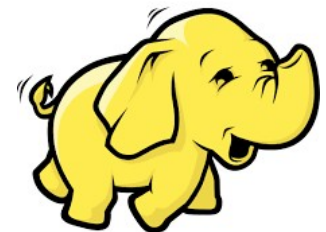


Introduction to Hadoop

Tushar B. Kute,
<http://tusharkute.com>



Hadoop

- 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.
- The Hadoop framework 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.

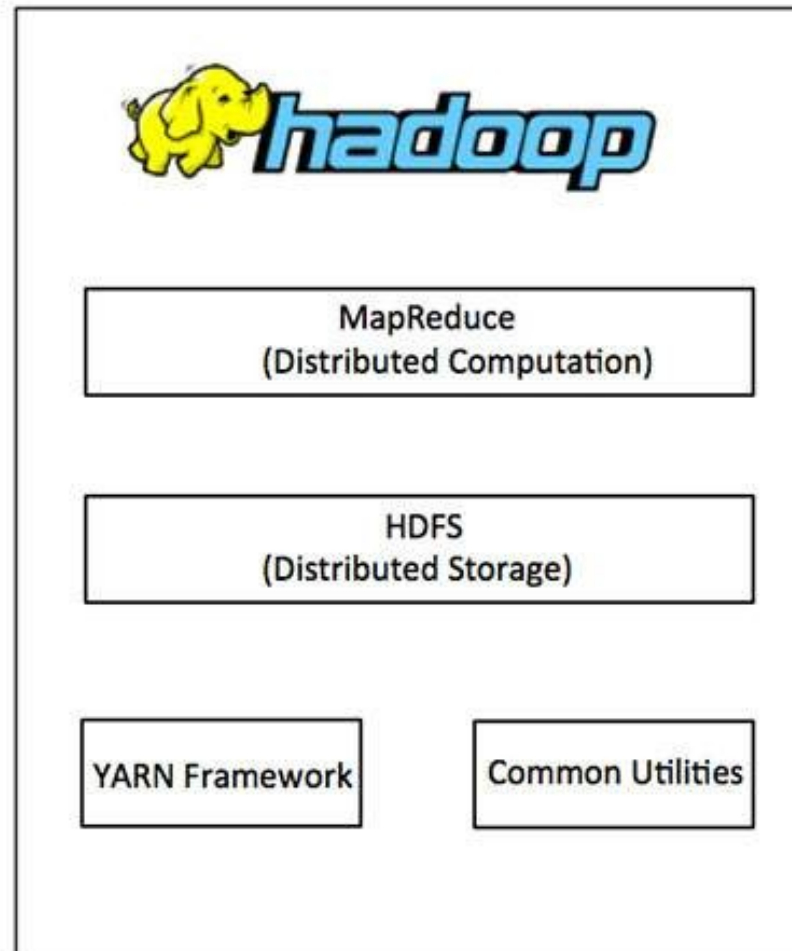
Hadoop

- Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume.
- Hadoop is written in Java and is not OLAP (online analytical processing).
- It is used for batch/offline processing. It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster.

Hadoop Architecture

- At its core, Hadoop has two major layers namely –
 - Processing/Computation layer (MapReduce), and
 - Storage layer (Hadoop Distributed File System).

Hadoop Architecture



Hadoop - Modules

- HDFS: Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.
- Yarn: Yet another Resource Negotiator is used for job scheduling and manage the cluster.

