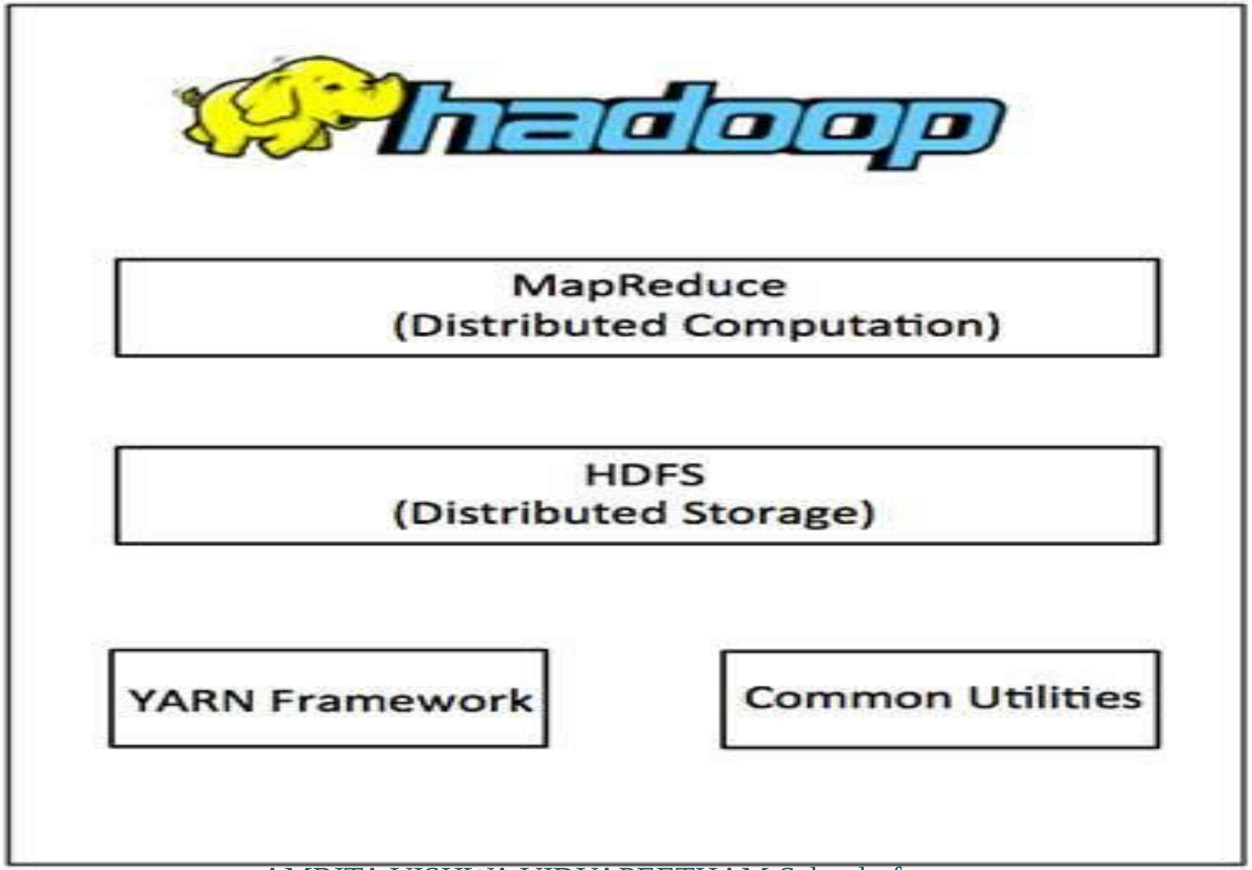
