International Research Journal of Multidisciplinary Science & Technology
www.irjmrs.com ISSN : 2455-930X

Hadoop: - An Overview

Shikha Mehta[1]
[1]*Librarian, Aricent Technologies Holding Ltd, Gurgaon*
E-Mail: shikhafeb@gmail.com

----------***----------

Abstract - *"In pioneer days they oxen for heavy pulling, and when one ox couldn't budge a log, they didn't to grow a larger ox. We shouldn't be trying for bigger computers but for more systems of computers"-Grace Hopper*

During the past several years there has been an increasing within business and academic circles that certain nations have evolved into information societies. These countries now rely heavily on knowledge and information to spur economic growth. Many leaders today believe the generation of Information has provided the foundation needed to increase the efficiency and productivity of a society. The computer has been the catalyst for this information revolution.

To gain an appreciation of the statements of business and academic leaders, one need only examine selected segments of society to understand what the computer has done to provide this vital information.

The nature and complexity of software have changed significantly in the last 30 years. In the 1970's applications ran on a single processor, produced alphanumeric output, and received their input from a linear source. Today's applications are far more complex; typically have graphical user interface and client-server architecture. They frequently run on two or more processors, under different operating systems and on geographically distributed machines.

Key Words: *Hadoop, HDFS, HDFS Cluster, Yarn, Map Reduc.*

1.INTRODUCTION

Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware. Hadoop is an Apache top-level project being built and used by a global community of contributors and users. It is licensed under the Apache License 2.0.

Hadoop was created by Doug Cutting and Mike Cafarella.It was originally developed to support distribution for the Nutch search engine project. Doug, who was working at Yahoo! at the time and is now Chief Architect of Cloudera,Hadoop is the most popular platform related to Big Data. Compared to relational databases.

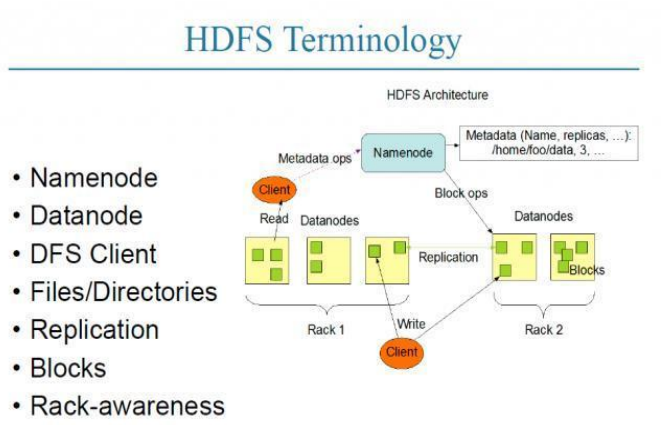
Pairing Apache Hadoop distributed file storage with hardware based trusted computing mechanisms has the potential to reduce the risk of data compromise. With the growing use of Hadoop to tackle big data analytics involving sensitive data, a Hadoop cluster could be a target for data exhilaration, corruption, or modification. By implementing open standards based Trusted Computing technology at the infrastructure and application levels; a novel and robust security posture and protection is presented.

2.The Apache Hadoop framework is composed of the following modules

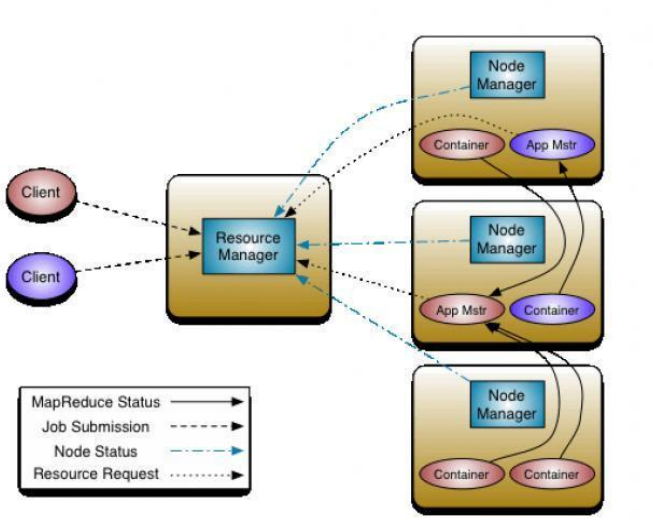
2.1.Hadoop Common: contains libraries and utilities needed by other Hadoop modules

