Big Data Architecture

Module - 1

(1) [redacted]

Big Data: is a high-velocity and or high variety information asset that requires new forms of processing for enhanced decision making, insights discovery and process optimization.

Industry analysts described 3Vs, volume, variety and velocity and variability.

A collection of data sets so large or complex that traditional data processing applications are inadequate.

Big Data Characteristics: 4 Vs.

- Volume – It defines the amount or quantity of data which is generated from an application. Size determines the processing considerations for handling data.
- velocity – The term velocity refers to the speed of generation of data. It is the measure of how fast the data generates & processes.

- Variety: Big Data comprises of a variety of data. Data is generated from multiple sources in a system. This also increases the complexity as data arrives from different platforms in different formats.

- Veracity: It is the quality of data captured, which can vary greatly, affecting its accurate analysis.

(2) Grid Computing:

- It refers to distributed computing in which a group of computers from several locations are connected with each other to achieve a common task. The resources are heterogeneously and geographically dispersed.

A grid can be used for various purposes.

- A single grid can be used particularly for one application

- It provides large-scale resource sharing which is flexible, co-ordinated and secure among its users

- It depends on sharing of resources to attain coordination and coherence among resources.

Drawback of Grid Computing

Grid computing is the single point, which leads to failure in case of performance or failure of any of the participating node.
A system's storage capacity varies with the number of users instances and the amount of data transferred at a given time. Sharing resources among a large number of users helps in reducing infrastructure cost and raising load capacities.

(3) Data Architecture Design:

i.) identification of data sources
ii.) acquisition, ingestion & preprocessing
iii.) Data storage
iv.) Data processing
v.) Data consumption

(4) ~~Phases of Architecture~~

layer 5 Data Consumption	Export of dataset to cloud, web etc.	Dataset usage.	Analytics, real time, Scheduled batch
layer 4 Data Processing	Processing technology Map Reduce, Hive, Pig	Processing in real-time Scheduled batches or hybrid	synchronous or asynchronous processing
layer 3 Data Storage	Considerations of types, formats, compression, frequency of incoming data, patterns of querying & data consumption	Hadoop distributed System Spark	NoSQL data Stores Hbase Mongo
layer 2 Data Ingestion and acquisition	Ingestion using Extract load & Transform (ELT)	Data semantics	Pre-Processing or Requirement
			Ingestion of data from sources in batches or real time
layer 1 Identification of internal & external sources of data	Sources for ingestion of data	Push or pull of data for ingestion	Data types of DB Data formats (structured unstructured Semi

4. Phases in Analysis

1. Descriptive analytics enables deriving the additional value from visualisation and reports

2. Predictive analytics in advanced analytics which enables extraction of new facts and knowledge & then predicts / forecasts

3. Prescriptive analytics enable derivation of the additional value and undertake better decisions for new option(s) to maximize the profits

4. Cognitive analytics enables derivation of the additional value and undertake better decisions.

5. Big Data Classification:

Structured: Conforms and associate with data schema and data models. They are in the form of tables.

We can carry the following operations on table:

i) Insert, delete, update & append.
ii) Indexing, Allows scalability
iii) Data processing operations.

iv) Transaction processing by following ACID rules.
v) Data security.

Semi Structured data: contains tags or other markers. They do not conform and associate with formal data model structure.
eg: XML, JSON.

Multistructured: Consists of multiple formats of data such as structured, semi structured and/or unstructured data - they are found in non-transactional systems.
eg: data from multiple sensors.

