Problem Statement

1. What is NoSQL data base?

Ans : A NoSQL database environment is, simply put, a non-relational and largely distributed database system that enables rapid, ad-hoc organization and analysis of extremely high-volume, disparate data types. NoSQL databases are sometimes referred to as cloud databases, non-relational databases, Big Data databases and a myriad of other terms and were developed in response to the sheer volume of data being generated, stored and analyzed by modern users (user-generated data) and their applications (machine-generated data).

2. How does data get stored in NoSQl database?

Ans: Graph stores are used to store information about networks of data, such as social connections. Graph stores include Neo4J and Giraph. Key-value stores are the simplest NoSQL databases. Every single item in the database is stored as an attribute name (or 'key'), together with its value.

3. What is a column family in HBase?

Ans. HBase is a column-oriented database and the tables in it are sorted by row. The table schema defines only column families, which are the key value pairs. A table have multiple column families and each column family can have any number of columns. A HBase table is comprised of one or more column families, each of which is stored in a separate set of regionfiles sharing a common key.

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

Ans. There is really no limit.

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

Ans. Column families need to be defined at the time of table creation. But there is no need to define columns since they can be created on the fly.

6. How does data get managed in HBase?

Data in Hbase is organized into tables. Any characters that are legal in file paths are used to name tables. Tables are further organized into rows that store data. Each row is identified by a unique row key which does not belong to any data type but is stored as a bytearray. Column families are further used to group data in rows. Column families define the physical structure of data so they are defined upfront and their modification is difficult. Each row in a table has same column families. Data in a column family is addressed using a column qualifier. It is not necessary to specify column qualifiers in advance and there is no consistency requirement between rows. No data types are specified for column qualifiers, as such they are just stored as bytearrays. A unique combination of row key, column family and column qualifier forms a cell. Data contained in a cell is referred to as cell value. There is no concept of data type when referring to cell values and they are stored as bytearrays. Versioning happens to cell values using a timestamp of when the cell was written.

Tables in Hbase have several properties that need to be understood for one to come up with an effective data model. Indexing and sorting only happens on the row key. The concept of data types is absent and everything is stored as bytearray. Only row level atomicity is enforced so multi row transactions are not supported.

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

Ans . While inserting data user will mention column family and col name , unique key and value. Habse inserts the timestamp that's how it can handle multiple versions. Hbase provide very fast key access . Records are sorted in key based.

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.

Region assignment, DDL (create, delete tables) operations are handled by the HBase Master. A master is responsible for: • Coordinating the region servers • Assigning regions on startup • Re-assigning regions for recovery or load balancing • Monitoring all RegionServer instances in the cluster (listens for notifications from zookeeper)