Explain in brief

● Differences between HBASE and HDFS.

Basically we can say they HBASE is a framework which sits on top of HDFS. HDFS(Hadoop distributed file system) is used to store the huge amount of data in a distributed way. HBASE is used to get extra functionality to HBASE.

HBASE is a non-relational database that can run on top of Hadoop and provides you random data access/querying capabilities. HDFS, by itself has no support for reads/writes at random location.

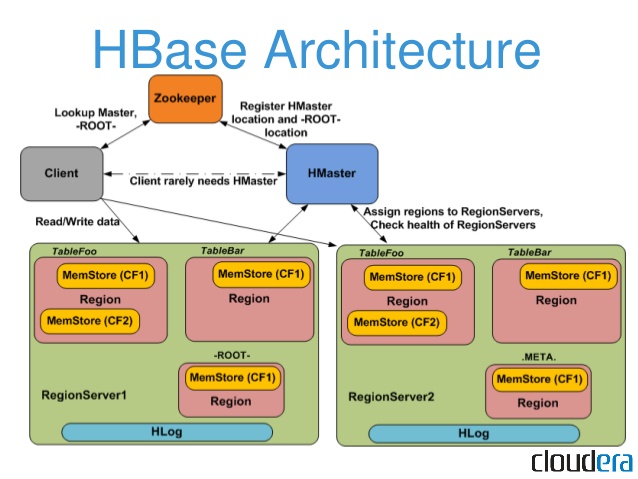
Following are the facilities that are provided by HBASE.

1. The basic difference in HDFS and HBASE is that the later has the flexibility of accessing the records randomly but HDFS cannot access the specific record.
2. Data storing is done in the columnar manner
3. Basically HABSE is NoSQL database. Hence the traditional select, insert statements do not work in this case.
4. Storing of the values in the HBASE is like a key value pair.
5. Defining the columns we do not need the datatypes. HBASE stores everything in the bytes. So the column is not restricted to only one datatype.
6. We can store different version of the same column which helps us to look into the historic data of the same column.
7. In the traditional RDBMS system if the column is having null values still some amount of space is allotted and basically we are wasting that space. But case of the HBASE the space is left blank and hence that amount of the space is saved and in the world of competitive data storage these plays very important.

Limitations of the HBASE

1. It can't be used for classic transactional applications or even relational analytics.
2. It is also not a complete substitute for HDFS when doing large batch MapReduce.
3. It doesn’t talk SQL, have an optimizer, support cross record transactions or joins.
4. It can't be used with complicated access patterns (such as joins)

● List and explain the main components of HBASE.



Basic Components of HBASE

1. HMASTER
2. REGION SERVER
3. ZOOKEEPER

HMASTER

It allocates the regions to every region server in the Hadoop cluster. It performs the function of monitoring of the whole Hadoop cluster and keeps the check on health of each region. DDL operations are all handled by Hmaster. Whenever a client wants to change the schema and change any of the metadata operations, HMaster is responsible for all these operations.

REGION SERVER

It handles the read write, update, and delete requests from user. Whenever Hmaster receives the request it first searches the place of the record and which region server is serving the node and handovers the query to the region server. Region Server runs on HDFS DataNode and consists of the following components.

Block Cache-This is the read cache. Most frequently read data is stored in the read cache

MemStore- This is the write cache and stores new data that is not yet written to the disk. Every column family in a region has a MemStore.

HFile- It is the actual storage file that stores the rows as sorted key values on a disk.

ZOOKEEPER

HBase uses Zookeeper as a distributed coordination service for region assignments and to recover any region server crashes by loading them onto other region servers that are functioning. Zookeeper is a centralized monitoring server that maintains configuration information and provides distributed synchronization.

● Does Hbase support sql?

HBASE is basically a NoSQL Database so it will not support the sql. Also the working of HBASE is different than that of the traditional RDBMS system and sql was used to query the RDBMS system. But new projects have been taken up so as to get the similar interface as that of the sql. By default Hbase is not supporting sql but we have to build up on HBASE to support sql some of them are

Phoenix, impala, hive ,etc.