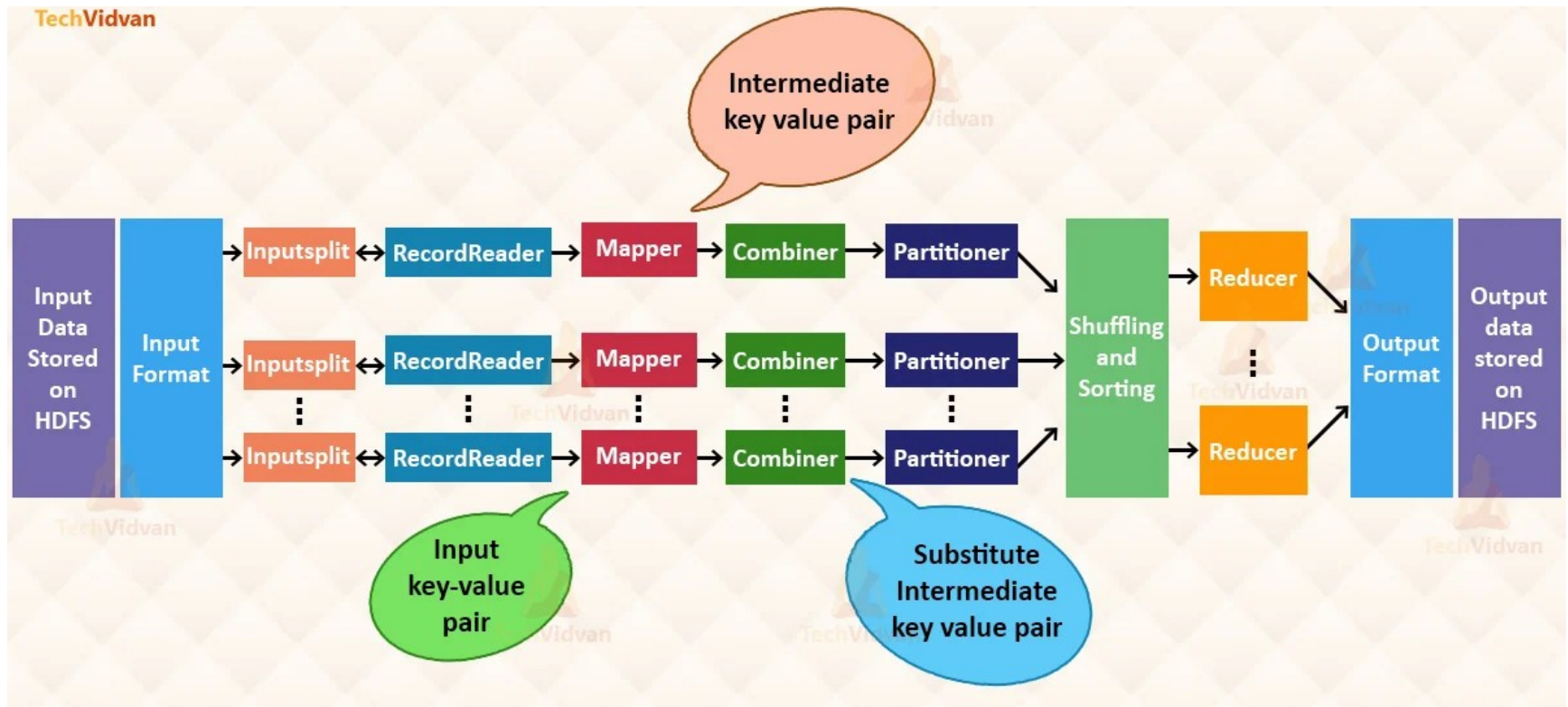
Hadoop - Modules

- Map Reduce: This is a framework which helps Java programs to do the parallel computation on data using key value pair.
 - The Map task takes input data and converts it into a data set which can be computed in Key value pair.
 - The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.
- Hadoop Common: These Java libraries are used to start Hadoop and are used by other Hadoop modules.

Mapreduce

- MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.
- The MapReduce program runs on Hadoop which is an Apache open-source framework.

