1.HDFS

The hadoop distributed file system is a distributed ,scalable and portable file system written in java for the Hadoop framework.HDFS stores large files(GB to TB) across multiple machines .It also provides fault-tolerant storage at low cost.Files stored on HDFS are decomposed into blocks and each block is written to more than one servers.

It achieves reliability by recplicating the data across multiple host(3 by default) .

It was designed for 3 basic features –

1. Very large files – might go upto petabytes.

2. Streaming data access – Data is wriiten once but read many times.

3. Commodity hardware – thus low on cost.

Over the years storage capacities of hard drives have increased but the access speeds have not been increased with the same pace. Thus HDFS divides the data over multiple machines for faster parallel processing of the data , hence the name distributed file system comes into existence.

2.Hadoop cluster

A Hadoop cluster is a special type of computational cluster (number of commodity machines) designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment. These clusters run on low-cost commodity machines. One machine in the cluster is called the Name Node (the master,) and others are knows as DataNodes (salves). Hadoop clusters are knows for boosting the speed of data analysis applications. They are highly scalable. If at some point cluster’s processing power is overwhelmed the cluster can always be scaled out to accommodate more machines. Clusters are also fault-tolerant as the data is copied to several machines (Redundancy) in case of failures of datanodes .

3.HDFS blocks

HDFS file system stores data in terms of blocks. Blocks can be defined as quantized data retainers. Block size in HDFS is 128 mb , also these blocks are continuous in nature to minimize the head seek time of a HDD. As the blocks are of fixed size it is very easy to calculate the no. of blocks that can be stored on a disk. The namenode maintains the metadata of all the blocks present on datanodes. Suppose the file size is less than the block size, then the file does not occupy the complete block storage.

Blocks are also easy to replicate between datanodes and thus provide fault-tolerance and high availability.