2.2.Hadoop distributed file system

Hadoop Distributed File System (HDFS): A distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster. TheHadoop distributed file system (HDFS) is a distributed, scalable, and portable file-system written in Java for the Hadoop framework. Each node in a Hadoop instance typically has a single name node, and a cluster of data nodes form the HDFS cluster. The situation is typical because each node does not require a data node to be present. Each data node serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses the TCP/IP layer for communication. Clients use Remote procedure call (RPC) to communicate between each other. HDFS stores large files (typically in the range of gigabytes to terabytes) across multiple machines. It achieves reliability by replicating the data across multiple hosts, and hence does not require RAID storage on hosts. With the default replication value, 3, data is stored on three nodes: two on the same rack, and one on a different rack. Data nodes can talk to each other to rebalance data, to move copies around, and to keep the replication of data high. HDFS is not fully POSIX-compliant, because the requirements for a POSIX file-system differ from the target goals for a Hadoop application. The tradeoff of not having a fully POSIX-compliant file-system is increased performance for data throughput and support for non-POSIX operations such as Append.



2.3.Hadoop YARN: a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications



3.What YARN does

YARN enhances the power of a Hadoop compute cluster in the following ways:

a) Scalability: The processing power in data centers continues to grow quickly. Because YARN Resource Manager focuses exclusively on scheduling, it can manage those larger clusters much more easily.

b) Compatibility with Map Reduce: Existing Map Reduce applications and users can run on top of YARN without disruption to their existing processes.

c) Improved cluster utilization: The Resource Manager is a pure scheduler that optimizes cluster utilization according to criteria such as capacity guarantees, fairness, and SLAs. Also, unlike before, there are no named map and reduce slots, which helps to better utilize cluster resources.

d) Support for workloads other than Map Reduce: Additional programming models such as graph processing and iterative modeling are now possible for data processing. These added models allow enterprises to realize near real-time processing and increased ROI on their Hadoop investments.

e) Agility: With Map Reduce becoming a user-land library, it can evolve independently of the underlying resource manager layer and in a much more agile manner.

4.How YARN works

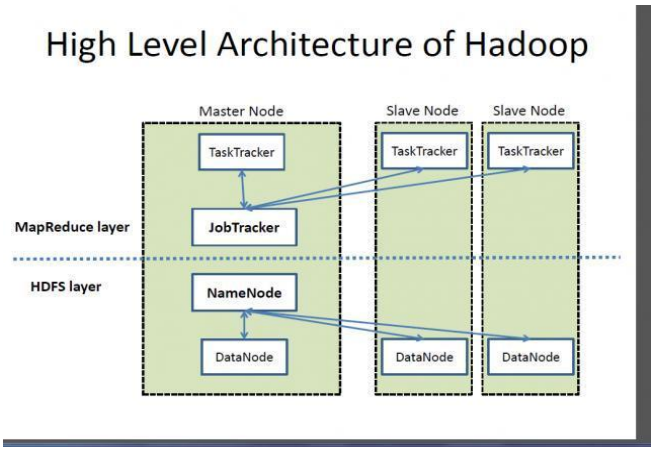
The fundamental idea of YARN is to split up the two major responsibilities of the Job Tracker/Task Tracker into separate entities:

1. a global Resource Manager
2. a per-application Application Master
3. a per-node slave Node Manager and
4. a per-application container running on a Node Manager

