Volume 5, Issue 4, 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Hadoop Technology and its Role in Information Technology

Anusha B. Dhakite[1] , Prof. Sameer Y. Thakur[2]
[1]Research Scholar, [2]Assistant Professor,
[1, 2,] P.G. Department of Computer Science and Technology DCPE, (Autonomous college),
HVPM, Amravati, Maharashtra, India

Abstract- Hadoop is the popular open source implementation of Map Reduce, a powerful tool designed for deep analysis and transformation of very large data sets. Hadoop enables you to explore complex data, using custom analyses tailored to your information and questions. Hadoop is the system that allows unstructured data to be distributed across hundreds or thousands of machines forming shared nothing clusters and the execution of Map/Reduce routines to run on the data in that cluster. Hadoop has its own file system which replicates data to multiple nodes to ensure if one node holding data goes down, there are at least 2 other nodes from which to retrieve that piece of information. This protects the data availability from node failure, something which is critical when there are many nodes in a cluster. Hadoop has its origins in Apache Nutch, an open source web search engine, itself a part of the Lucene project. Building a web search engine from scratch was an ambitious goal, for not only is the software required to crawl and index websites complex to write, but it is also a challenge to run without a dedicated operations team, since there are so many moving parts. It's expensive too: Mike Cafarella and Doug Cutting estimated a system supporting a 1-billion-page index would cost around half a million dollars in hardware, with a monthly running cost of $30,000.

Keywords-Namenode,HDFS,Map Reduce,Petabyte,Datanode

I. INTRODUCTION

Apache Hadoop is an open source framework for developing distributed applications that can process very large amounts of data. It is a platform that provides both distributed storage and computational capabilities. Hadoop has two main layers:

A.Computation layer: The computation tier uses a framework called ***Map Reduce****.*

B.Distributed storage layer: A distributed file system called ***HDFS*** *provides storage.*

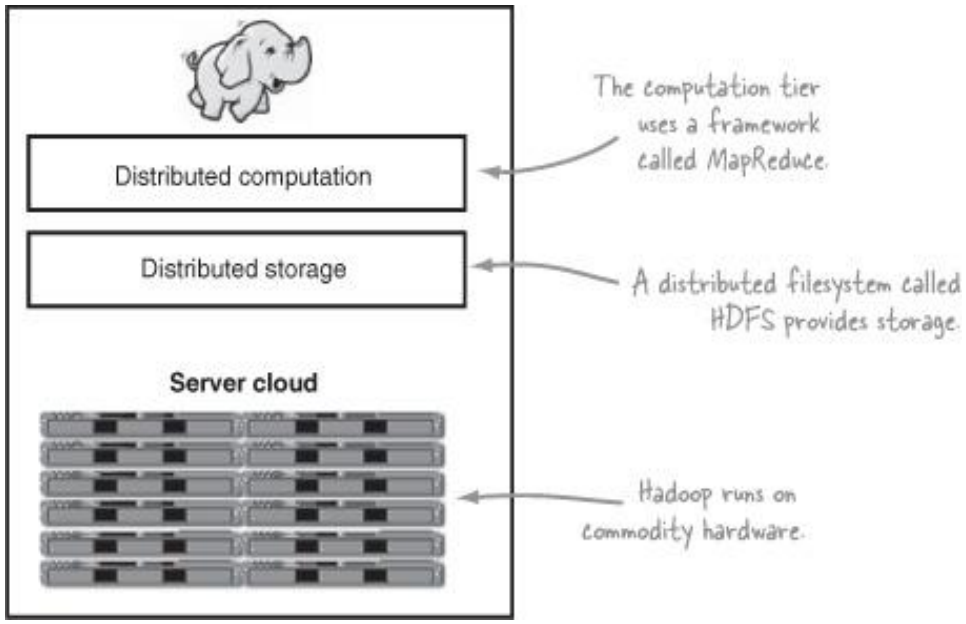
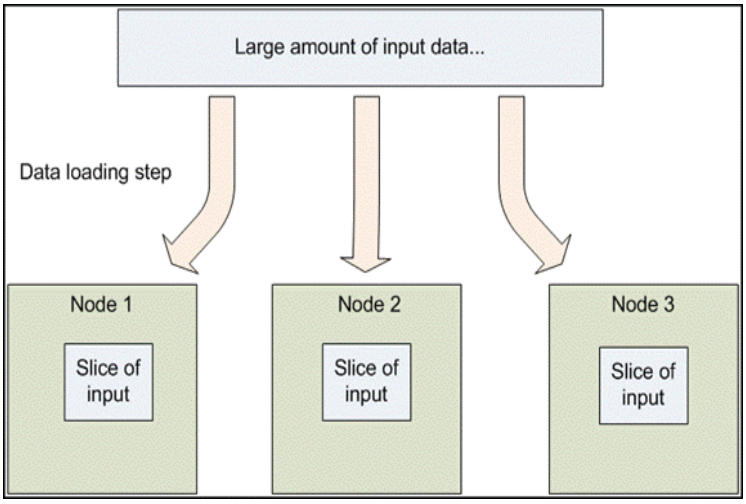


