configuration. Specific factors such as HDFS and its architecture are discussed in data storage. The deployment phase of the Hadoop deployment is analyzed. HDFS maintains data consistency across the cluster, taking features such as preserving transaction logs into consideration. Another feature is the validation of an effective error detection technique, by which a transmitted message will be given numeric value based on number of bits. HDFS maintains replicated data block copies to avoid file duplication due to server failure. This paper also discusses MapReduce, an aggregation of various functions for sorting, storing and analyzing big data.

Future research includes the use of various technologies in large data sets to optimize and improve performance. The test results to be analyzed with different tools and experimental configurations.

References

- [1] Botta, A., de Donato, W., Persico, V., Pescapé, A., 2016. Integration of cloud computing and internet of things: a survey. Future Gener. Comput. Syst. 56, 684–700.
- [2] BaoRong Chang, Yo-Ai Wang, Yun-Da Lee, and Chien- FengHuang, "Development of Multiple Big Data Analysis Platforms for Business Intelligence", Proceedings of the 2017 IEEE International Conference on Applied System Innovation
- [3] Chu-Hsing Lin, Jung-Chun Liu, Tsung-Chi Peng, "Performance Evaluation of Cluster Algorithms for Big Data Analysis on Cloud", Proceedings of the 2017 IEEE International Conference on Applied System Innovation
- [4] Chen, M., Mao, S., Liu, Y., 2014a. Big data: a survey. Mobile Networks App. 19, 171–209.
- [5] Kankanhalli, A., Hahn, J., Tan, S., Gao, G., 2016. Big data and analytics in healthcare: introduction to the special section. Inf. Syst. Front. 18, 233–235.
- [6] Usama, M., Liu, M., Chen, M.: Job schedulers for big data processing in Hadoop environment: testing real-life schedule with benchmark programs. Digit. Commun. Networks. 3(4), 260–273 (2017)
- [7] Chang L, Wang Z, Ma T, Jian L, Ma L, Goldshuv A, Lonergan L, Cohen J, Welton C, Sherry G et al (2014) HAWQ: a massively parallel processing SQL engine in hadoop. In: Proceedings of the 2014 ACM SIGMOD
- [8] Ms.Preeti Narooka, Dr.Sunita Choudhary, "Optimization of the Search Graph Using Hadoop andLinux Operating System", 2017 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017) IEEE-ICASI 2017.
- [9] https://intellipaat.com/tutorial/hadooptutorial/introductio n- hadoop/
- [10] Apache Hadoop. http://hadoop.apache.org/