Hadoop Framework

Chetna Kachhwaha
Computer Application, Jodhpur National University, Jodhpur, India
chetna_1978@yahoo.com

***Abstract*—Hadoop is an open source software framework that provides support system for processing and storing extremely large data sets in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. It is Java-based programming framework for distributed storage and distributed processing of very large data sets on computer clusters built from hardware. The most important point taken into account by Hadoop is that hardware failures are common and should be automatically handled by the framework.**

Keywords - Hadoop, HDFS,MapReduce,Cluster

I. INTRODUCTION

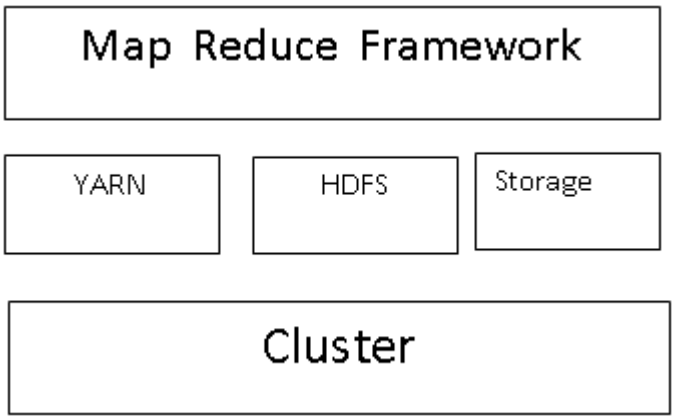
HADOOP makes it possible to run applications on systems with thousands of commodity hardware nodes, and handle thousands of terabytes of data. It has distributed file system which facilitates transferring of data rapidly among the nodes and allows the system to operate continuously in case of a node failure also. Thus reducing the risk of catastrophic system failure and unexpected data loss, even when significant number of nodes become inoperative.

II. HISTORY

Hadoop was created by computer scientists Doug Cutting and Mike in January 2006 to support distribution for the Nutch search engine. This was inspired by Google's MapReduce which is a software framework in which an application is broken down into numerous small parts. Any of these parts, which are also called fragments or blocks, can run on any of the node present in the cluster. After years of development within the open source community, Hadoop 1.0 became publically available in November 2012 as part of the Apache project sponsored by the Apache Software Foundation. It provides a reliable and efficient way of analyzing the data both in structured and unstructured patterns with the capability of handling large amount of data sets.

III. HADOOP ARCHITECTURE

Hadoop framework can mainly be classified in five building blocks. The runtime environment of Hadoop is shown in figure

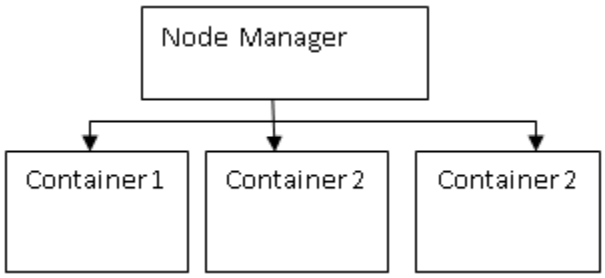


A. Cluster
This is the set of host machines (nodes).These Nodes can further partitioned into racks. This is the hardware part of the infrastructure.

B. YARN Infrastructure
YARN stands for Yet Another Resource Negotiator which is the framework that provides the computational resources (e.g., CPUs, memory, etc.) required to execute an application. There are two important elements of YARN ,these are:

- Resource Manager
- Node Manager

a) Resource Manager: Each cluster has one resource manager which is the master. It contains the details of the slaves such as s where the slaves are located and how many resources the slaves have. Several services are driven by it but the most important is the **Resource Scheduler** which decides how to assign the resources.
b)Node Manager : Multiple node managers can be there for each cluster. It is the slave of the infrastructure. When it starts, it reports to the Resource Manager. Also it gives the updates to the Resource Manager periodically. Each Node Manager offers some resources to the cluster. At run-time, the Resource Scheduler decides how to use this capacity. A container is a fraction of the Node Manager capacity and it is used by the client for running a program



C.HDFS Federation

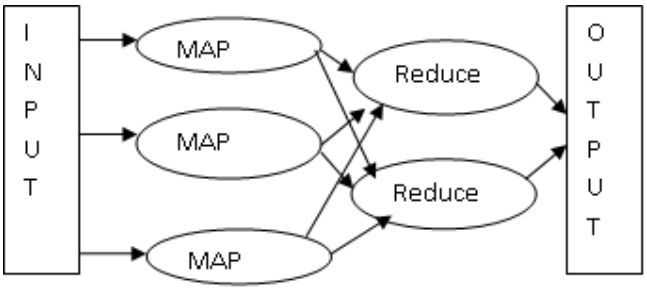
This is the framework that provides permanent, reliable and distributed storage. This is typically used for storing inputs and output but no intermediate data is stored here.

