Assignment 3.7

1. What is NoSQL data base?

NOSQL stands for Not Only SQL Database. It provides a DBMS system that does not use the conventional tabular relations used in RDBMS. It is an alternative to the RDBMS.

NoSQL is especially useful for use towards working with large datasets of distributed data. It is a complete different approach towards database design which can accommodate a wide variety of data model including columnar, key-value, graphs, documents.

It allows for great flexibility as well as operational performance.

2. How does data get stored in NoSQl database?

Different NoSQL databases use different ways to store data. There are mainly 4 types of NoSQL databases:

a. **Key-Value** : Leverages hashtable key-values pairs to store data. That means it stores every value as a key and value related to that key
b. **Column based** : Each store block contains data from one column.
c. **Document based** : Documents made up of tagged elements.
d. **Graph based** : Network based, edges and nodes to represent and store data.

3. What is a column family in HBase?

HBase uses column families as base storage mechanism.

A HBase table is made up of one or more column families, and each of those column families are stored in separate region files but they do share a common key that relates them. All column members of a column family uses the same prefix.

Column families must be declared at the schema definition. The columns inside those can be arbitrarily added later on.

4. How many maximum number of columns can be added to HBase table?

Usually 3 column families are recommended, but there is no such hard limit to the number of columns, they can be as many as they could.

5. Why columns are not defined at the time of table creation in HBase?

HBase uses the logical and physical distribution of column families and each column family can have one or more columns. Columns in one family is kept separate from column of another family.

HBase uses Column families which must be defined at the table creation, however there can be one or more columns in each column family.

6. How does data get managed in HBase?

a. WAL : Write ahead log is a file on the distributed file system. The WAL keeps in it's store the data until it is written to the permanent storage. In case the system fails to write the data to permanent storage, WAL will be able to recover it.

b. BlockCache : It is the cache memory used to read data in memory. It runs cleanup of old data which is not necessary at the moment.

c. MemStore : It is the cache memory that is used to write data. There is one MemStore per column family per region.

d. HFiles : It stores the rows as sorted key-value on the disk.

7. What happens internally when new data gets inserted into HBase table?

The client fetches the information on it's Region Server that hosts the meta table via zookeeper.

It will query the meta server to get the region server corresponding to the row key that it needs to access which will be cached by the client along with the meta table location. It will get the row from the given region server.

The client uses the cache to retrieve meta location and previously read row key and over time it does not need to query meta table repetitively unless there is some region move.