Fig1:Layers of Hadoop

In a Hadoop cluster, data is distributed to all the nodes of the cluster as it is being loaded in. The Hadoop Distributed File System (HDFS) will split large data files into chunks which are managed by different nodes in the cluster. In addition to this each chunk is replicated across several machines, so that a single machine failure does not result in any data being unavailable. An active monitoring system then re-replicates the data in response to system failures which can result in partial storage. Even though the file chunks are replicated and distributed across several machines, they form a single namespace, so their contents are universally accessible. Data is conceptually **record-oriented** in the Hadoop programming framework. Individual input files are broken into lines or into other formats specific to the application logic. Each process running on a node in the cluster then processes a subset of these records. The Hadoop framework then schedules these processes in proximity to the location of data/records using knowledge from the distributed file system. Since files are spread across the distributed file system as chunks, each compute process running on a node operates on a subset of the data.



Which data operated on by a node is chosen based on its locality to the node: most data is read from the local disk straight into the CPU, alleviating strain on network bandwidth and preventing unnecessary network transfers. This strategy of **moving computation to the data,** instead of moving the data to the computation allows Hadoop to achieve high data locality which in turn results in high performance. Hadoop is written in the Java programming language and is an Apache top-level project being built and used by a global community of contributors. Hadoop and its related projects (Hive, HBase, Zookeeper, and so on) have many contributors from across the ecosystem. Though Java code is most common, any programming language can be used with "streaming" to implement the "map" and "reduce" parts of the system.It provides a distributed file system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both map/reduce and the distributed file systems are designed so that node failures are automatically handled by the framework. It enables applications to work with thousands of computation-independent computers and pet bytes of data. The entire Apache Hadoop "platform" is now commonly considered to consist of the Hadoop kernel, Map Reduce and Hadoop Distributed File System (HDFS), as well as a number of related projects – including Apache Hive, Apache HBase, and others.

II. WHY HADOOP?

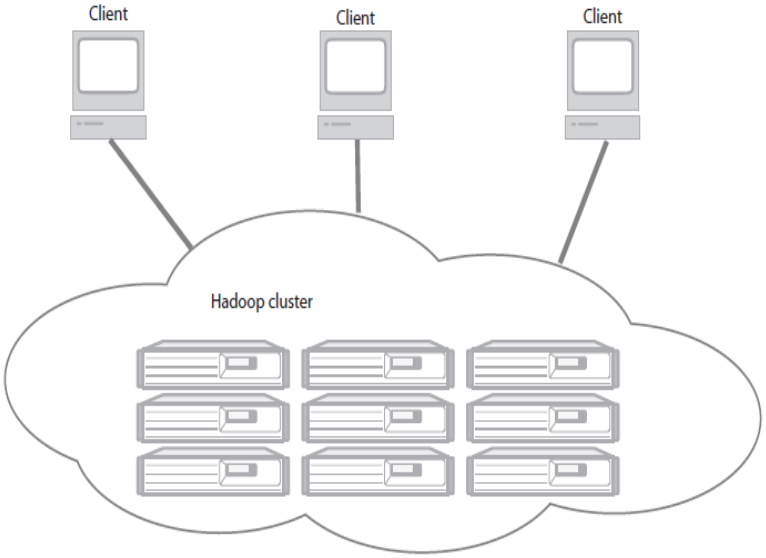
Building bigger and bigger servers is no longer necessarily the best solution to large-scale problems. Nowadays the popular approach is to tie together many low-end machines together as a single functional distributed system.
For example,
i.A high end machine with four I/O channels each having a throughput of 100 MB/sec will require three hours to read a 4 TB data set! With Hadoop, this same data set will be divided into smaller (typically 64 MB) blocks that are spread among many machines in the cluster via the Hadoop Distributed File System (HDFS).
ii. With a modest degree of replication, the cluster machines can read the data set in parallel and provide a much higher throughput. Moreover its cheaper than one high-end server!

