**ACD\_BDD\_Session\_17\_Assignment\_2\_Main**

**Problem Statement**

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

**1. What is the difference between MemStore and HFile in HBase?**

-> MemStore:

The MemStore stores updates in memory as sorted Key-Values, the same as it would be stored in an HFile. There is one MemStore per column family. The updates are sorted per column family.

HFile:

Data is stored in an HFile which contains sorted Key-Values. When the MemStore accumulates enough data, the entire sorted Key-Value set is written to a new file in HDFS. This a sequential write. It is very fast as it avoids moving the disk drive head.

**2. Describe compactions in HBase.**

-> When something is written to HBase it is written to an in-memory store (MemStore), once this MemStore reaches a certain size, it is flushed to disk in a store file (everything is also written immediately to a log file for durability). The store files created on disk are immutable. Sometimes the store files are merged together, this is done by a process called compaction.

There are 2 types of compactions:

Minor compactions: These are triggered each time a MemStore is flushed and will merge some of the store files determined by some algorithm.

Major compactions: These run about every 24 hours (after the currently oldest file was written), and merge together all files into one. Major compactions can also be triggered manually.

**3. List and explain the logical entities in HBase.**

**4. What will happen if we do not create a row key while inserting the data?**

-> A row key is an unique key in Hbase table, without row key we cannot retrieve the data of a single row using "get" from Hbase table.

**5. How can filters be applied in HBase and what are the benefits?**

-> When reading data from HBase using Get or Scan operations, you can use custom filters to return a subset of results to the client. While this does not reduce server-side IO, it does reduce network bandwidth and reduces the amount of data the client needs to process. Filters are generally used using the Java API, but can be used from HBase Shell for testing and debugging purposes.

**6. What are the data model operations in HBase?**

-> The four primary data model operations are -

* Get
* Put
* Scan
* Delete

**7. How can MapReduce be used with HBase?**

-> Data stored in HDFS can be used for offline or batch analysis but not necessarily for real-time analysis.

Hbase addresses this by being both a key/value store for real-time analysis and supporting MapReduce for batch analysis.

**8. What is Region Server?**

-> The region servers have regions and store -

Regions does the following:

* Communicate with clients and handle data related operations
* Handle read and write requests for all the regions under it
* Decide the size of the region by following region size thresholds

Store:

It contains memory store and HFiles. Memstore is like a cache memory. Anything that is stored in the Hbase is stored here initially. Later, the data is transferred and saved in HFiles as blocks and the memstore is flushed.