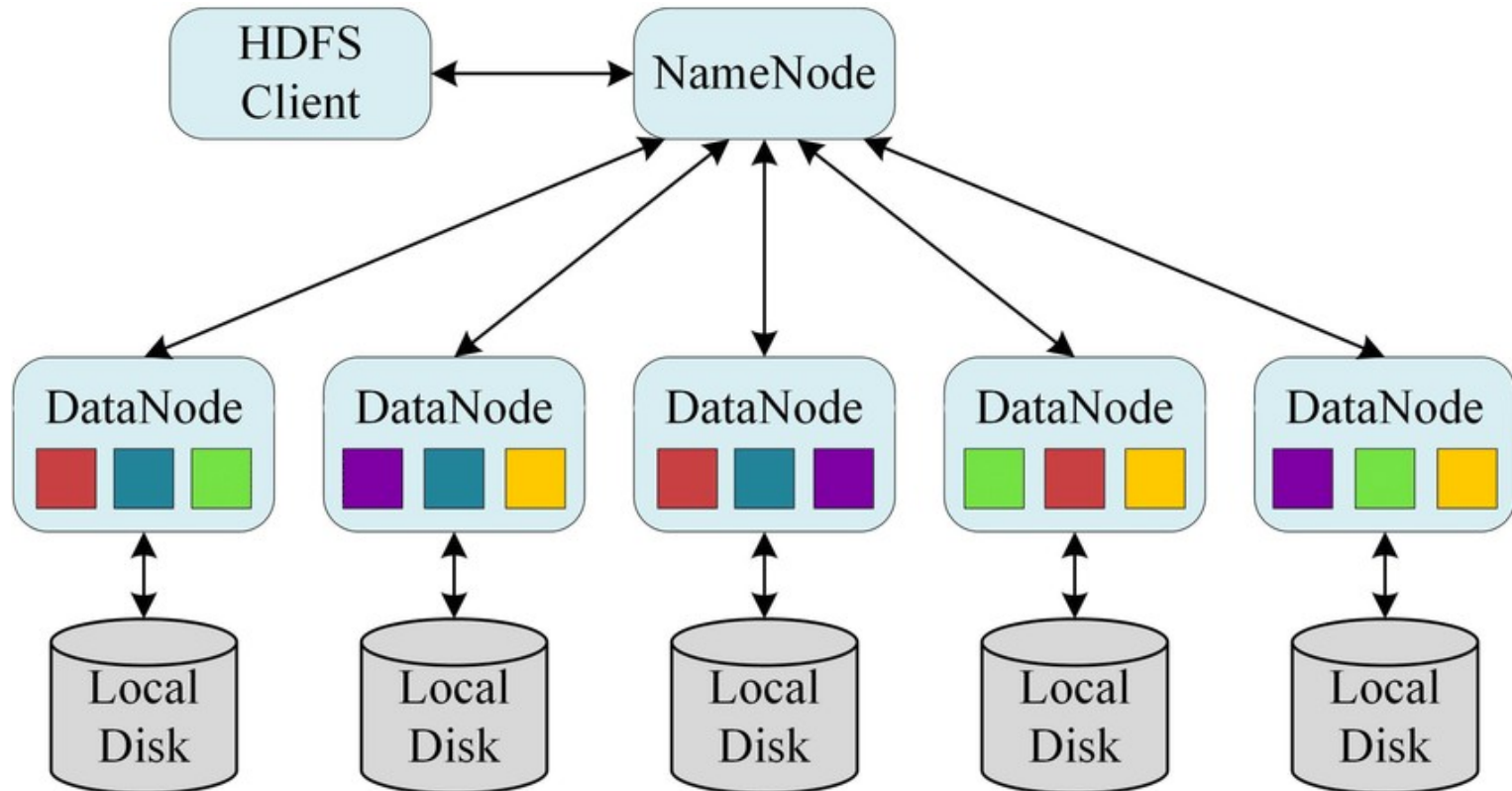
Mapreduce



HDFS

- 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 commodity hardware.
- It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant.
- It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.

