Scientific Journal of Impact Factor (SJIF): e-ISSN (O): 2348-4470

International Journal of Advance Engineering and Research Development

Volume 3, Issue 2, February -2016

A Study: Hadoop Framework

Devateja G[1], Kashyap P V B[2], Suraj C[3], Harshavardhan C[4], Impana Appaji[5]

[1234]*Computer Science & Engineering, Academy for Technical and Management Excellence college of Engineering, Mysore, Karnataka, India,*

[5]*Assistant Professor, Computer Science & Engineering, Academy for Technical and Management Excellence college of Engineering, Mysore, Karnataka, India,*

Abstract: *In recent years, the information that are retrieved from large datasets – known as Big Data. It's difficult to transfer larger files, For these reasons, we need to manipulate (e.g. edit, split, create) big data files to make them easier to move and work with them and even split big data files to make them more manageable. For this we use Apache hadoop software. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Which is based on distributed computing having HDFS file system. This file system is written in Java and designed for portability across various hardware and software platforms. Hadoop is very much suitable for storing high volume of data and it also provide the high speed access to the data of the application which we want to use. But hadoop is not really a database: It stores data and you can pull data out of it, but there are no queries involved - SQL or otherwise. Hadoop is more of a data warehousing system - so it needs a system like Map Reduce to actually process the data.*

KEYWORDS*: Big Data, Apache Hadoop, HDFS, Map Reduce, Distributed storage.*

I. INTRODUCTION

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

1.1 Hadoop Architecture:

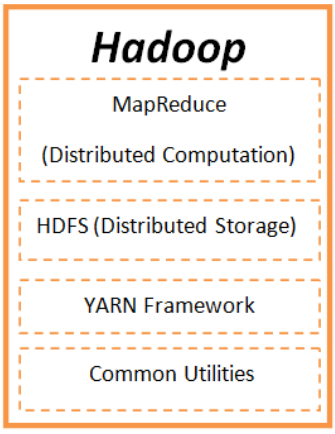

Fig.1.1 Hadoop Architecture.

Hadoop framework includes following four modules:

- **Hadoop Common:** These are Java libraries and utilities required by other hadoop modules. These libraries provide file system and OS level abstractions and contains the necessary Java files and scripts required to start hadoop.
- **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.
- **Hadoop Distributed File System (HDFS):** A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce:** This is YARN-based system for parallel processing of large data sets.

MapReduce: Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.The term MapReduce actually refers to the following two different tasks that hadoop programs perform:

- **Map Task:** This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples.
- **Reduce Task:** This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.

The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks.

The slaves TaskTracker execute the tasks as directed by the master and provide task-status information to the master periodically.

The JobTracker is a single point of failure for the Hadoop MapReduce service which means if JobTracker goes down, all running jobs are halted.

1.2 Hadoop Distributed File System:

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters of small computer machines in a reliable, fault-tolerant manner. HDFS uses this method when replicating data for data redundancy across multiple racks. This approach reduces the impact of a rack power outage or switch failure; if one of these hardware failures occurs, the data will remain available

HDFS uses a master/slave architecture where master consists of a single NameNode that manages the file system metadata and one or more slave DataNodes that store the actual data. A file in an HDFS namespace is split into several blocks and those blocks are stored in a set of DataNodes. The NameNode determines the mapping of blocks to the DataNodes. The DataNodes takes care of read and write operation with the file system. They also take care of block creation, deletion and replication based on instruction given by NameNode.

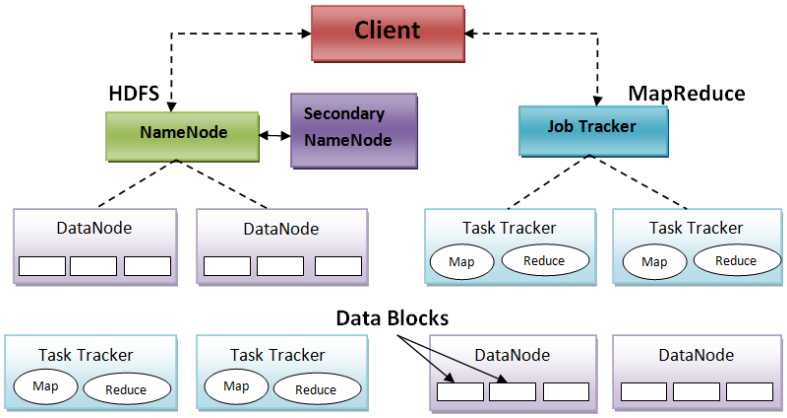

Fig.1.2 Working: Hadoop Distributed File System

Hadoop is supported by GNU/Linux platform and its flavors. Hadoop framework can be configure in following three modes: Standalone Mode, Pseudo-Distributed Mode and Fully-Distributed Mode. By default, hadoop is configured to run in a non-distributed mode, as a single Java process. This is useful for debugging. It can also be run on a single-node in a pseudo-distributed mode where each daemon runs in a separate Java process. Pseudo-distributed mode is also called as single node cluster.

