**ACD\_BDD\_Session\_17\_Assignment\_1\_Main**

**Problem Statement**

**Give a brief elaboration of the below questions.**

**1. Give a brief difference between HBASE and HDFS.**

|  |  |
| --- | --- |
| **HDFS** | **HBase** |
| HDFS is a distributed file system suitable for storing large files. | HBase is a database built on top of the HDFS. |
| HDFS does not support fast individual record lookups. | HBase provides fast lookups for larger tables. |
| It provides high latency batch processing; no concept of batch processing. | It provides low latency access to single rows from billions of records (Random access). |
| It provides only sequential access of data. | HBase internally uses Hash tables and provides random access, and it stores the data in indexed HDFS files for faster lookups. |

**2. List the main components of HBASE.**

In HBase there are three main components: Master, Region Server and Zookeeper. The other components are MemStore, HFile and WAL(Write Ahead Log).

**3. Does HBase support SQL?**

No, HBase does not support SQL.

**4. When should we use HBASE, list some of the scenarios for the same.**

HBase has mainly 3 characteristics: Fault tolerant, Fast and Usable.

So we can use HBase in the following scenarios:

* When we need high fault tolerance because HBase provides high availability through automatic failover, automatic sharing and load balancing of tables.
* When we need fast lookup because HBase provides near real time lookups, in memory caching via block cache and bloom filters, server side processing via filters and co-processors.
* When we need to reuse existing data models because HBase can use data models via Java API's as well as Thrift & Rest gateway API's.

**5. What are the different modes in which HBase can be run?**

HBase has 2 modes: Standalone and Distributed.

**6. Why is zookeeper needed in HBase?**

-> A distributed HBase relies completely on Zookeeper (for cluster configuration and management). In apache HBase, Zookeeper coordinates, communicates and shares state between the Master and Region Servers. HBase has a design policy of using Zookeeper only for transient data (e.g; for coordination and state communication). Thus, if the HBase's Zookeeper data is removed, only the transient operations are affected - data can continue to be written and read to/from HBase.

**8. HBase is a schema less database, what does it mean?**

HBase is schema less, that means it does not have the concept of fixed columns schema; it defines only column families.

**9. What is the minimum number of column family every HBase table should have?**

HBase should have minimum 1 column family.

**10. What is the benefit of using connection pool in HBase?**

Connection pool provides high end multi threading access.