Unit 2

RDBMS vs No-SQL

Relational Database	No-SQL Database
(1) Relational database supports a powerful query language	① NOSQL database supports a very simple query language
② Relational Database has a fixed schema	② No-SQL Database has no fixed schema
③ Relational database follows acid properties (Atomicity, Consistency, Isolation & Durablity)	③ NO-SQL database is only eventually consistent.
④ Relational database supports transactions (also complex transactions with joins)	④ NOSQL database do not support transactions (support only simple transactions.
⑤ RDMS manages only Structured data	⑤ NO-SQL database can manage Structured Unstructured & Semistructured data
⑥ Relational databases has Centralized structure	● NO-SQL databases can handle big data or data in a very high volume
⑦ Relational Database are used to handle moderate volume of data	⑥ NO SQL database has decentralized structure
⑧ Relational Database have a single point of failure with fail over	⑦ No-SQL database have no single point of failure

Types Of No-SQL Database (Architecture pattern

① Key Value Stores :- The main idea is using a hash table where there is a unique key & a pointer to a particular item of data.

But it is inefficient when you are only interested in querying or updating part of a value, among other disadvantages. Key-value pair storage database stores data as a hash table where each key is unique & the value can be a JSON, BLOB (binary large objects) string etc.

Eg

Key	Value
Name →	Preeti
Birthday →	21/12/2001

(2) Column family stores :- These were created to store and process very large amounts of data distributed over many machines there are still keys but they point to multiple columns. The columns are arranged by column family.

Eg Cassandra, HBase, Hypertable
these database are mainly used to manage data warehouse business intelligence, CRM, library card catalogs.

(3) Document Database :- These were inspired by lotus notes & are similar to key value stores. The model is basically versioned documents that are collection of other key-value collections. The semi-structured document are stored in formats like JSON.
Document database support querying more efficiently.
Eg Couch DB, Mongo DB, Amazon, Simple DB, Riak lotus notes.

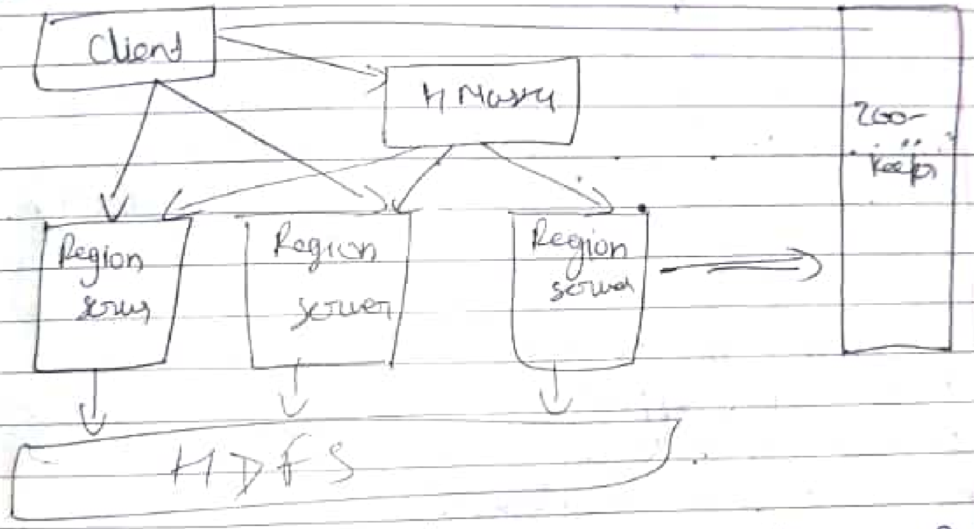
(4) Graph Databases :- Instead of tables of rows & columns & the rigid structure of SQL, a flexible graph model is used which again can scale across multiple machine. Graph base databases mostly used for social network, logistics spatial data.

Eg Neo4J , Infogrid , Infinite Graph etc.

(5) Multi model Database :- These are designed to handle multiple data models against a single integrated backend. they are a brand-new in the No-SQL world, there will be much more buzz around this type of database in the future.

eg OrientDB is a multi-model database, combining No-SQL types, Orient DB is graph database where each node is a document.

