14/2/23

Unit-III

Cassandra

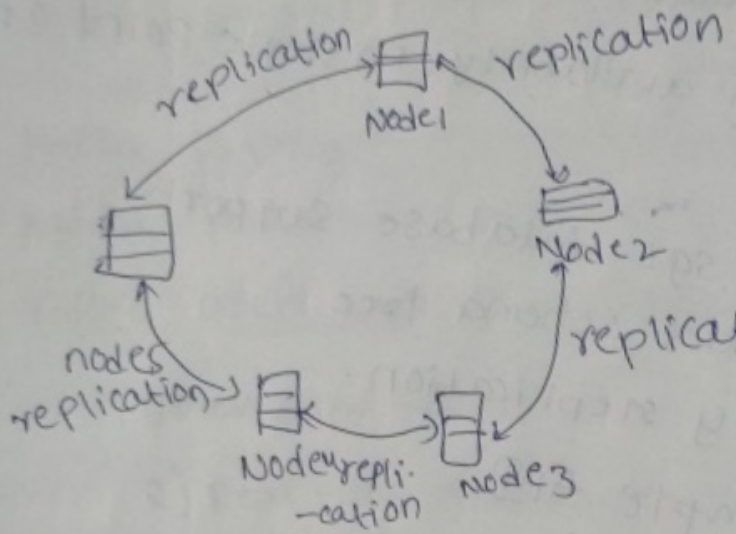
* Cassandra is a distributed database, highly scalable, designed to manage large volume of structured data.
* provides high availability, no single point of failure.
* It is a NoSQL database support.
  - provides schema free
    - → easy replication
    - → simple API
    - → consistent.
* simplicity
* horizontal scaling
* It is scalable, fault tolerant
* column oriented database
* created by facebook.
* used by fb, twitter, cisco, Rickspace

Features

* Elastic scalability
* Always on Architecture
* Flexible data storage
* Easy data distribution
* Transaction support.

Goal :- To handle bigdata work load across multiple nodes

- All nodes in the cluster plays same role
- Each node in the cluster perform read and write request



* if the node reponted with a out of data value. cassandra, will return the most recent value to the client.

* constand. using Gossip protocal for detect error nodes.

* After returning the most resned value cassendra performs a need repair in the background to update the state value.

(create crecent)

* Components of cassendra :-

1. Nodes
2. Data centre

3. cluster
4. commit log
5. mem table
6. SS table
7. Bloom filter.

Components of cassandra.

1. Nodes - It is the place where the data is stored.
2. Data center – collection of related nodes.
3. cluster - contains one or more data center.
4. commit log - is a crash recovery mechanism
5. mem table - is a memory resident data structure
6. SS table - it is a disk file to which the data is flushed from mem table.
7. Bloom filter - is a algorithm for testing whether an element is a member of a set or not.

Cassandra Query Language (CQL)

* CQL treats the database as a container of tables
* programmers works with cqlsh as a prompt to work.

write operation

* Every write activity is captured through the commit log.
* Later the data will be captured and stored in the memtable.
* whenever the memtable is full the data is written into ss table datafile.

Read operation

* During read consstand gets value from memtable and checks the bloom better. to find appropriate sstable that hold the actual data.

supports the following data.

* Time series data.
* marketing data.
* Financial data
* IOT data
* graph data.

Statistics about facebook

mysql >50 GB data.
write average ~ 300ms
Read average ~ 350ms

20/2/23

Mongo DB:

* Mongo DB is a cross platform document oriented db that provides high performance, high availability and easy synchronization and scaling which works on collection of documents.

Features

* It is an intelligent operational db platform
* It belongs to a aggregation frame work.
* It efficiently Map reduces the data.
* Converts table into Json document, then further converts it into Bson.doc.
* Allows Sharding
* Schema less
* Capped Collection.
* Stores data in cloud efficiently.

Advantages

* schema less
* No complex joints
* Easy to scale out
* Deep Querying ability.

Where MongoDB is used?

* Big data
* Content Management and delivery.

* Mobile and Social infrastructure * Data hub

* Data Models in Mongo DB:

1) Embedded data model:

2) Normal data model.

Data types in Mongo DB:

* String * Integer * Boolean * null * Symbol,

* Binary * Regular expression.