A. For Computationally intensive work,

Most of the distributed systems (e.g. SETI@home) are having approach of moving the data to the place where computation will take place and after the computation, the resulting data is moved back for storage. This approach works fine for computationally intensive work.

B. For data-intensive work,

We need other better approach; Hadoop has better philosophy toward that because **Hadoop focuses on moving code/algorithm to data instead data to the code/algorithm**. The move-code-to-data philosophy applies within the Hadoop cluster itself, and data is broken up and distributed across the cluster, and computation on a piece of data takes place on the same machine where that piece of data resides. **Hadoop philosophy of move-code-to-data makes more sense** as we know the code/algorithm are always smaller than the Data hence code/algorithm is easier to move around.



A Hadoop cluster has many parallel machines that store and process large data

IV. HADOOP'S COMPONENTS

A. Hadoop Distributed File System:

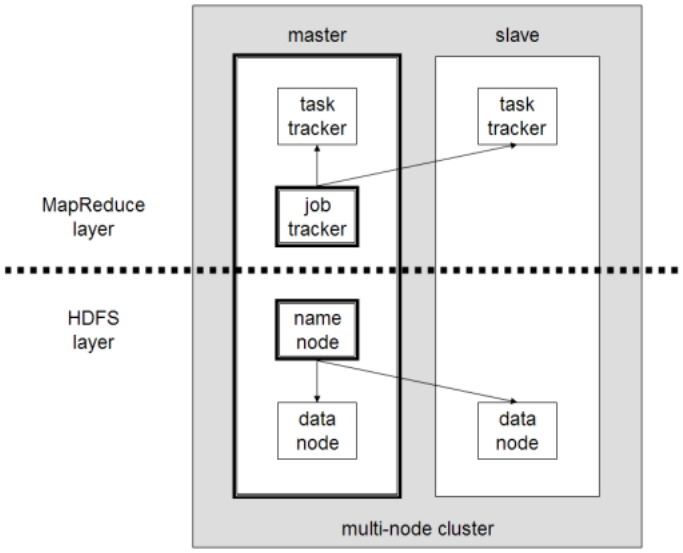
HDFS, the storage layer of Hadoop, is a distributed, scalable, Java-based file system adept at storing large volumes of unstructured data. Map Reduce is a software framework that serves as the compute layer of Hadoop. Map Reduce jobs are divided into two (obviously named) parts. The "Map" function divides a query into multiple parts and processes data at the node level. The "Reduce" function aggregates the results of the "Map" function to determine the"answer" to the query. Hive is a Hadoop-based data warehousing-like framework originally developed by Facebook. It allows users to write queries in a SQL-like language called HiveQL, which are then converted to Map Reduce. This allows SQL programmers with no Map Reduce experience to use the warehouse and makes it easier to integrate with business intelligence and visualization tools such as Microstrategy, Tableau, Revolutions Analytics, etc.Pig Latin is a Hadoop-based language developed by Yahoo. It is relatively easy to learn and is adept at very deep, very long data pipelines (a limitation of SQL.)

HBase is a non-relational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes. EBay and Facebook use HBase heavily. Flume is a framework for populating Hadoop with data. Oozie is a workflow processing system that lets users define a series of jobs written in multiple languages – such as Map Reduce, Pig and Hive -- then intelligently link them to one another. Oozie allows users to specify, for example, that a particular query is only to be initiated after specified previous jobs on which it relies for data are completed.Flume is a framework for populating Hadoop with data. Ambari is a web-based set of tools for deploying, administering and monitoring Apache Hadoop clusters. Its development is being led by engineers from Hortonworoks, which include Ambari in its Hortonworks Data Platform. Avro is a data serialization system that allows for encoding the schema of Hadoop files. It is adept at parsing data and performing removed procedure calls. Mahout is a data mining library. It takes the most popular data mining algorithms for performing clustering, regression testing and statistical modeling and implements them using the Map Reduce model. Sqoop is a connectivity tool for moving data from non-Hadoop data stores – such as relational databases and data warehouses – into Hadoop. HCatalog is a centralized metadata management and sharing service for Apache Hadoop. BigTop is an effort to create a more formal process or framework for packaging and interoperability testing of Hadoop's sub-projects and related components with the goal improving the Hadoop platform as a whole.

V. WORKING OF HADOOP ARCHITECTURE

Hadoop is designed to run on a large number of machines that don't share any memory or disks. That means you can buy a whole bunch of commodity servers, slap them in a rack, and run the Hadoop software on each one. When you want to load all of your organization's data into Hadoop, what the software does is bust that data into pieces that it then spreads across your different servers. There's no one place where you go to talk to all of your data; Hadoop keeps track of where the data resides. And because there are multiple copy stores, data stored on a server that goes offline or dies can be automatically replicated from a known good copy.

