1. **What is NoSQL data base?**

NoSQL stands for Not Only SQL,or having better capabilities than SQL.

It is an alternative to the Structured Query Language (SQL)  .

They don’t have SQL like constructs. They are schema-less don’t use the relational model and run well on clusters

They provide lower-level API and meant for automatic scaling-out.

It is an approach to databases , away from traditional relational database management systems (RDBMS).

It is particularly useful for storing unstructured data, which does not fit the relational schemas of RDBMS.

Common types of unstructured data include: user and session data; chat, messaging, and log data; time series data such as IoT and device data; and large objects such as video and images.

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

There are 4 main NoSQL models designed to store data:

* Key-value stores, also known as hashmaps. Eg:Cassandra.  Every single item in the database is stored as an attribute name (or 'key'), together with its value
* Document store, which stores a big set of bytes under a key.
* Pair each key with a complex data structure known as a document. Documents can contain many different key-value pairs, or key-array pairs, or even nested documents.EG:Mongo DB.
* Wide Column oriented, (Cassandra and HBase) are optimized for queries over large datasets, and store columns of data together, instead of rows.The partitioning of data is simple.

The columns can be stored on different computers, so the queries can be done all in parallel and faster.

* Graph database store information about networks of data, such as social connections. Graph stores include Neo4J and Giraph.

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

Column families are physical and logical grouping of columns**.**

HBase tables are organized by column, rather than by row. The columns are organized in groups called *column families*.

When creating a HBase table,we must define the column families before inserting any data,but we can defind the column later while the table is up and running.

All column members of a column family have the same prefix. For example, the columns subject:english and subject:kannada are both members of the courses column family.

The colon character (:) delimits the column family from the column family qualifier.The column family prefix must be composed of printable characters. The qualifying tail, the column family qualifier, can be made of any arbitrary bytes.

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

There is no limit as such.

But there might be limitations to consider such as

* Lock Granularity (When you do an operation within a row, the RegionServer code briefly holds a lock on that row while applying the changes and the lock is held until the changes are complete)
* Region Distribution (The region servers handles load balanciong an d distribution. A hot row will never be split over 2 region servers)

There is no specific limit on the number of columns in a column family. Actually you can have millions of columns in the single column family.

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

Column families define the physical structure of data so they are defined upfront and their modification is difficult.

Physically, all column family members are stored together on the filesystem. Because tunings and storage specifications are done at the column family level, it is advised that all column family members have the same general access pattern and size characteristics.

1. **How does data get managed in HBase?**

* The Data Model in HBase is designed to accommodate semi-structured data that could vary in field size, data type and columns.
* The layout of the data model makes it easier to partition the data and distribute it across the cluster.
* The Data is managed in HBase is made of different logical components such as Tables, Rows, Column Families, Columns, Cells and Versions.
* HBase can manage structured and semi-structured data with its built-in features such as scalability, versioning, compression and Hadoop’s MapReduce capabilities.
* Since its uses write-ahead logging and distributed configuration, it can provide fault-tolerance and quick recovery from individual server failures.

**HBase Architecture**

The HBase Physical Architecture consists of servers in a Master-Slave relationship as shown below. Typically, the HBase cluster has one Master node, called HMaster and multiple Region Servers called HRegionServer. Each Region Server contains multiple Regions – HRegions.

Just like in a Relational Database, data in HBase is stored in Tables and these Tables are stored in Regions. When a Table becomes too big, the Table is partitioned into multiple Regions. These Regions are assigned to Region Servers across the cluster.

Each Region Server contains a Write-Ahead Log (called HLog) and multiple Regions. Each Region in turn is made up of a MemStore and multiple StoreFiles (HFile). The data lives in these StoreFiles in the form of Column Families (explained below). The MemStore holds in-memory modifications to the Store (data).

The mapping of Regions to Region Server is kept in a system table called .META. When trying to read or write data from HBase, the clients read the required Region information from the .META table and directly communicate with the appropriate Region Server. Each Region is identified by the start key (inclusive) and the end key (exclusive)

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

When the client gives a command to Write or insert new data or by import of any kind, the following steps occur:

* It is directed to Write Ahead Log and first to important logs .This is for the purpose of Fault Tolearnce so that if any error occurs while writing data, HBase always has WAL to look into.
* Once the log entry is done, the data to be written is forwarded to MemStore which is actually the RAM of the data node. All the data is written in MemStore which is faster than RDBMS (Relational databases).
* Later, the data is dumped in HFile, where the actual data is stored in HDFS. If the MemCache is full, the data is stored in HFile directly.
* Once writing data is completed, ACK (Acknowledgement) is sent to client as a confirmation of task completed.