Assignment-31.1:

**Differences between HBASE and HDFS.**

|  |  |
| --- | --- |
| **HBASE** | **HDFS** |
| HBase - open source, distributed, versioned, column-oriented,  No-SQL / Non-relational database management system is needed to run on top of Hadoop. | HDFS –is called as distributed file system which will distribute data over cluster of machines and will take care of the redundancy. |
| HBase will store data as key/value pairs as in a column database.  Each key will hold a value. | HDFS is used to store data in the form of flat files. |
| HBASE has the ability of random read/writes of data. | Instead HDFS will not support random reads. |
| HBase will provide flexible data model**.** | HDFS in contrast will not provide with any flexible data model. |
| HDFS files will be written once and read many times.  And there is no option for random write or read. | Data will be indexed by row key  Will be provided with very flexible data model  data will stored inside the hashed table and will be accessed in a random manner |
| HDFS will never support fast individual record lookups | Whereas HDFS will support faster data lookup in the tables |

***List and explain the main components of HBASE.***

HBase is mainly comprises of 3 types of servers in a master slave manner:

1. **Region server:** will serve data for read and write process.
2. **HBase Master**: will process to handle Region assignment, DDL (create, delete tables) related operations.
3. **Zookeeper** : used to maintain the live cluster state.

*HBase master:*

* HBase HMaster acts as a lightweight process
* It will assign region to region servers in Hadoop cluster for the process of load balancing.

*Responsibilities of HMaster are–*

* Hadoop Cluster will be Managed and Monitored.
* used to Perform Administration like Interface for creating, updating as well deleting tables.
* to Control the failover
* DDL operations will be handled by HMaster.
* Will assign regions to region server
* And will seek help of Apache Zoo Keeper for this task.

*Region Server:*

* Region server perform like worker nodes
* And used to handle read, write, update, and delete those requests from client end.
* Region Server process will runs on all the node in hadoop cluster.
* Region Server will run on HDFS DataNode
* As well it will consists these components –

**Block Cache –**

* Block cache is used to read cache.
* All the read data will be stored in read cache
* Once the block cache becomes full, recently used data will get evicted.

**MemStore-**

* Write cache will stores new data .
* Which is not yet written to disk.
* All the column family inside a region will have MemStore.
* *Write Ahead Log (WAL)* is a file used to store new data.
* And it will not be persisted to the permanent storage.

## Zookeeper:

* Zookeeper acts as open-source project.
* Services like maintaining configuration information, naming, providing distributed synchronization, etc. will be provided.
* Zookeeper will have ephemeral nodes which represents different region servers.
* Master servers will use those nodes to find all those available servers.
* Adding to availability, all nodes were used to track server failures or network partitions.
* Clients will communicate with region server via zookeeper.
* pseudo and standalone modes HBase were used to take care of the zookeeper.
* HBase will make use of Zoo Keeper which will act as distributed coordination service in order to maintain server state in cluster.

Various services provided by Zookeeper are:

1. Establishing client communication with region servers.
2. Tracking server failure and network partitions.
3. Maintain Configuration Information.

**Does Hbase support sql?**

Native Hbase will not support Sql like advance queries

to have Sql like queries you need phoenix on top of hbase.

HBase non-relational (NoSQL) database that runs on top of HDFS

It will acts as a open source

SQL database that provide real-time read/write to access large datasets.