**ASSIGNMENT NO.:**

**Aim:** Create BIG DATA applications using Hadoop through twitter app.

.**Objective:** To learn Hadoop system and handle large data.

**Theory:**

**Hadoop**

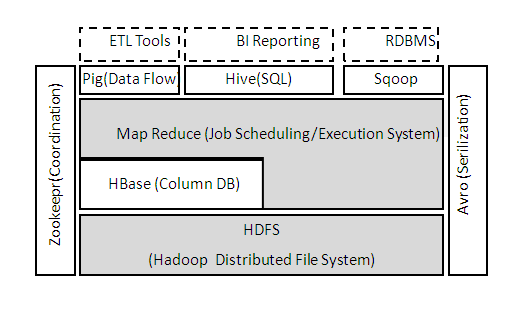
* Hadoop is a software framework for *distributed processing* of *large datasets* across *large clusters* of computers
  + ***Large datasets*** 🡪 Terabytes or petabytes of data
  + ***Large clusters*** 🡪 hundreds or thousands of nodes
* Hadoop is open-source implementation for Google ***MapReduce***
* Hadoop is based on a simple programming model called *MapReduce*
* Hadoop is based on a simple data model, *any data will fit*

**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.

The Apache Hadoop framework is composed of the following modules:

* *Hadoop Common* – contains libraries and utilities needed by other Hadoop modules.
* *Hadoop Distributed File System (HDFS)* – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
* *Hadoop YARN* – a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.
* *Hadoop MapReduce* – a programming model for large scale data processing.

**Hadoop ecosystem components**

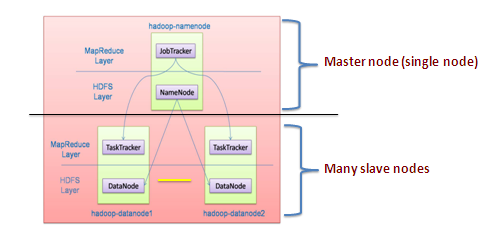


* ***HDFS***
  + A distributed filesystem that runs on large clusters of commodity machines.
* ***MapReduce*** 
  + A distributed data processing model and execution environment that runs on large

clusters of commodity machines.

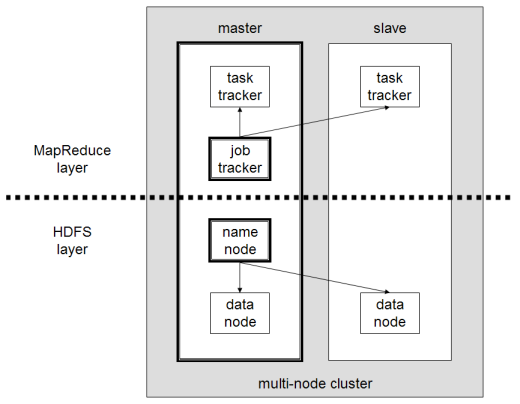
* ***Pig***
  + A data flow language and execution environment for exploring very large datasets.
* ***Hive***
  + A distributed data warehouse. query language based on SQL (and which is translated by the runtime engine toMapReduce jobs) for querying the data.
* ***HBase*** 
  + A distributed, column-oriented database.
* ***ZooKeeper*** 
  + A distributed, highly available coordination service.
* ***Sqoop*** 
  + A tool for efficiently moving data between relational databases and HDFS.

**Hadoop Master/Slave Architecture**

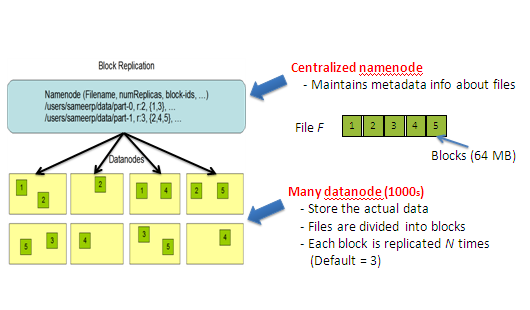


**Hadoop framework consists of two main layers**

* + Distributed file system (HDFS)
  + Execution engine (MapReduce)



**Hadoop Distributed File System (HDFS)**



**Main Properties of HDFS**

* ***Large:*** A HDFS instance may consist of thousands of server machines, each storing part of the file system’s data
* ***Replication:***Each data block is replicated many times (default is 3)
* ***Failure:*** Failure is the norm rather than exception
* ***Fault Tolerance:*** Detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS
  + Namenode is consistently checking Datanodes

**Input:** Live tweets from twitter.

**Output:** Analysis on Big Data .

**Platform:** OS-64 bit Ubuntu 12.01,Hadoop

**Conclusion:** Thus, we handled BIG Data.