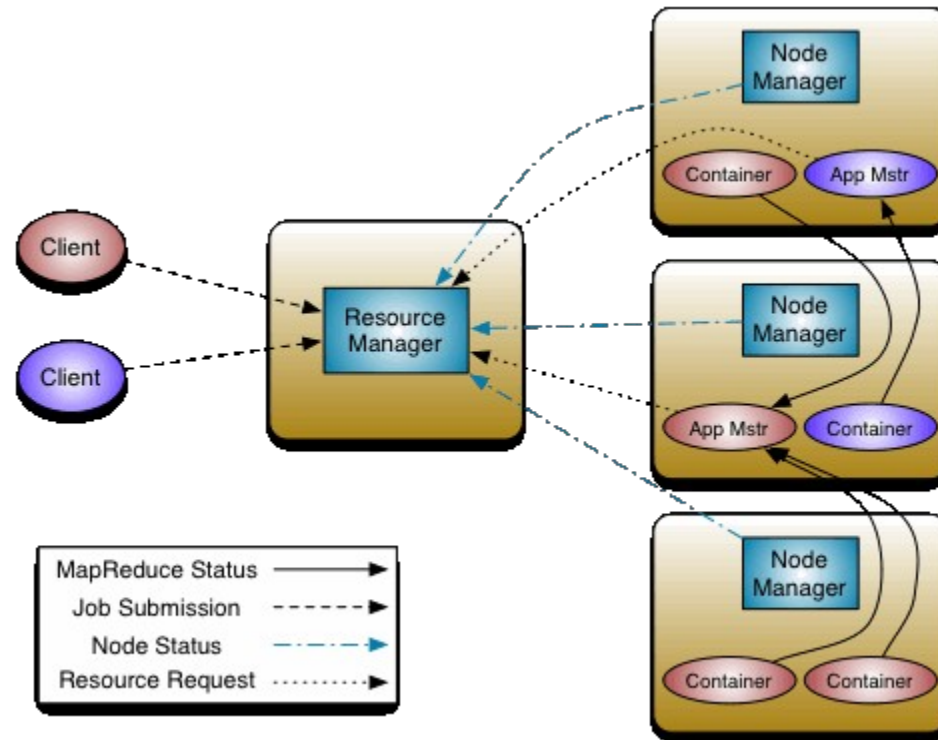
HDFS



Hadoop - Common and YARN

- Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules –
 - Hadoop Common – These are Java libraries and utilities required by other Hadoop modules.
 - Hadoop YARN – This is a framework for job scheduling and cluster resource management.

