Chapter 8

Structured and unstructured data:

Structured data: Stored in tabular format, clearly defined. For example excel files, SQL Database etc. the rows and columns are related to each other. DBMS is needed for managing data.

Unstructured data: No pre-defined structured, no data model, data is irregular and ambiguous. Easiest to extract data. 80-90% of data is unstructured. It is a combination of text, numbers, audio, video, images, messages, social media etc. unstructured data is the most useful kind of data. It provides a lot of information.

No Sql: A NoSql database provides mechanism for storage and retrieval of data that is modelled in means other than tabular relations used in relational database.

Types of databases in NoSql: There are four types of databases in NoSql

a. Key values database
b. Wide column database
c. Document database
d. Graph database.

A key-value database is a type of non-relational database, that uses a simple key-value method to store data. Here key serves as a unique identifier. For example:

Sid	Name	Address	Phon.	Hobby
1	Shyam		786898876	Tennis
2	Sita	Khajura		

Document database: Here databases are used in the form of document. Key value could be number, character or any symbols. For example: Information about movies, bank transaction.

Wide column database: Column database stores data in the form of column. It focuses on column unlike relation database which focus on rows.

Graph database: When a data is highly inter-connected than graph database is used. For example: social media graph database.

Advantages of NoSql:

1. High performance.
2. Flexibility. With SQL databases, data is stored in a much more rigid, predefined structure.
3. Availability.

4. High functional.
5. Scalability. Instead of scaling up by adding more servers, NoSQL databases can scale out by using commodity hardware.

HBase Architecture: HBase has mainly three different parts. These are HMaster, RegionServer and the ZooKeeper.

HMaster: HMaster in HBase is a process which helps to assign the region servers. It balances the loads by assigning the regions.

- Hmaster manages the Hadoop clusters.
- Helps to create, modify and deletes tables in the database.
- It also cares about different tasks when the client wants to change the schema or metadata.

Region server: The region servers are the main working nodes. It handles the read, write, modify requests from the clients. The region server runs on every node in the Hadoop cluster.

- It has a read cache called Block cache, read data are stored in the read cache and when the cache is full, recently used data is removed.
- Another cache is present here called MemStore. It is write cache.
- It has the actual storage file called HFile. It stores the actual data in the disk.

ZooKeeper: This is an open-source server which enables the reliable distributed cordination. ZooKeeper is a centralized service that maintains the distributed synchronization. It keeps track on all regions servers in HBase.

MongoDb: MongoDB is a document-oriented, no sequel (NoSQL) database.

Why MongoDB is used:

1. Flexibility.
2. Flexible query model.
3. Native aggregation.
4. Schema-less model

Characteristics of MongoDB:

1. General purpose database.
2. Flexible schema design.
3. Scalability and load balancing
4. Aggregation framework.
5. Native replication
6. Security features

Working of MongoDB:

Documents store data with the help of key-value pairs. A collection is a group of documents then these collections are stored in the MongoDB database.

Application of MongoDB:

1. Internet of things
2. Mobile applications
3. Real time analysis.
4. Personalization
5. Catalog management.
6. Content management.