In a centralized database system, you've got one big disk connected to four or eight or 16 big processors. But that is as much horsepower as you can bring to bear. In a Hadoop cluster, every one of those servers has two or four or eight CPUs. You can run your indexing job by sending your code to each of the dozens of servers in your cluster, and each server operates on its own little piece of the data. Results are then delivered back to you in a unified whole. That's Map Reduce you map the operation out to all of those servers and then you reduce the results back into a single result set. Architecturally, the reason you're able to deal with lots of data is because Hadoop spreads it out.



And the reason you're able to ask complicated computational questions is because you've got all of these processors, working in parallel, harnessed together.Hadoop implements a computational paradigm named Map/Reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system (HDFS) that stores data on the compute nodes, providing very high aggregate bandwidth across the that node failures are automatically handled by the framework. Hadoop Common is a set of utilities that support the other Hadoop subprojects. Hadoop Common includes File System, RPC, and serialization libraries.

VI. HADOOP KEY FEATURES

Distributed computing is the very vast field but following key features has made Hadoop very distinctive and attractive.

A. Accessible:

Hadoop runs on large clusters of commodity machines or on cloud computing services such as Amazon's Elastic Compute Cloud (EC2).

B. Robust:

As Hadoop is intended to run on commodity hardware, It is architected with the assumption of frequent hardware malfunctions. It can gracefully handle most such failures.

C. Scalable:

Hadoop scales linearly to handle larger data by adding more nodes to the cluster.

D. Simple:

Hadoop allows users to quickly write efficient parallel code. Hadoop's accessibility and simplicity give it an edge over writing and running large distributed programs.

VII. HADOOP ADVANTAGES

i.Hadoop is an open source, versatile tool that provides the power of distributed computing.
ii.By using distributed storage & **transferring code instead of data**, Hadoop avoids the costly transmission step when working with large data sets.
iii.Redundancy of data allows Hadoop to recover from single node fail.
iv.Ease to create programs with Hadoop As it uses the Map Reduce framework.
v.You did not have to do worry about partitioning the data, determining which nodes will perform which tasks, or handling communication between nodes as It is all done by Hadoop for you.
vi.Hadoop leaving you free to focus on what is most important to you and your data and what you want to do with it.

VII. APPLICATIONS

1.A9.com – Amazon:

To build Amazon's product search indices; process millions of sessions daily for analytics, using both the Java and streaming APIs; clusters vary from 1 to 100 nodes.

2.Yahoo! :

More than 100,000 CPUs in ~20,000 computers running Hadoop; biggest cluster: 2000 nodes (2*4cpu boxes with 4TB disk each); used to support research for Ad Systems and Web Search

3.AOL :

Used for a variety of things ranging from statistics generation to running advanced algorithms for doing behavioral analysis and targeting; cluster size is 50 machines, Intel Xeon, dual processors, dual core, each with 16GB Ram and 800 GB hard-disk giving us a total of 37 TB HDFS capacity.

4.Facebook:

To store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning; 320 machine cluster with 2,560 cores and about 1.3 PB raw storage;

5.FOX Interactive Media :

3 X 20 machine cluster (8 cores/machine, 2TB/machine storage) ; 10 machine cluster (8 cores/machine, 1TB/machine storage); Used for log analysis, data mining and machine learning

6. Cornell University Web Lab:

To generate web graphs on 100 nodes (dual 2.4GHz Xeon Processor, 2 GB RAM, 72GB Hard Drive)

7.NetSeer:

Up to 1000 instances on Amazon EC2; Data storage in Amazon S3; Used for crawling, processing, serving and log analysis.

8.The New York Times :

Large scale image conversions; EC2 to run hadoop on a large virtual cluster

as many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is part of the Apache Hadoop Core project.

REFERENCES

[1] Apache Hadoop! (hadoop.apache.org)
[2] Hadoop on Wikipedia(http://en.wikipedia.org/wiki/Hadoop)
[3] www.guruzon.com/6/introduction/ Hadoop
[4] Free Search by Doug Cutting (http://cutting.wordpress.com)
[5] www.pentaho.com
[6] Cloudera - Apache Hadoop for the Enterprise (http://www.cloudera.com)
[7] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler Yahoo! Sunnyvale, California USA ©2010 IEEE

