Bigdata Assignment 3.7

1. What is NoSQL data base?

Ans - NOSQL stands for Not Only SQL Database. They are nonrelational databases optimized for scalable performance and schemaless data models.It provides a DBMS system that does not use the conventional tabular relations used in RDBMS instead they use a variety of data models, including columnar, document, graph, and in-memory key-value stores.

These databases are a great fit for many big data, mobile, and web applications that require greater scale and higher responsiveness than traditional relational databases. Due to simpler data structures and horizontal scaling, NoSQL databases typically respond faster and are easier to scale than relational databases.

2. **How does data get stored in NoSQl database?**

Ans - There are of 4 types in which data gets stored :-

a) Key-value store NoSQL database - Key-value stores are the simplest NoSQL data stores to use. The client can either get the value for the key, assign a value for a key or delete a key from the data store. The key-value database uses a hash table to store unique keys and pointers (in some databases it’s also called the inverted index) with respect to each data value it stores. Key-value databases give great performance and can be very easily scaled as per business needs. Examples are MUMPS , Oracle No SQL , Zookeeper, etc.

b) **Document store NoSQL database** – It is similar to key-value databases in that there’s a key and a value. Data is stored as a value. Its associated key is the unique identifier for that value. The difference is that, in a document database, the value contains structured or semi-structured data. This structured/semi-structured value is referred to as a document and can be in XML, JSON or BSON format. Examples are RangoDB , MongoDB , etc

c) **Column store NoSQL database -** Here, data is stored in cells grouped in columns of data rather than as rows of data. Columns are logically grouped into column families. Column families can contain a virtually unlimited number of columns that can be created at runtime or while defining the schema. Read and write is done using columns rather than rows. Examples are Cassandra , Hbase , etc.

d) **Graph base NoSQL database –** Thses databases are basically built upon the Entity – Attribute – Value model. Entities are also known as nodes, which have properties. It is a very flexible way to describe how data relates to other data. Nodes store data about each entity in the database, relationships describe a relationship between nodes, and a property is simply the node on the opposite end of the relationship. Examples are Apache Giraph , OrientDB , etc.

3. **What is a column family in HBase?**

Ans - 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?**

Ans - Usually 3 column families are recommended. Because flushing and compactions are done on a per Region basis so if one column family is carrying the bulk of the data bringing on flushes, the adjacent families will also be flushed though the amount of data they carry is small. When many column families the flushing and compaction interaction can make for a bunch of needless i/o loading.

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

Ans - 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?**

Ans -
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.

BlockCache : It is the cache memory used to read data in memory.. It stores frequently read data in memory. Least Recently Used data is evicted when full.

MemStore : It is the cache memory that is used to write data. It stores new data which has not yet been written to disk. It is sorted before writing to disk. There is one MemStore per column family per region.

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

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

Ans - From the reigon server client fetches the information that hosts the metatable via zookeeper. It will query the meta server from there it will go to reigon server corresponding to th row key . It 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