HBase Architecture



H-Base Architecture has 3 Components, HMaster, Region Server, Zookeeper.

(1) HMaster ⇒ the implementation of master server in HBase is HMaster. It is a process in which regions are assigned to region server as well as DDL (create, delete table) operations It monitor all Region Server instances present in the cluster.

It has many features like controlling, load balancing, failover etc.

(2) Region Server :- HBase Tables are divide horizontally by row key range into Regions. Regions are the basic building elements of HBase Cluster that consists of the distribution of tables & are composed of Colowmn families. Region Server runs on HDFS data Node which is present in Hadoop Cluster. The default size of Region is 256 MB.

(3) Zookeeper :- It is like a coordinator in HBase.
It provides services like maintaining configuration information, naming, providing distributed synchronization, server failure notification etc.
Clients Communication with region servers via Zookeeper.

Advantage of HBase

1 Can store large data sets
2 Database can be shared.
3 cost-effective from gigabytes to petabytes.
4 High availability through failover & replication.

Disadvantages of HBase

1 No Support SQL structure
2 No transaction support
3 sorted only on key.
4 memory issues on the Cluster.

CAP Theorem . (also called Brewer's Theorem)

Network partition means there is a break in the network. or there is no connection b/w the data stores so the senario where there is a break b/w different nodes in a distributed data stores is called partition tolerance. It is nothing but no data replication for a particular node or maybe group of nodes or we can call it isolated data so that way database or data store is isolated from other databases in the distributed system because these are partition tolerance.

- Consistency :- means all Clients see the same data at the same time no matter which node they connect to for this is to happen whenever data is written to one node it must be instantly provided or replicated to all the other node in the system before the write is derived successfully.
- Availability :- means that any client making a sequence for data gets a response even if one or more node are down another way to state this all working nodes in the distributed system return a valid response for any request without any exception.
- Partition Tolerance :- Acc. to Cap theorem partition tolerance is must so when a network fails we have to choose Consistency or availability as a distributed database System is bound to have partition in a real world system due to network failure or some other reason therefore partition tolerance is a property we can not avoid while building our system. So a distributed system will either choose to give up consistency or avalablity but not partition toleranc

Eg In a distributed system if a partition occurs b/w 2 nodes It is impossible to provide Consistant data on both the nodes & availablity of complet data therefore in such

a scanario we either choose to Compromise on Consisten
or on availability.

Distributed database is either Characterised as:-

CP or AP

HDFS (Hadoop Distributed file system)

It is used for storing massive data

The system Contains memory store & hfiles, memory st
is fast like the cache memory anything that is entered
into the h base is stored here initially later the data
is transfered & saved in hfiles as blocks & the memstor
is flushed.

HDFS - It is the Component of Hadoop which is respo
for storing the data it has 2 Components.

(1) Name code :- as data is segmented into various block
so name node stores the meta data of the various blocks
the info which this meta data holds is the name of the
blocks, size of the block, location of the block & th
Original file.

(2) Datanode :- It holds the actual data i.e. block

In map Reduce layer - it has 2 components

(1) Job Tracker :- Divides the Computer into various task. It assighs
these tasks to the various task trackers. Once the Computati
is performed by the various task trackers. The result is
accumulated & Output is generated by the job tracker,
it also takes the feeback from the various task trackers
as well as it monitors the status of all these task
trackers.

(3) Task Tracker :- It performs the actual Computations.

A features of Hadoop

(1) fault tolrant : as the data is stored in various nodes of the cluster & a replicated copy of the original block is also stored somewhere in the clusture so if a particular system crashes then the data is not lost.

(2) Scalable :- It means that no of system can be increased or decreased on the basis of requirement of the application.

(3) low Cost :- As hadoop is open source i.e it is freely available & it is build on low cost comodity hardware which makes the entire cost very low.

(4) Can store unstructured data.

(5) faster Computation :- as hadoop distributes processing among the various node in the clusters i.e why it performs faster computation.