ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

Hadoop Architecture and Applications

Kusum Munde[1], Nusrat Jahan[2]

Assistant Professor, Department of Computer Engg, S.R.C.O.E, Savitribai Phule Pune University, Maharashtra, India[1]

Assistant Professor, Department of Computer Engg, S.R.C.O.E, Savitribai Phule Pune University, Maharashtra, India[2]

ABSTRACT: Hadoop has been specially designed to manage Big Data. It changes the way enterprises store, process, and analyze data. Hadoop provides scalable, reliable and flexible architecture. It allows organizations to store and analyze data at very high speed. Hadoop provides services for large data sets such as data processing, data access, data governance, security etc.

KEYWORDS: Hadoop Framework, Hadoop Architecture, Hadoop file system, Hadoop MapReduce Process, Hadoop Applications.

I. INTRODUCTION

Traditional approach works well where we have less volume of data that can be accommodated by standard database servers, or up to the limit of the processor. But when it has to manage huge amounts of data, it is really a challenge.

Doug Cutting, Mike Cafarella and team took the solution provided by Google and started an Open Source Project called HADOOP in 2005 and Doug named it after his son's toy elephant. Now Apache Hadoop is a registered trademark of the Apache Software Foundation [1].

Apache Hadoop is an open-source software framework. It is used for distributed storage and distributed processing of very large data sets on computer clusters. It uses commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common and should be automatically handled by the framework [2]. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop is a programming model which allows users to quickly write and test software in distributed systems.

II. HADOOP : FRAMEWORK

The base Apache Hadoop framework is composed of the following modules shown in Fig.1:

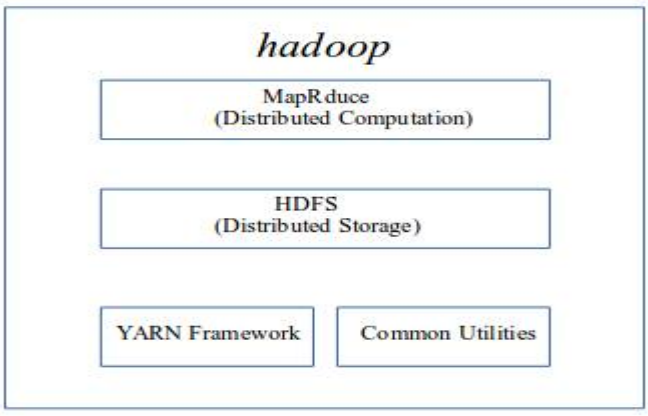

Fig.1 Hadoop Framework Components

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

A. **Hadoop Common**: It contains Java libraries and utilities needed. These libraries provide file system and OS level abstractions. It contains the necessary Java Archive files and scripts to start Hadoop.

B. **Hadoop Distributed File System (HDFS)**: It is a distributed file-system that stores data on commodity machines, providing very high bandwidth across the cluster. It provides high fault tolerance and native support of large data sets. It stores data on commodity hardware.

C. **Hadoop YARN**: This is a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users applications [3, 4].

D. **Hadoop MapReduce**: It is an implementation of the MapReduce programming model for large scale data processing. It is used for parallel processing of large datasets.

Hadoop is an open source project hosted by Apache Software Foundation. Hadoop consists of [5]:

Hadoop distributed file system-It is a File System.
Map Reduce-It is a Programming Paradigm. I t is introduced by Google for processing and generating large data sets on clusters of computers.

III. HADOOP ARCHITECTURE

The following Fig.2 shows the basic Hadoop Architecture.

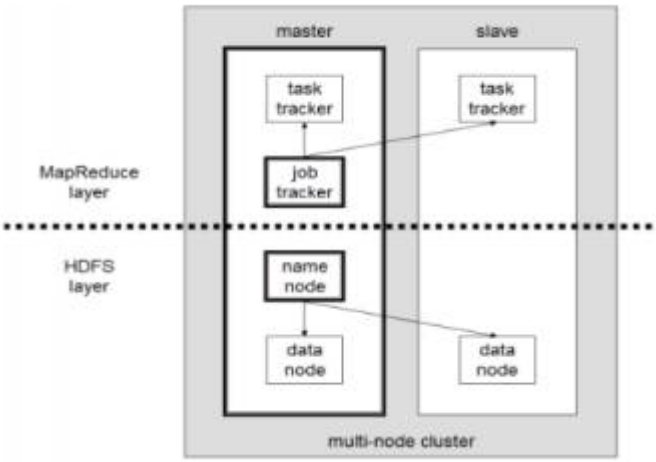

Fig2. Basic Architecture of Apache Hadoop

A small Hadoop cluster includes a single master node and multiple worker nodes. The master node consists of a JobTracker, TaskTracker, NameNode and DataNode. A slave or worker node acts as both a DataNode and TaskTracker, though it is possible to have data-only worker nodes and compute-only worker nodes. These are normally used only in nonstandard applications [6].

