**31.1**

**Problem Statement:**

Explain in brief

● Differences between HBASE and HDFS.

● List and explain the main components of HBASE.

● Does Hbase support sql?

**Solution:**

* **Differences between HBASE and HDFS:**
* **HBase:**

HBase is an open source, distributed, versioned, column-oriented, No-SQL / Non-relational database management system that runs on the top of Hadoop.

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.

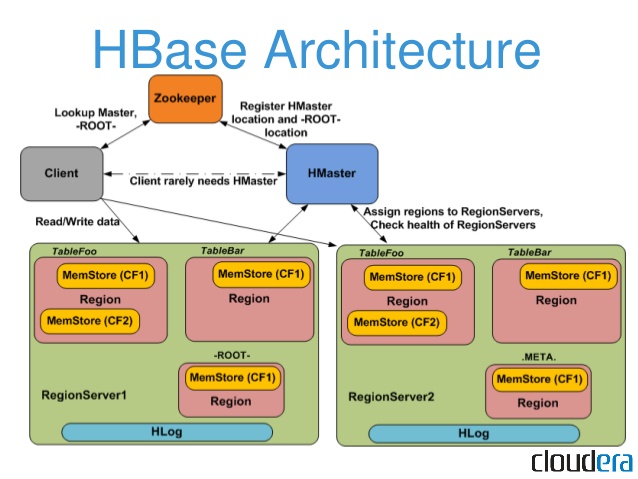
* **HDFS:**

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.

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.

* **Components of Apache HBase Architecture:**

****

**Fig: Apache HBase Architecture**

HBase architecture has 3 important components-

* HMaster
* Region Server and
* ZooKeeper
* **HMaster:**

HBase HMaster is a lightweight process that assigns regions to region servers in the Hadoop cluster for load balancing.

Responsibilities of HMaster –

* Manages and Monitors the Hadoop Cluster
* Performs Administration (Interface for creating, updating and deleting tables.)
* Controlling the failover
* **Region Server:**

These are the worker nodes which handle read, write, update, and delete requests from clients. Region Server process, runs on every node in the hadoop cluster. 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 and whenever the block cache is full, recently used data is evicted.
* 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.
* **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.

In case of node failure within an HBase cluster, ZKquoram will trigger error messages and start repairing failed nodes.

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.
* **Does Hbase support sql?**

HBase does not support sql.

HBase is a column-oriented database management system that runs on top of hdfs. It is well suited for sparse data sets, which are common in many big data use cases.

For example, if the table is storing diagnostic logs from servers in your environment, where each row might be a log record, a typical column in such a table would be the timestamp of when the log record was written, or perhaps the server name where the record originated..

Just as HDFS has a NameNode and slave nodes, and MapReduce has JobTracker and TaskTracker slaves, HBase is built on similar concepts. In HBase a master node manages the cluster and region servers store portions of the tables and perform the work on the data.