1: HBASE:

HBase is an open source, distributed, versioned, column-oriented, No-SQL / Non-relational database management system that runs on the top of Hadoop (To Install Hadoop Follow this installation Guide). It adds transactional capability to hadoop, allowing users to update data records. Hadoop is designed for batch processing of large dataset, but with HBase on the top of Hadoop we can process real time dataset.In HBase a master node manages the cluster and region servers store portions of the tables and perform the work on the data. An HBase system comprises a set of tables. Each table contains rows and columns, much like a traditional database. Each table must have an element defined as a Primary Key, and all access attempts to HBase tables must use this Primary Key. An HBase column represents an attribute of an object.

HDFS

HDFS is a distributed file system which provides redundant storage space for storing files which are very huge in sizes; files which are in the range of Terabytes and Petabytes. In HDFS data is stored reliably. Files are broken into blocks and distributed across nodes in a cluster. After that each block are replicated, means copies of blocks are created on different machines. Hence if a machine goes down or get crashed, then also we can easily retrieve and access our data from different machines. By default 3 copies of a file are created on different machines. Hence it is highly fault tolerant. HDFS provides faster file read and write mechanism, as data is stored in different nodes in a cluster. Hence user can easily access the data from any machine in a cluster. Hence HDFS is highly used as a platform for storing huge volume and different varieties of data worldwide.

2: Component of HBASE:

a. HMaster

b. HRegion Server

c. Region

d. Zookeeper

3: No Hbase does not support SQL commands, it has own set of commands to create, insert, read data.

4: HBase should be used when the big data application has –

A variable schema

When data is stored in the form of collections

If the application demands key based access to data while retrieving.

5: Modes in which Hbase can run are standalone and Distributed.

Standalone mode: By default Hbase run in standalone. It does not require HDFS but uses local filesystem instead. In this mode, all HBase daemons and local zookeeper run in same JVM.

Zookeeper binds to a well known port so that client can talk to HBase directly

Distributed mode: Distributed mode is also divided into pseudo-distributed and fully distributed mode. In pseudo-distributed mode all daemons run in the same node while in fully distributed mode daemons are spread across the nodes in the cluster.

Distributed mode requires an instance of HDFS.

6: HBase uses ZooKeeper as a distributed coordination service to maintain server state in the cluster.

Zookeeper maintains which servers are alive and available, and provides server failure notification. Zookeeper uses consensus to guarantee common shared state. Note that there should be three or five machines for consensus.

ZooKeeper service keeps track of all the region servers that are there in an HBase cluster- tracking information about how many region servers are there and which region servers are holding which DataNode.

HMaster contacts ZooKeeper to get the details of region servers.

Various services that Zookeeper provides include –

Establishing client communication with region servers.

Tracking server failure and network partitions.

Maintain Configuration Information

Provides ephemeral nodes, which represent different region servers.

8: HBase is a schema less database since it does not define fixed number of columns as against the RDBMS defined schema with fixed columns. Column families are defined which may contain any number of columns. It is column oriented database. Tables in it are sorted by row where each row contain column families.

9: Minimum number of column family should be one. We should introduce second and third column family only when data access is column scoped i.e. when we query one column family or other but usually not both at same time.

10: When applications require high end multithreaded access like application server or web server that serves many application threads in a single JVM, connection to database are created earlier for good performance so that it can directly use existing open connection from pool and there is no need to create new connection.