In a larger cluster, HDFS nodes are managed through a dedicated NameNode server to host the file system index, and a secondary NameNode. A JobTracker server can manage job scheduling across nodes. When Hadoop MapReduce is used with an alternate file system, the NameNode, secondary NameNode, and DataNode architecture of HDFS are replaced by the file-system-specific equivalents.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

IV. HADOOP: FILE SYSTEM

It is suitable for the distributed storage and processing. Hadoop provides a command interface to interact with HDFS. HDFS provides file permissions and authentication. An HDFS cluster is comprised of a NameNode, which manages the cluster metadata, and DataNodes that store the data.

Files and directories are represented on the NameNode by means of inodes. Inodes record attributes like permissions, access and modification times, namespace and disk space. The file is divided into large blocks (typically 128 megabytes), and each block of the file is independently replicated at multiple DataNodes. These blocks are stored on the local file system on the DataNodes.

The Namenode monitors the number of replicas of a block. When a replica of a block is lost due to a DataNode failure or disk failure, the NameNode creates another replica of the block. The NameNode maintains the namespace tree and the mapping of blocks to DataNodes, holding the entire namespace image in RAM.

The NameNode does not directly send requests to DataNodes. It sends instructions to the DataNodes by replying to heartbeats sent by those DataNodes. The instructions include commands to: replicate blocks to other nodes, remove local block replicas,re-register and send an immediate block report, or shut down the node [7]. Following Fig.3 shows the HDFS Architecture.

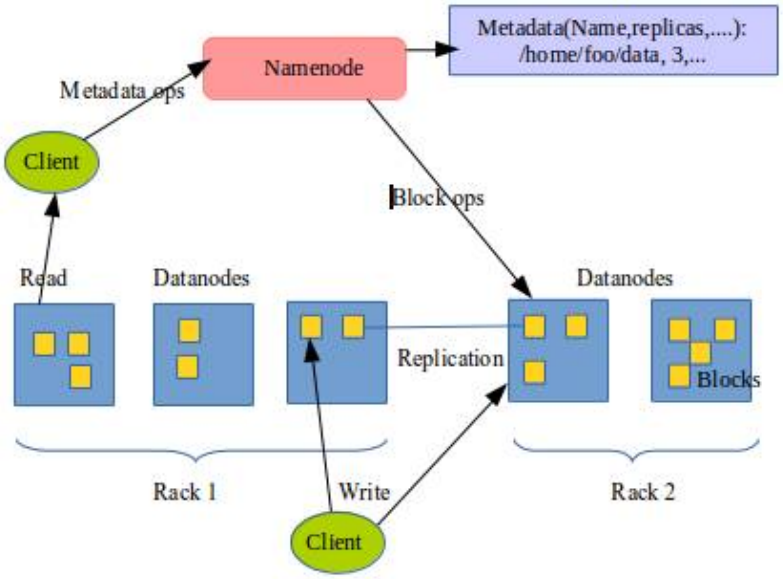

Fig.3 HDFS Architecture

V. HADOOP: MAP REDUCE PROCESS

The Map function divides the input into ranges by the InputFormat and creates a map task for each range in the input. The JobTracker distributes those tasks to the worker nodes. The output of each map task is partitioned into a group of key-value pairs for each reduce.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

The Reduce function then collects the various results and combines them .Each reduce pulls the relevant partition from the machines where the maps executed, then writes its output back into HDFS. Thus, the reduce is able to collect the data from all of the maps for the keys and combine them to solve the problem [8].

Following Fig.4 shows the Hadoop Map Reduce Process.

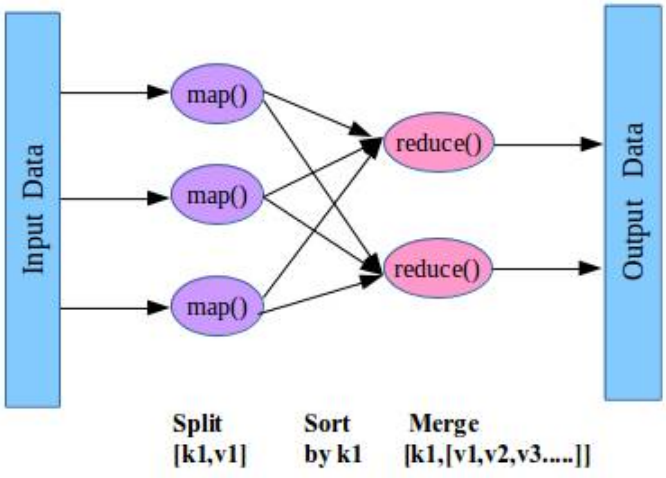

Fig. 4 Hadoop Map Reduce Process

The MapReduce framework operates on <key, value> pairs. The framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job.