The Resource Manager and the Node Manager form the new, and generic, system for managing applications in a distributed manner. The Resource Manager is the ultimate authority that arbitrates resources among all the applications in the system. The per-application Application Master is a framework-specific entity and is tasked with negotiating resources from the Resource Manager and working with the Node Manager(s) to execute and monitor the component tasks. The Resource Manager has a scheduler, which is responsible for allocating resources to the various running applications, according to constraints such as queue capacities, user-limits etc. The scheduler performs its scheduling function based on the resource requirements of the applications. The Node Manager is the per-machine slave, which is responsible for launching the applications' containers, monitoring their resource usage (CPU, memory, disk, network) and reporting the same to the Resource Manager. Each Application Master has the responsibility of negotiating appropriate resource containers from the scheduler, tracking their status, and monitoring their progress. From the system perspective, the Application Master runs as a normal container.

5.Hadoop Map Reduce:

A programming model for large scale data processing. Hadoop Map Reduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. AMap Reduce *job* usually splits the input data-set into independent chunks which are processed by the *map tasks* in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the *reduce tasks*. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks. Typically the compute nodes and the storage nodes are the same, that is, the Map Reduce framework and the Hadoop Distributed File System are running on the same set of nodes. This configuration allows the framework to effectively schedule tasks on the nodes where data is already present, resulting in very high aggregate bandwidth across the cluster.



6. Use of the Hadoop in the field of Library and Information Science:

Hadoop is well known in now a days for big data for very long without running into small parts. It's proven to be very helpful in storing and managing vast amounts of data cheaply and efficiently.

As in Libraries we need back-up and in big libraries we have large-scale computers and staff so we need it exactly. Because, it's a way sorting enormous data sets across distributed clusters of servers and then running distributed cluster of server and then running distributed analysis application in each cluster.

For a big university libraries we have lots of data of books, journals and we have to lots no of staff andeveryone doing his/her work in own way. So they added all data in different format in different computers. But the problem is all computers had efficient balance to accept all Data? In that case we can user hadoop, because big data application will continue to run even individual servers or clusters- fail and its designed to be efficient, because itsdoesn't require our application to settle huge volumes of data across our network.

Hadoop software library is a framework that allows for the distributed processing of large data sets across cluster of computer using simple programming models. Its designed to scale up from single servers to thousands of machine, each offering local computation and storage.rathen then rely on hardware to deliver high availability, the library itself is designed to detect and handle failure at the application layers. So delivering a highly available service on top of cluster of computers, each of which may be prone to failure.

7.CONCLUSION

The lure of using big data for our Libraries is a strong one, and there is no brighter lure these days than Apache Hadoop, the scalable data storage platform that lies at the heart of many big data solutions. But as attractive as Hadoop is, there is still a steep learning curve involved in understanding what role Hadoop can play for an organization, and how best to deploy it.

If you remember nothing else about Hadoop, keep this in mind: It has two main parts - a data processing framework and a distributed file system for data storage. There's more to it than that, of course, but those two components really make things go.

Perhaps it's also a "**right place at the right time**" situation...
Hadoop's development coincided with the onset of weblog-generated data that was high in volume and unstructured, to boot. Add to that that Hadoop was open source and relatively inexpensive, and it seems a good mix for success.

We use Hadoop to data mine nearly 300 million bibliographic records for library items and we can attest that it has quickly become essential to much that we do in OCLC Research. We located a unique string in 129 records out of 300 million in fewer than 5 minutes using a simple string search. Just try doing that outside of Hadoop.

REFERENCES

[1] White, Tom(2012)ed.3 "Hadoop: the definitive guide:O'reilly publisher, pp. 270-480, 2014
[2] https://opensource.com/life/14/8/intro-apache-hadoop-big-data-[Online]. Available-Communications-Economy.pdf.
[3] Bar-Ilan, J., "Data collection methods on the Web for informetrics purposes: a review and analysis". Scientometrics .50.1 (2001).
[4] Ingwersen, P. "The calculation of Web impact factors." Journal of Documentation. 54.2 (1998).
[5] Aggarwal K.K, and Singh, Yogesh. (2008) ed.3, Software Engineering, New Age International Publisher. P.238-269.