The Apache Hadoop Module:

Hadoop Common: this includes the common utilities that support the other Hadoop modules

HDFS**:** the Hadoop Distributed File System provides unrestricted, high-speed access to the application data.

Hadoop YARN: this technology accomplishes scheduling of job and efficient management of the cluster resource.

MapReduce**:** highly efficient methodology for parallel processing of huge volumes of data.

Then there are other projects included in the Hadoop module that are no less important:

Apache Ambari**:** it is a tool for managing, monitoring and provisioning of the Hadoop clusters. Ambari supports the HDFS and MapReduce programs. Some of the major highlights of Ambari are:

- It makes managing of the Hadoop framework highly efficient, secure and consistent
- It manages the cluster operations with an intuitive web UI and a robust API
- The installation and configuration of Hadoop cluster are highly simplified
- It supports automation, smart configuration and recommendations
- Advanced cluster security set-up comes along with this tool
- The entire cluster can be regulated using metrics, heatmaps, analysis and troubleshooting
- Increased levels of customization and extension makes Ambari highly valuable

Cassandra**:** it is a distributed system to handle extremely large amounts of data that is stored across several commodity servers. The hallmark of this database management system is high availability with no single point of failure.

HBase**:**it is a non-relational, distributed database management that works very well on sparse data sets and it is highly scalable.

Apache Spark**:** it is an extremely agile, scalable and secure Big Data compute engine that is versatile enough to work on a wide variety of applications like real-time processing, machine learning, ETL and so on.

Hive:it is a data warehouse tool for analyzing, querying and summarizing of data on top of the Hadoop framework.

Pig: a high-level framework that can be used to work in coordination either with Apache Spark or MapReduce to analyze the data. The language to program for this platform is called Pig Latin.

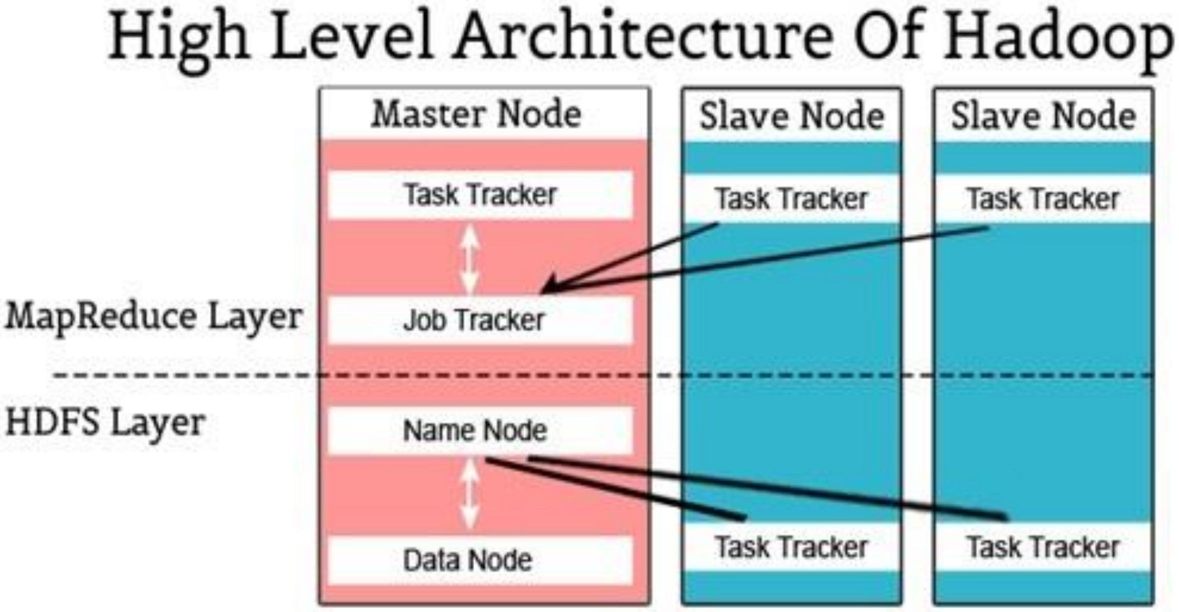
Sqoop: a framework for transferring data to Hadoop from relational databases. This application is based on a command-line interface.

Oozie: it is a scheduling system for workflow management, executing workflow routes for successful completion of the task in a Hadoop set-up.