Single node cluster:

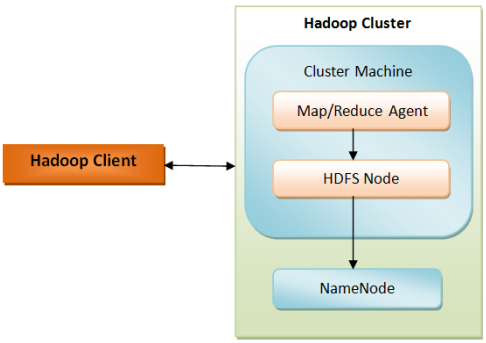

Fig.1.3 Single node cluster

A server cluster is a collection of servers, called nodes that communicate with each other to make a set of services highly available to clients.

Multi node cluster: A small Hadoop cluster includes a Master and its Secondary master. The master node consists of a JobTracker, TaskTracker, NameNode, and DataNode. A slaves acts as both a DataNode and TaskTracker, though it is possible to have data-only worker nodes and compute-only worker nodes

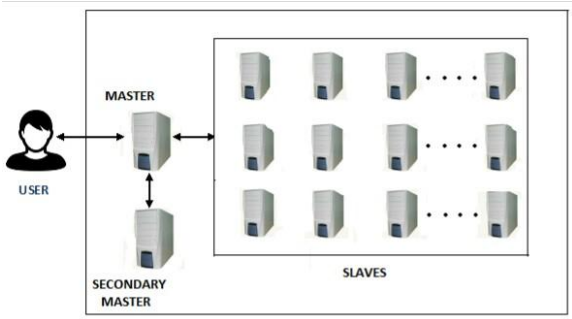

Fig.1.4 Multi node cluster

- **DataNode:** It stores data in the hadoop file system. A functional file system has more than one DataNode, with the data replicated across them.
- **NameNode:** It keeps the directory of all files in the file system, and tracks where across the cluster the file data is kept. It does not store the data of these file (Meta data of files).
- **Jobtracker:** It services within hadoop that forms out MapReduce to specific nodes in the cluster, ideally the nodes that have the data.
- **TaskTracker:** A node in the cluster that accepts tasks- MapReduce and Shuffle operations, from a Job Tracker.
- **Secondary Namenode:** Its purpose is to have a checkpoint in HDFS. It is just a helper node for NameNode.

II. ADVANTAGES

- **Scalable**

Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional database systems (RDBMS) that can't scale to process large amounts of data.

- **Cost effective**

Hadoop offers a cost effective storage solution for businesses exploding data sets. In an effort to reduce costs, many companies in the past would have had to down-sample data and classify it based on certain assumptions as to which data was the most valuable. Hadoop offers computing and storage capabilities for hundreds of pounds per terabyte.

- **Flexible**

Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, and market campaign analysis and fraud detection.

- **Fast**

Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing. If they dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes, and petabytes in hours.

- **Resilient to failure**

A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use. When it comes to handling large data sets in a safe and cost-effective manner, Hadoop has the advantage over relational database management systems, and its value for any size business will continue to increase as unstructured data continues to grow.

III. DISADVANTAGES

- **Security Concerns**

Hadoop is missing encryption at the storage and network levels, which is a major selling point for government agencies and others that prefer to keep their data under wraps.

- **Vulnerable By Nature**

Speaking of security, the very makeup of Hadoop makes running it a risky proposition. The framework is written almost entirely in Java. It has been heavily exploited by cybercriminals and as a result, implicated in numerous security breaches. For this reason, several experts have suggested dumping it in favor of safer, more efficient alternatives.

- **Not Fit for Small Data**

Due to its high capacity design, the Hadoop Distributed File System lacks the ability to efficiently support the random reading of small files. As a result, it is not recommended for organizations with small quantities of data.

IV. APPLICATIONS

The HDFS file system is not restricted to MapReduce jobs. It can be used for other applications, many of which are under development at Apache. The list includes the HBasedatabase, the Apache Mahout, Machine learning system, and the Apache Hive, Data Warehouse system. Hadoop can in theory be used for any sort of work that is batch-oriented rather than real-time, is very data-intensive, and benefits from parallel processing of data.

Commercial applications of Hadoop included:

- Marketing analytics
- Machine learning
- Image processing
- Processing of XML messages
- Web crawling

V. CONCLUSION

Everyday a large amount of data is getting dumped into machines. The major challenge is not to store large data sets in our systems but to retrieve and analyze the big data in the organizations that too data present in different machines at different locations. Hadoop offers a proven solution to the modern challenges facing legacy systems. Hadoop is an open-source software platform by the Apache Foundation for building clusters of servers for use in distributed computing. Hadoop can handle large volumes of structured and unstructured data more efficiently than the traditional enterprise data warehouse.

REFERENCES

[1] Apache Hadoop: https://hadoop.apache.org/
[2] Wikipedia: https://en.wikipedia.org/wiki/Apache_Hadoop
[3] https://opensource.com/life/14/8/intro-apache-hadoop-big-data
[4] https://radar.oreilly.com/2012/02/what-is-apache-hadoop.html
[5] https: //www .nobius.org/ dbg/practical-file-systemdesign.pdf
[6] https: //www.maxi-pedia.comiwhat+is+DFS.

