**Session 17**

**Assignment 2**

**Q 1. What is the difference between memstore and hfile in HBase?**

**Answer:** The **MemStore** is a write buffer where HBase accumulates data in memory before a permanent write.Its contents are flushed to disk to form an HFile when the MemStore fills up.  
It doesn’t write to an existing HFile but instead forms a new file on every flush.   
There is one MemStore per column family. (The size of the MemStore is defined by the system-wide property in hbase-site.xml called hbase.hregion.memstore.flush.size) Whereas

The **HFile** is the underlying storage format for HBase. HFiles belong to a column family and a column family can have multiple HFiles. But a single HFile can’t have data for multiple column families

**Q 2. Describe compactions in HBase.**

**Answer:** As more and more data is written to Hbase, many HFiles get created.HBase will try to combine HFiles to reduce the maximum number of disk seeks needed for a read. This process is called compaction. Compactions choose some files from a single store in a region and combine them. This process involves reading KeyValues in the input files and writing out any KeyValues that are not deleted, are inside of the time to live (TTL), and don’t violate the number of versions. The newly created combined file then replaces the input files in the region.

**Two Types of Comaction**

**Minor :** A minor compaction folds HFiles together, creating a larger HFile from multiple smaller HFiles

**Major:** When a compaction operates over all HFiles in a column family in a given region, it’s called a major compaction. Upon completion of a major compaction, all HFiles in the column family are merged into a single file

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

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

**Answer:** Every interaction in Hbase is going to start with rowkey only as they are treated as primary key. If there is no rowkey we are not able to distinguish between data. Rowkey is the first coordinate to locate the piece of data within the cell.

It enables the application to define the desired sort order. It also allows logical grouping of cells and make sure that all cells with the same rowkey are co-located on the same server.

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

**Answer:** When reading data from HBase using Get or Scan operations, you can use custom filters to return a subset of results to the client. Filters are generally used using the Java API, but can be used from HBase Shell for testing and debugging purposes.

**Benefits:**

a) Filters are used to get specific data form a Hbase table rather than all the records.

b) Filter is a space-efficient mechanism to test whether a StoreFile contains a specific row or row-column cell. Without Bloom Filter, the only way to decide if a row key is contained in a StoreFile is to check the StoreFile's block index, which stores the start row key of each block in the StoreFile.

**Q 6. What are the data model operations in hBase?**

**Answer:**

Get:- returns attributes for a specified row

Put: - either adds new rows to a table (if the key is new) or can update existing rows (if the key already exists).

Delete :- removes a row from a table

Scan: - allow iteration over multiple rows for specified attributes.

Increment

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

**Answer:** Apache MapReduce is a software framework used to analyze large amounts of data, and is the framework used most often with Apache Hadoop. HBase can be used as a data source, TableInputFormat, and data sink, TableOutputFormat or MultiTableOutputFormat, for MapReduce jobs. Writing MapReduce jobs that read or write HBase, it is advisable to subclass TableMapper and/or TableReducer.

**Q 8. What is regionserver?**

**Answer:** It is responsible for serving and managing regions. In a distributed cluster, a RegionServer runs on a DataNode.