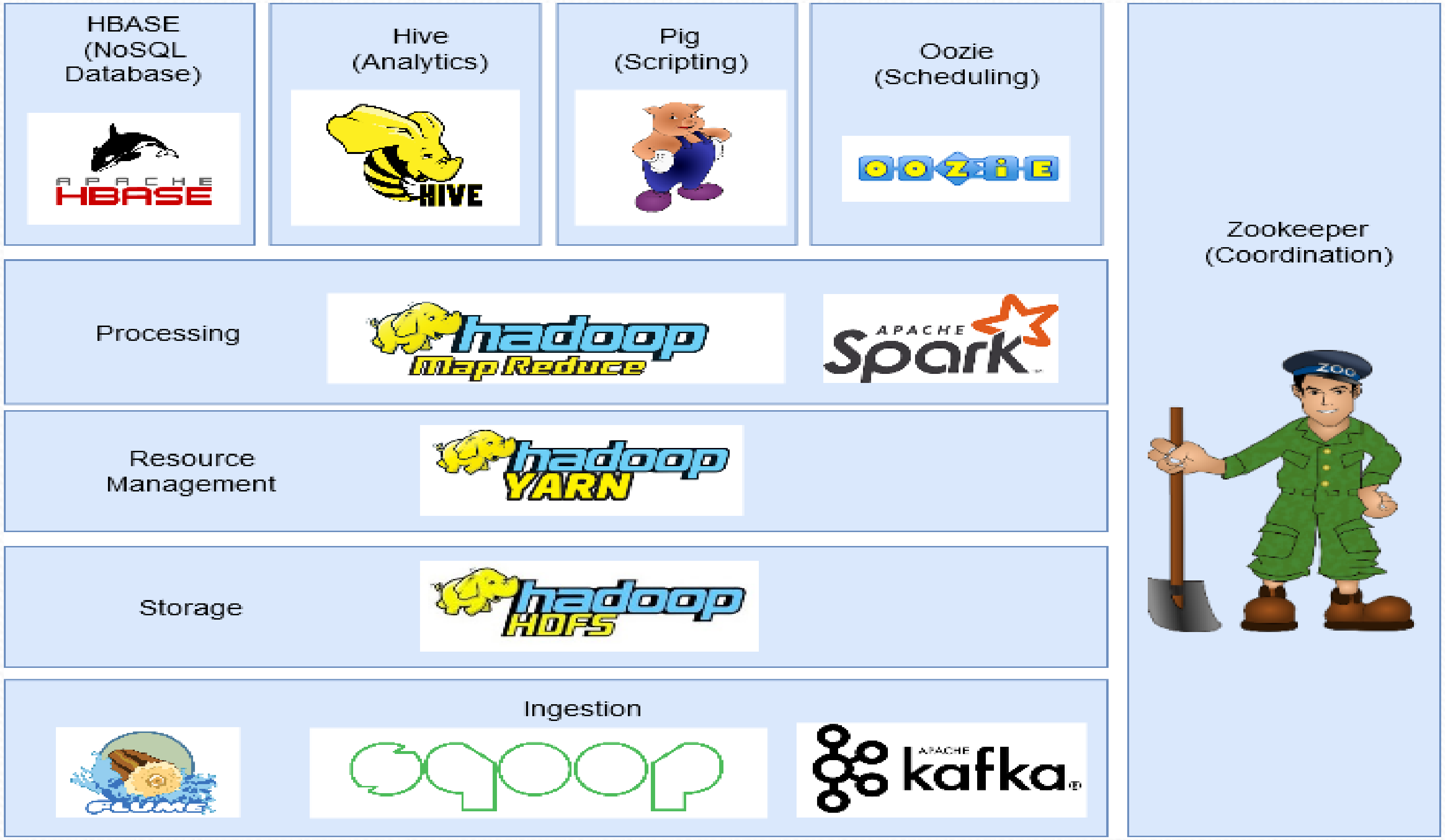
# HADOOP