VI. HADOOP: BENEFITS

There are number of reasons, why organizations use Hadoop. Hadoop has ability to store, manage and analyze structured and unstructured at low-cost.

- **Scalability and Performance** –Its distributed processing of data local to each node in a cluster enables Hadoop to store, manage, process and analyze data at petabyte scale. Thus, Hadoop is built to process large amount of data from terabytes to petabytes and beyond it.

- **Reliability** – Hadoop is fundamentally resilient – when a node fails processing is re-directed to the remaining nodes in the cluster and data is automatically re-replicated in preparation for future node failures.

- **Flexibility** –You can store data in any format, including semi-structured or unstructured formats, and then parse and apply schema to the data when read [8].

- **Low Cost** – Hadoop is open source and runs on low-cost commodity hardware. It is cost effective.

- **Fault-tolerant**. Fault tolerance is one of feature for using Hadoop. Even if individual nodes experience high rates of failure when running jobs on a large cluster, data is replicated across a cluster so that it can be recovered easily in the face of disk, node or rack failures [9].

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

VII. HADOOP: APPLICATIONS

1. Amazon S3 is a simple data storage service. You are billed every month for storage and data transfer. Transfer in between S3 and AmazonEC2 is free of cost. This makes use of S3 useful for Hadoop users who run clusters on EC2.

2. Yahoo has launched the world's largest Apache Hadoop production application recently. The Yahoo Search. Webmap is a Apache Hadoop application that runs on a more than about 10,000 core Linux cluster and introduces data that is now used in every Yahoo query.

3. Facebook's engineering team has given some details on the tools it is using to analyze the large data sets it collects. Hadoop that makes it easy to analyze large amounts of data [10].

4. In addition to above hadoop is also used in various areas such as marketing analytics, machine learning, image processing, processing of XML messages, web crawling and/or text processing [11].

5. Yahoo is using Hadoop for content optimization, search index, ads optimization and content feed processing. Biggest development has happened in the media industry. New York Times uses Hadoop to make PDF file from published articles. E-Commerce companies are using Hadoop to track user behaviour. Hadoop is also used in customer segmentation and experience analysis, credit risk assessment, targeted services, etc. Hadoop addresses most common industry challenges like fraud, financial crimes and data breaches effectively [12].

VIII. CONCLUSION

Hadoop is a big data processing platform. Hadoop stores unstructured data such as images, sensor data, Media files, log files, data from the internet and so on. Hadoop supports distributed storage and distributed computing .It uses Hadoop distributed file system (HDFS) for storage and distributed computing through a programming model called MapReduce by processing query in parallel. Hadoop is cost effective, scalable. As the data is stored in HDFS which uses block replicas, hadoop is also fault tolerant.

REFERENCES

[1]Vance, Ashlee (2009-03-17). "*Hadoop, a Free Software Program, Finds Uses Beyond Search*". The New York Times. Archived from the original on August 30, 2011. Retrieved 2010-01-20.
[2]"*Welcome to Apache Hadoop*!". hadoop.apache.org. Retrieved 2016-08-25.
[3]Resource (Apache Hadoop Main 2.5.1 API)". apache.org. Apache Software Foundation. 2014-09-12. Retrieved 2014-09-30.
[4]Murthy, Arun (2012-08-15). "*Apache Hadoop YARN – Concepts and Applications*". hortonworks.com. Hortonworks. Retrieved 2014-09-30.
[5]Jaimesh J Patel Harshal Shah,"*A Survey: Data Retrieval in Hadoop MapReduce*",International Journal of Advanced Research in Computer Science and Software Engineering,Volume 5, Issue 4, 2015 ISSN: 2277 128X.
[6]Twinkle Antony, Shaiju Paul,"*Addressing Big Data With Hadoop*",IJCSMC, Vol. 3, Issue. 2, February 2014, pg.459 – 462,ISSN 2320–088X.
[7]"*Hadoop Distributed File System (HDFS)*", http://hortonworks.com/hadoop/hdfs/.
[8]"*MapReduce*", http://hortonworks.com/apache/mapreduce/ .
[9]Apache Hadoop | Big Data Software - Map,Rhttps://www.mapr.com/products/apache-hadoop.
[10]Ruchira A. Kulkarni, Manisha K. Damade, Shahista Bano,"*Hadoop Architecture: A distributed file system*",International Journal of Innovative and Emerging Research in Engineering,Volume 2, Issue 3, 2015,e-ISSN: 2394 – 3343,p-ISSN: 2394 – 5494.
[11]How 30+ enterprises are using Hadoop", in DBMS2". Dbms2.com. 10 October 2009. Retrieved 2013-10-17.
[12]"Real life example how Big Data & Hadoop is changing the life for a..", http://www.manipalprolearn.com/.