**Hadoop**

Hadoop is an open-source framework. It’s provided by Apache to process and analyze very huge volume of data. It is written in Java and is used by Google, Facebook & Twitter.

It is used for batch processing or offline processing. Moreover, it can be scaled by adding more nodes in the cluster.

90% of todays data was generated in the past 3 years. Sources of this type of data are,

1. Social networking websites
2. E-commerce site
3. Weather station
4. Telecom company
5. Share market

3V’s of Big Data:

1. Velocity
2. Variety
3. Volume

**Write Once Read Many**

Using HDFS (Hadoop Distributed File System) data can be stored in distributed fashion by employing commodity hardware to form clusters. To analyze the data, we can use Pig or Hive.

**Modules of Hadoop**

Mainly consists of the following modules

1. HDFS

It states that the files can be broken into blocks and stored in nodes over the distributed architecture.

1. YARN

Yet Another Resource Navigator is used for job scheduling and managing the cluster.

1. 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 key value pair. The output of the Map task is fed to Reduce task and it outputs the desired result.

1. Hadoop Common

These Java libraries used by Hadoop and used to start Hadoop and are used by other Hadoop modules.

Hadoop Architecture

The Hadoop architecture is a package of file system, MapReduce engine & HDFS. A Hadoop cluster consists of a single master and slave nodes.

The master node includes a

1. Job Tracker
2. Task Tracker
3. Data Node
4. Name Node

whereas a slave node only consists of

1. Task Tracker
2. Data Node

