Session17 Assignment 2- answers

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

MemStore – is a buffer that holds in-memory modifications (till it is flushed to store files) to the Store. Modifications are [KeyValue](https://hbase.apache.org/0.94/apidocs/org/apache/hadoop/hbase/KeyValue.html)s. Memstore is kept in Region Server main memory.

H files  – the file which is created when MemStore fills up, and due to Memtore flush, are called Hfiles. when certain thresholds are met (obviously, main memory is well-limited) Memstore data gets flushed into HFile. HFiles are written to HDFS.

1. Describe compactions in HBase.

Compaction, the process by which HBase cleans up after itself, A compaction asynchronously reads two or more existing HFiles and rewrites the data into a single new HFile. The source HFiles are then deleted. There are 2 types:

Minor compactions:

* Bigger Hfile are created by combining smaller Hfiles.
* Hfile keeps the deleted file with them.
* Increases space in memory, useful to store more data.
* Merge sorting is used in process.

Major Compactions:

* Data present per column family in one region is accumulated to 1 Hfile.
* During this process, all deleted files or expired cells are deleted permanently
* Increase read performance of newly created Hfile.
* Accepts lots of I/O.
* Possibilities for traffic congestion.
* The Major compaction process is also known as Write Amplification Process.
* This process must be scheduled at a minimum bandwidth of network I/O.

1. List and explain the logical entities in HBase.

|  |  |
| --- | --- |
| Entity | Description |
| Tables | Data is organized as tables consistingof rows. |
| Row Key | An array of bytes to form a unique key to reference each row. Rows are sorted lexicographically by Row Keys. For Example: Keys 1,2,3,10,15 will get sorted as 1,10,15,2,3. |
| Column Family | Columnsare grouped into families. |
| Column Qualifier | Used to identify column. |
| Cell | Contains a value = Table+ Row Key+Family+Column+Timestamp. |

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

**Row-key**-Row-key in hbase is mandator field which serves as the unique identifier for every record.

If it is not mentioned explicitly while loading any data file, the first column is considered as rowKey.

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

HBase can query data very quickly on demand but specific use cases may require to only return a subset of the scan results. Instead of scanning the entire dataset only to return a subset to the client, we can use Filters to get the data closer to what we need in less amount of time.

Thus, HBase has a set of predefined Filters as well as custom filters that we can use to scan and get filtered results from the HBase database.

There are two prominent ways to read data from HBase.

* **Get** is simply a Scan limited by the API to one row.
* **Scan** fetches zero or more rows of a table. By default, a Scan reads the entire table from start to end. We can limit our Scan results in several different ways, which affect the Scan’s load in terms of IO, network, or both, as well as processing load on the client side.

1. What are the data model operations in hBase?

* [**Get**](http://hbase.apache.org/apidocs/org/apache/hadoop/hbase/client/Get.html) returns attributes for a specified row.
* [**Put**](http://hbase.apache.org/apidocs/org/apache/hadoop/hbase/client/Put.html) either adds new rows to a table (if the key is new) or can update existing rows (if the key already exists).
* [**Scan**](http://hbase.apache.org/apidocs/org/apache/hadoop/hbase/client/Scan.html) allow iteration over multiple rows for specified attributes.
* [**Delete**](http://hbase.apache.org/apidocs/org/apache/hadoop/hbase/client/Delete.html) removes a row from a table.

1. How can MapReduce be used with HBase?

MapReduce can be used with HBase as a data source and a sink (the destination for the output):

* HBase provides a TableInputFormat, to which you provided a table scan, that splits the rows resulting from the table scan into the regions in which those rows reside.
* The map process is passed an ImmutableBytesWritable that contains the row key for a row and a Result that contains the columns for that row.
* The map process outputs its key/value pair based on its business logic in whatever form makes sense to your application.
* The reduce process builds its results but emits the row key as an ImmutableBytesWritable and a Put command to store the results back to HBase.
* Finally, the results are stored in HBase by the HBase MapReduce infrastructure. (You do not need to execute the Put commands.)

8. What is regionserver?

Region Server - HBase Tables are divided horizontally by row key range into “Regions.” A region contains all rows in the table between the region’s start key and end key. Regions are assigned to the nodes in the cluster, called “Region Servers,” and these serve data for reads and writes. A region server can serve about 1,000 regions.