Unit 5:- Big data

Big data is defined as data that is huge in size. Big data is a term used to describe a collection of data that is huge in size & yet growing exponentially with time.

* Big data characteristics:

i) ~~velocity~~ volume
ii) ~~variety~~ velocity
iii) variety
iv) veracity
v) validity
vi) volatility

i) volume

It refers to amout of data (size of data). Today data size has increased to size of terabytes in the form of records or transactions.

ii) velocity

The velocity is the speed at which the data is created, stored & visualized. In the past when batch processing was comman practice, it was normal to receive an update from the database every night.

iii) variety

variety refers to the many sources & types of data. In the past, all the data ~~was~~ created was structured data, it neatly fitted in coloumms & rows.

Nowadays 90% of the data is generated by organizations is unstructured data.

Data comes in today in diffrent formats:

i) structured

ii) semi-structured

iii) unstructured.

iv) Complex structured

iv) veracity

veracity refers to the biases, noise, abnormalities, ambiguity & latency in data.

~~v) valid~~

* Data types.

i) structured data.

~~Ay~~ Any data that can be stered, accessed & processed in the form of fixed format is termed as structured data.

ii) semi-structured data.

Semi-structured data is information that doesn't reside in a relational database ~~thaat~~. but that does have some organizational properties that make it easier to analyze. With some process ~~to~~ you can store them in relation database.

Ex. XML, JSON documents.
NOSQL

iii) Unstructured data

Any data with unknown form or unknown structure is classified as unstructured data. It often includes text & multimedia content.

Ex. email, word-processing documents, videos, photos, audio-files, presentations, webpages

* Important 2v's of Big data.

i) Validity

The correct data & accurate data is intended to use for taking decisions.

ii) volatility

Big data volatility refers to how long is data valid & how long should it be stored.

Page No.
Date

In this world of realtime data you need to decide when the data is irrelevant to the current analysis.

* Distributed System

A distributed system contains multiple nodes that are physically seprate but linked using the network.

* Drawbacks of single system.

i) High chance of system feilure.
ii) Limited bandwidth.
iii) High complexity.
iv) High dependance on single system.
v) NOT Scalable.

* Hadoop → High avalibility distributed object oriented platform.

Hadoop is an open-source software framework for storing data & running applications on clusters of commodity hardware.

* Key characteristics of Hadoop.

i) system feilure..
ii) programming cmplexity.
iii) Bandwidth
iv) scalable
v) Stores multiple copies.
vi) Reliable

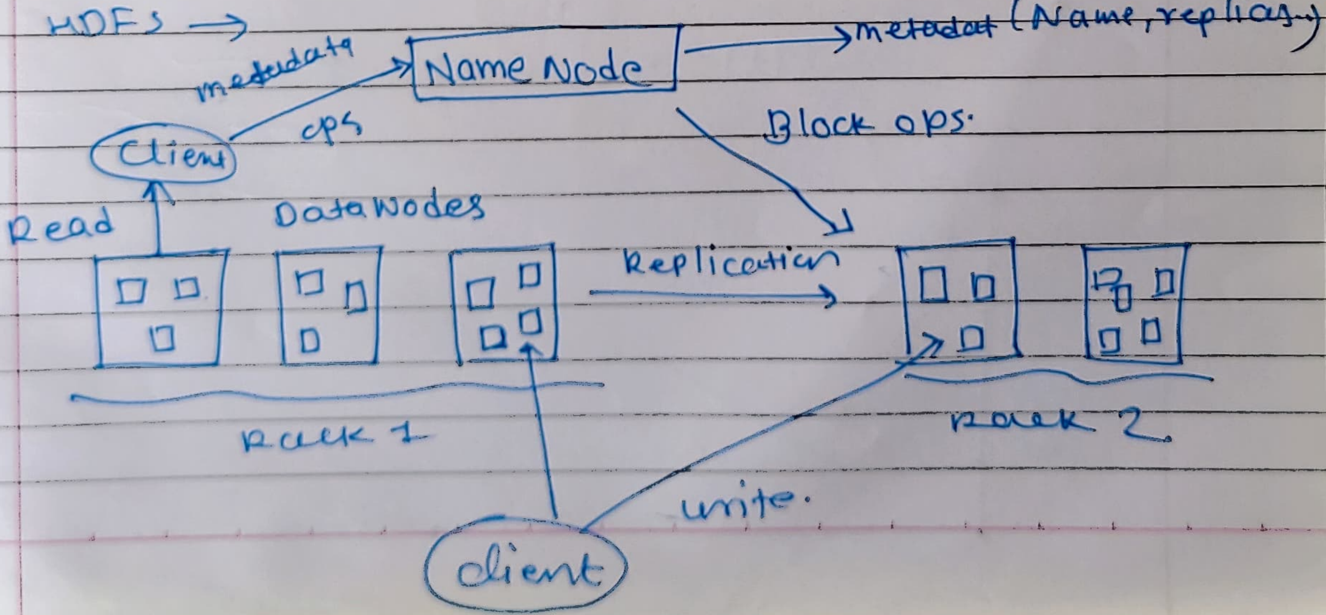
* Hadoop core components

map-reduce (data processing)
YARN (Yet another Resource Negoticiter)
HDFS

i) map-reduce

i/p data → mapper → shuffle + sort → reducer

HDFS →



* Name node

Keep track of all the files or datasets in HDFS.