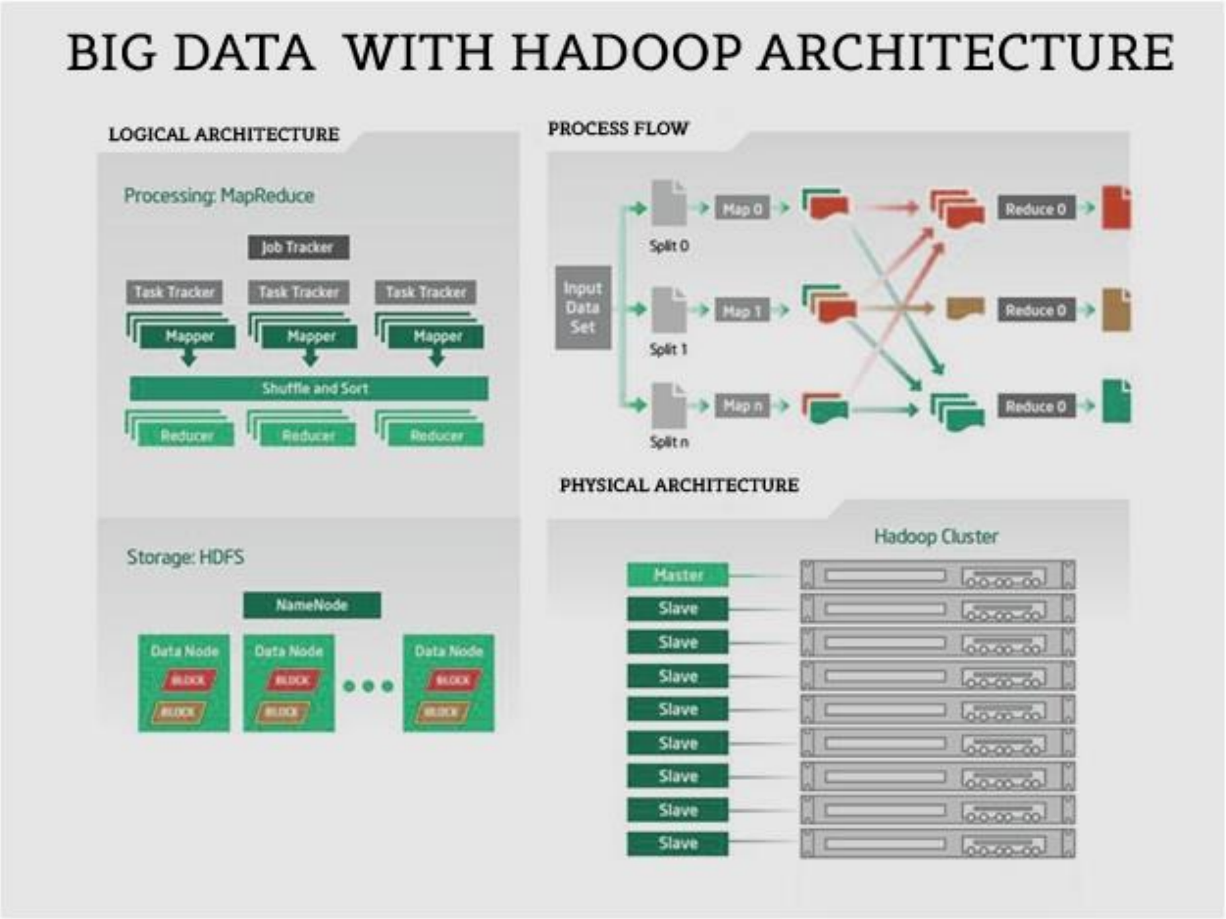
Zookeeper: it is an open source centralized service that is used to provide coordination between distributed applications of Hadoop. It offers naming registry and synchronization service on a massive level.

The Hadoop High-level Architecture:

Hadoop Architecture based on two most vital components viz. MapReduce and HDFS



Different Hadoop Architectures based on the Parameters chosen:


BIG DATA WITH HADOOP ARCHITECTURE
LOGICAL ARCHITECTURE
Processing: MapReduce
Job Tracker
Task Tracker
Task Tracker
Task Tracker
Mapper
Mapper
Mapper
Shuffle and Sort
Reducer
Reducer
Reducer
Storage: HDFS
NameNode
Data Node
Data Node
Data Node
BLOCK
BLOCK
PROCESS FLOW
Input Data Set
Split 0
Split 1
Split n
Map 0
Map 1
Map n
Reduce 0
Reduce 0
Reduce 0
PHYSICAL ARCHITECTURE
Hadoop Cluster
Master
Slave
Slave
Slave
Slave
Slave
Slave
Slave
Slave
Slave