D. Storage

This provides alternative storage solutions just as an example Amazon uses the Simple Storage Service(S3).

E. Map Reduce Framework

This is the software layer that implements the Map Reduce paradigm. Map reduce framework works on MapReduce technique which is a processing technique and a program model for distributed computing based on java. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.



In short we can say that YARN infrastructure and the HDFS federation are completely decoupled and independent. YARN provides resources for running an application while the HDFS provides storage. The MapReduce framework is only one of many possible framework which runs on top of YARN.

IV. HADOOP ON CLOUD

Hadoop can be deployed in a traditional onsite datacenter as well as in the cloud. The cloud allows organizations to deploy Hadoop without hardware to acquire or specific setup expertise. Hadoop has come to be an incredibly important technology for many big data projects and applications. But without the proper time or training, it can be difficult to leverage this technology. Hadoop-as-a-service has grown to satisfy the need created by this situation. With its unlimited scale and on-demand access to compute and storage capacity, cloud computing is the perfect match for big data processing. Few vendors who are currently offering the cloud include Microsoft, Amazon, IBM, Google and Oracle. Hadoop as a Service offering has several advantages and disadvantages over on-premise solutions.

Advantages of Hadoop:

A) On-demand Elastic Cluster :

Unlike static, on-premise clusters, Hadoop clusters in the cloud scale up or down depending on data processing requirements. Nodes are automatically added to or removed from clusters depending on data size. This elastic property helps to improve performance.

B) Integrated Big Data Software:

Hadoop as a Service includes full integration with the Hadoop ecosystem. Connectors for data integration and creating data pipelines provide a complete solution that works with the current pipeline.

C) Simplified Cluster Management:

Hadoop provides a fully managed cluster, which eliminates the need of extra time and resources that are required to manage nodes, setting up clusters and scaling the infrastructure.

D) Lower Costs:

Hadoop in the Cloud does not require any upfront investment in on-site hardware or in IT support. Instant expenditure reduces up to 90% compared to on-demand instances. Amount is paid for the space only when it is used with auto-scaling clusters.

Hadoop also supports a range of related projects that can complement and extend Hadoop's basic capabilities. Such complementary software packages are:

- Apache Flume
- Apache HBase
- Apache Hive
- Cloudera Impala
- Apache Oozie
- Apache Phoenix
- Apache Pig
- Apache Sqoop
- Apache Spark
- Apache Storm
- Apache ZooKeeper

Disadvantages of Hadoop:

A) Security Concerns

At the storage and network levels encryption is missing which brings into picture security concerns.

B) Vulnerable By Nature

Hadoop framework is written almost entirely in Java, which is one of the most commonly used but controversial programming languages in existence. Java has been heavily exploited by cyber criminals and thus as an implication it brings in numerous security breaches.

C) Not Fit for Small Data

Hadoop is not suitable for small data needs.

Hadoop has high capacity design and Distributed File System which 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.

D) Potential Stability Issues

Like all open source software, Hadoop has had its fair share of stability issues. To avoid these issues, organizations are strongly recommended to make sure they are running the latest stable version, or run it under a third-party vendor equipped to handle such problems.

E) General Limitations

Apache Flume, MillWheel, and Google's own Cloud Dataflow are the platforms that have the ability to improve the efficiency and reliability of data collection, aggregation, and integration. Thus companies should use these platforms to get big benefits which they may miss out by using Hadoop alone.

IV. Conclusion

The emergence of Hadoop has changed the data landscape drastically . Hadoop helps to gain new or improved business insights from structured, unstructured and semi-structured data sources. Additionally large volumes of data which were previously very expensive to store among departments can be gathered and analysed in one place at an affordable price with Hadoop.

References

[1] https://en.wikipedia.org/wiki/Apache_Hadoop#Hadoop_hosting_in_the_cloud

[2] Jeffrey Dean, Sanjay Ghemawat (2004) MapReduce: Simplified Data Processing on Large Clusters, Google. This paper inspired Doug Cutting to develop an open-source implementation of the Map-Reduce framework. He named it Hadoop, after his son's toy elephant.

[3] https://www.thoughtworks.com/insights/blog/6-reasons-why-hadoop-cloud-makes-sense

[4] https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm

[5] http://www.infoworld.com/article/2607258/big-data/when-you-should-put-big-data-in-the-cloud.html

[6] http://www.itproportal.com/2013/12/20/big-data-5-major-advantages-of-hadoop/

[7] https://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm

[8] http://download.bigbata.com/ebook/oreilly/books/Hadoop.The.Definitive.Guide.3rd.Edition.May.2012.pdf