## Hadoop

- **Hadoop** is an open-source software framework used for storing and processing **Big Data** in a distributed manner on large clusters of commodity hardware.
- Hadoop is licensed under the Apache v2 license.


hadoop
MapReduce
(Distributed Computation)
HDFS
(Distributed Storage)
YARN Framework
Common Utilities



HBASE (NoSQL Database)

Hive (Analytics)

Pig (Scripting)

Oozie (Scheduling)

Zookeeper (Coordination)

Processing

Resource Management

Storage

Ingestion

## Hadoop Distributed File System

- 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.
- 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.

## MapReduce

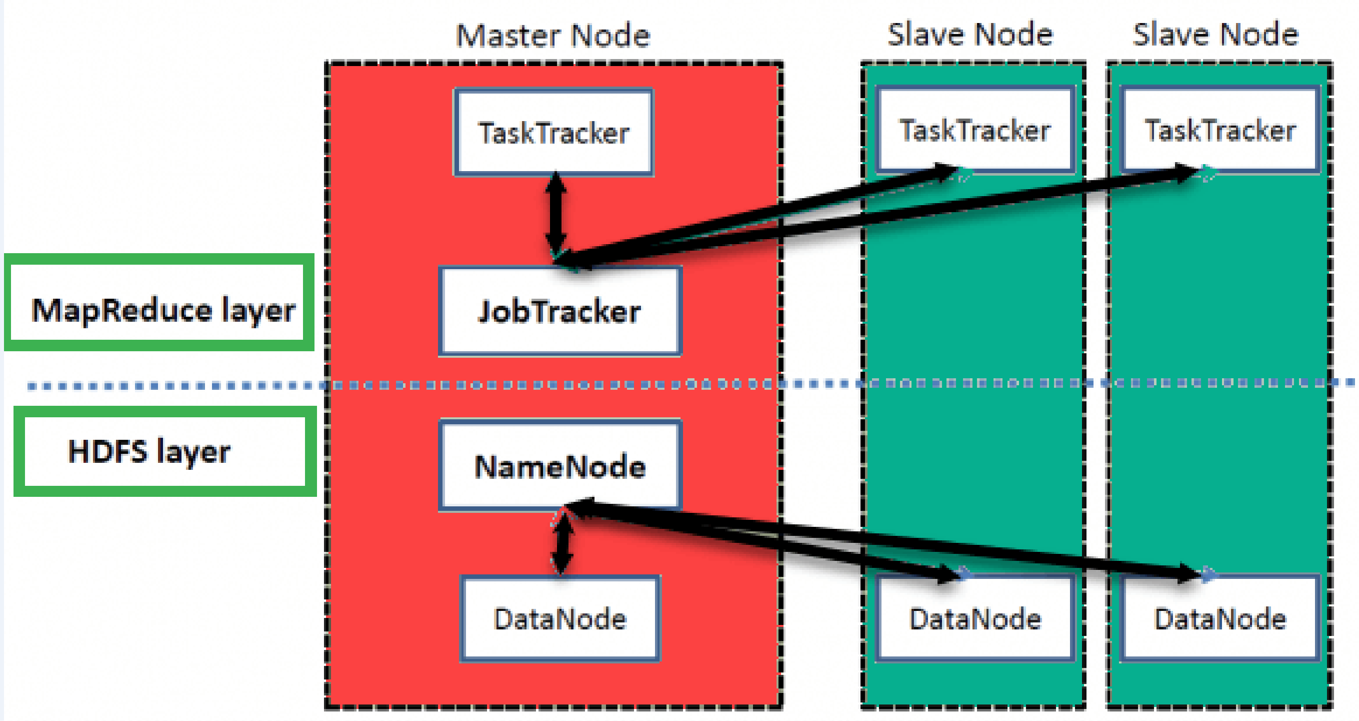
- MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.
- The MapReduce program runs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

## YARN

- **YARN** performs all your processing activities by allocating resources and scheduling tasks.
- Yet Another Resource Manager takes programming to the next level beyond Java , and makes it interactive to let another application Hbase, Spark etc. to work on it.
- Different Yarn applications can co-exist on the same cluster so MapReduce, Hbase, Spark all can run at the same time bringing great benefits for manageability and cluster utilization.
-

- Components Of YARN
- **Client:** For submitting MapReduce jobs.
- **Resource Manager:** To manage the use of resources across the cluster
- **Node Manager:**For launching and monitoring the computer containers on machines in the cluster.
- **Map Reduce Application Master:** Checks tasks running the MapReduce job. The application master and the MapReduce tasks run in containers that are scheduled by the resource manager, and managed by the node managers.

# High Level Hadoop Architecture


Master Node
Slave Node
Slave Node
TaskTracker
TaskTracker
TaskTracker
MapReduce layer
JobTracker
HDFS layer
NameNode
DataNode
DataNode
DataNode