Hadoop - Common and YARN

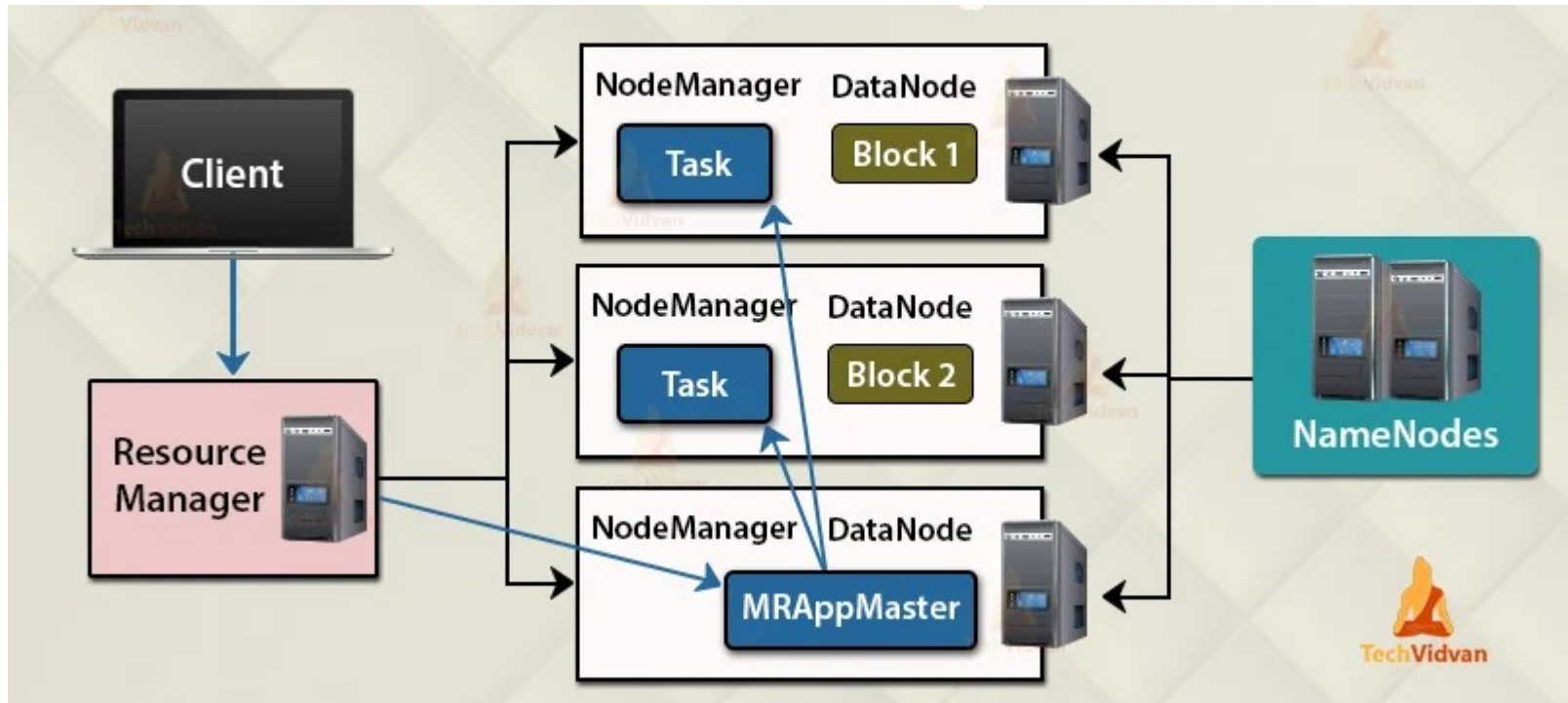


Ref: Apache Hadoop Official

Hadoop - How it works?

- It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput.
- Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines.

Hadoop - How it works?



Ref: TechVidvan

Hadoop - How it works?

- Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop performs –
 - Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).
 - These files are then distributed across various cluster nodes for further processing.
 - HDFS, being on top of the local file system, supervises the processing.
 - Blocks are replicated for handling hardware failure.
 - Checking that the code was executed successfully.
 - Performing the sort that takes place between the map and reduce stages.
 - Sending the sorted data to a certain computer.
 - Writing the debugging logs for each job.

Hadoop - Advantages

- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatically distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

Hadoop v1 vs. v2



Hadoop v1.0

MapReduce

Data Processing
& Resource Management

HDFS

Distributed File Storage



Hadoop v2.0

MapReduce

**Other Data
Processing
Frameworks**


YARN

Resource Management

HDFS

Distributed File Storage

Hadoop v1 vs. v2

Features	Hadoop 2.x	Hadoop 3.x 
Min Java Version Required	Java 7	Java 8
Fault Tolerance	Via replication	Via erasure coding
Storage Scheme	3x replication factor for data reliability, 200% overhead	Erasure coding for data reliability, 50% overhead
Yarn Timeline Service	Scalability issues	Highly scalable and reliable
Standby NN	Supports only 1 SBNN	Supports only 2 or more SBNN
Heap Management	We need to configure HADOOP_HEAPSIZE	Provides auto-tuning of heap

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License



@mitu_skillologies



/miTuSkillologies



@mitu_group



/company/mitu-
skillologies



MITUSkillologies

Web Resources

<https://mitu.co.in>
<http://tusharkute.com>

contact@mitu.co.in
tushar@tusharkute.com