15/2/23

Spark

* Spark is a opensource cluster computing framework, purpose is to handle real time generated data.
* started in 2009 in US Berkeley RandD Lab 2010 - Open source, 2013 - Apache Foundation adopted.

Features

* used to perform batch processing, stream processing - using Apache storm/s4.
* used for interactive processing (use of Neo4j apache graph).
* Fast - use DAG scheduler a query optimizer physical extution engine.
* Easy touse - write application in Java scala python, R, and SQL provides high level operations
* Generality - provides collection of libraries.
* Light weight - it is a unified analytics engine used for large scale data processing

* Runs everywhere - run on hadoop Apache Menos Kuberhetics, standalone and cloud.

Usage of Spark.

1. Data Integration - to fetch consistent data from System, use ETL (Extract, Transform, Load)

2. Stream processing - work with realtime generated data such as log files.

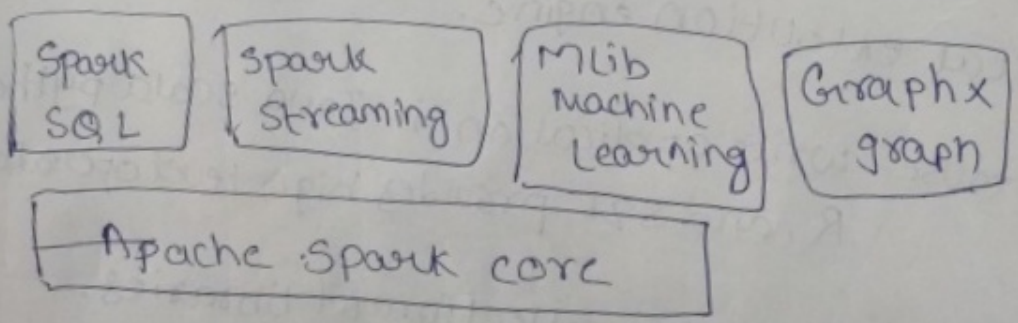
3. Machine learning.

4. Interactive analysis → handle the data interactively.

Why Spark

Support multiple languages, Realtime processing, fault tolerance in-memory processing, Rensability, cost efficient, Graph x.

Components of Spark



1. Spark core :- provides execution platform for all Spark applications.
- provides in-memory processing.

2. Spark SQL: provides new data abstraction called schema RDD which provides support for structured and semi structured data.

16/2/23

* Components of Spark

1. Spark core
2. Spark SQL
3. Spark Streaming
4. MLib
5. Graph.

Resillient Distributed Data set (RDD)

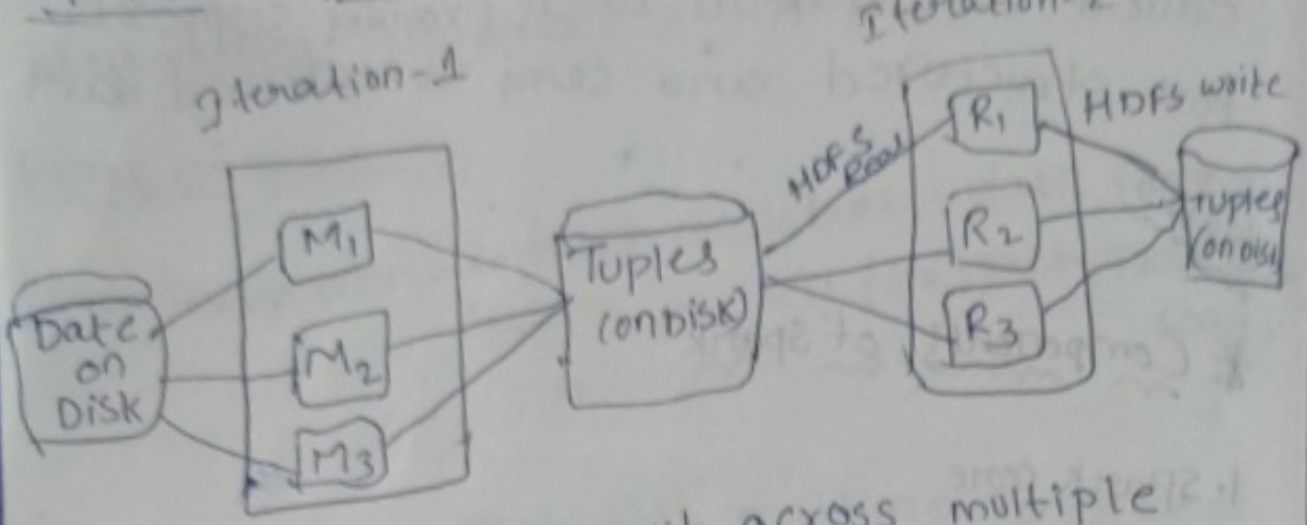
* It is a fundamental unit of data in Spark to perform parallel operation
* It is immutable, Readonly, partitioned collection of records.

Two way to create RDD

1. parallising an existing collection in your driver program
2. Referencing dataset in an external storage system such as HDFS, Hbase.

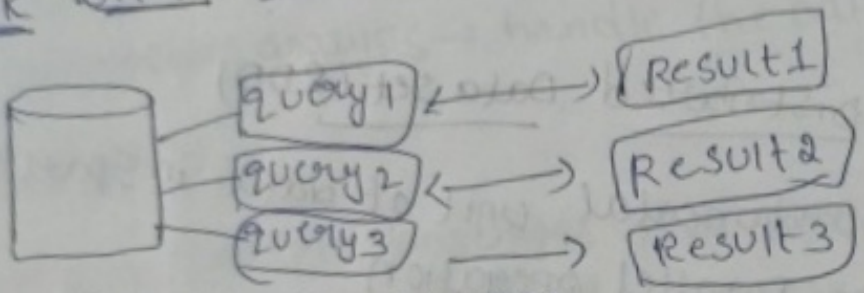
* Both Iterative and Interactive application requires faster data sharing across parallel jobs.

Iterative operations on map reduce:



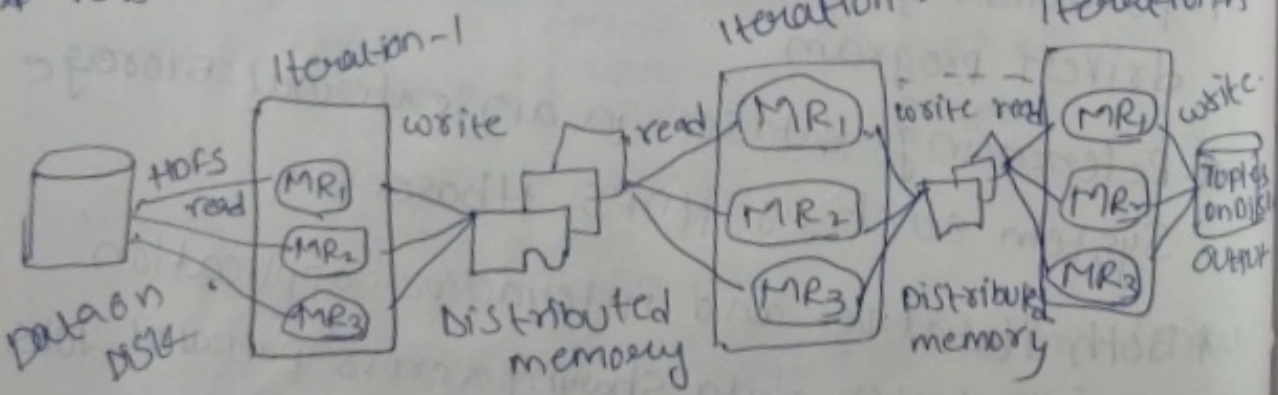
* Reuse intermediate result across multiple computation in multistage application.

Work with interactive queries:



Input from stable storage.

* In RDD supports in memory processing computation
* It stores the state of memory as an object across the jobs and object is shorable b/w
* jobs.



Q6. Data Visualization, It's importance, Tableau tools

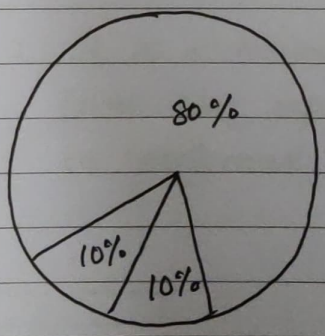
Data Visualization is the process of creating a graphical representation of data using various methods like pie charts, bar charts etc.

It is used to reveal the patterns and trends hiding under complex data

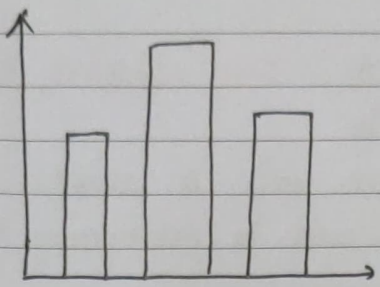
It leads to better and faster decision making

Different visualization Techniques include the following-

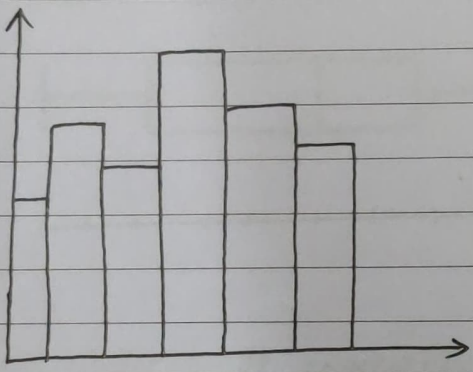
- Pie chart : It is divided into sections to represent the proportions of the total amount.



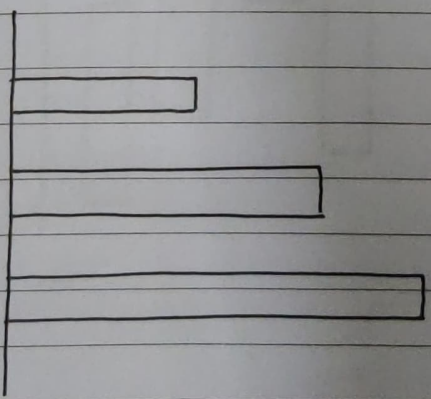
- Bar chart : It is used to represent different categories with the help of vertical bars



- **Histogram** : It is also divided into vertical bars representing how often something happens within a range.



- **Gantt chart** : It is a horizontal bar chart Each bar represents a task that is performed in a certain time

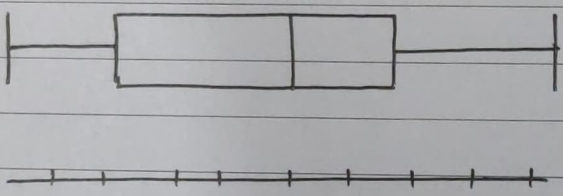


- **Heat map** : It uses different colours to show the contribution of each category.

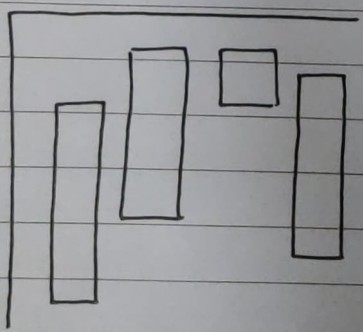
- **Box and whisker plot :** It is in the form of distribution of data

  It shows a five number summary that comprises of the following

  → minimum
  → first quartile (Q1)
  → median
  → third quartile (Q3)
  → maximum



- **Waterfall chart :** It represents the contribution or decrease in value with time in the form of vertical bars



Importance of data visualization →

- Using graphs and charts to visualize large amounts of data is more comfortable than studying spreadsheets and reports
- It helps to identify patterns and trends that are hidden in complex data
- It helps to make faster and better decisions
- It helps to understand the story in a single glance
- It helps us explore business insights
- It helps us identify errors in the data.

Tableau →

Tableau is a powerful visualization tool used to analyze raw data by presenting it in a visual manner.

No technical or programming skills is required to work with tableau

Tableau tools →

1. Tableau Desktop → establishes connectivity ~~with~~ btwn the data warehouse and other kinds of files

based on the connectivity, it is of 2 types
→ Tableau desktop personal
→ Tableau desktop professional

2. Tableau Public → Work created cannot be stored locally.
It is kept on cloud where it can be accessed and viewed by all.

3. Tableau Online → Data is stored in servers hosted on the cloud, which is maintained by the tableau group

4. Tableau Server → It can share the workbook visualization which is created by tableau desktop.

5. Tableau Reader → It allows us to view the visualisations, but we cannot edit or write to the data.