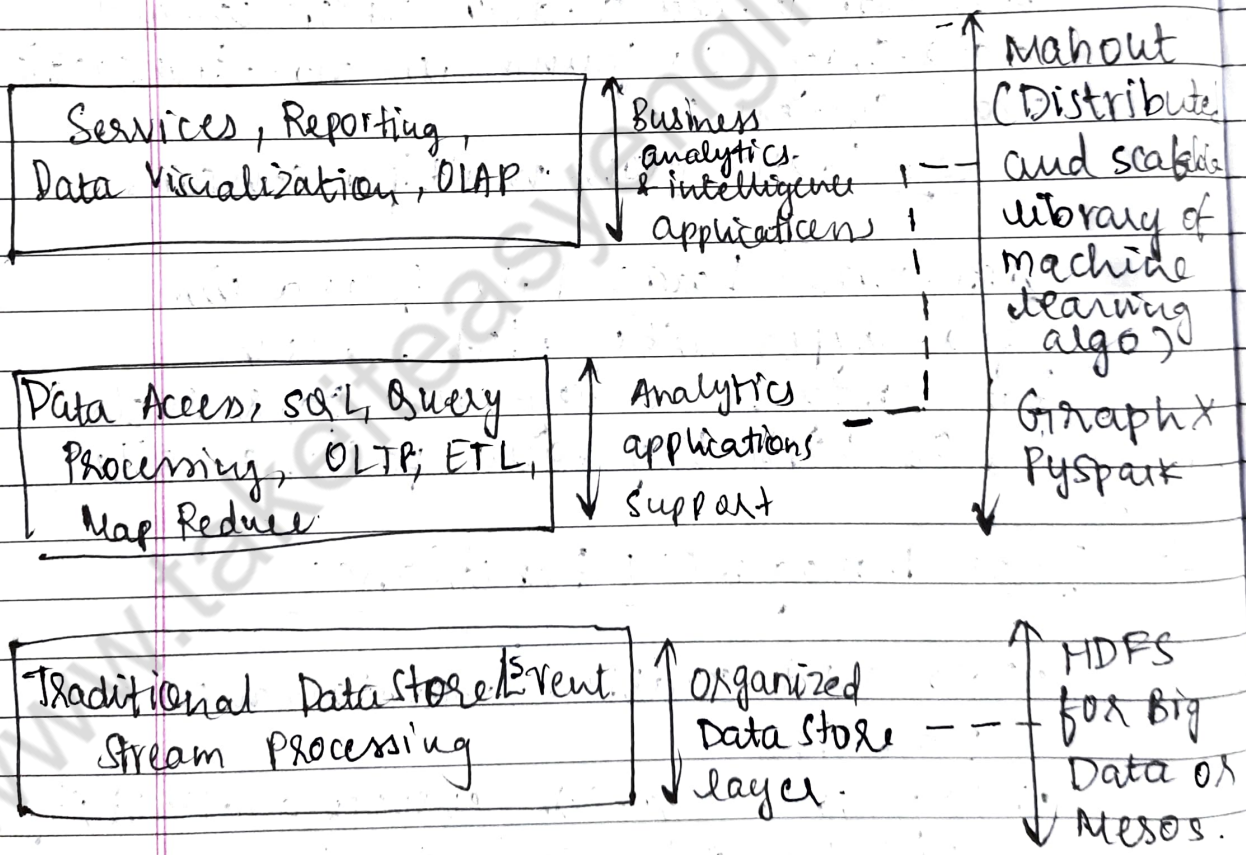
Unstructured Data: Data found in file types like .txt, .csv.
Data may be a key-value pair such as hash-key value pairs, email

(6) Data Management Functions:

- Data asset creation, maintainance & protection.
- Datagovernance which ensures availability, usability, integrity, security & high quality data.
- Data architecture creation, modelling and analysis
- Database maintainance, administration system

- Managing data security, data access control, deletion, privacy and security.
- Data collection using ETL processing
- Data application Integration
- Integrated Data Management.
- Maintainance of Business Intelligence.
- Data mining and analytics algorithm.

(7) ~~Data Store Export to cloud~~
Phases of Big Data Analytics.



(8) Big Data to Manage Credit Risk.

- Financial institutions in many countries face credit risk like;
  - loan defaults
  - Timely return of interest and principal amount.

- Financial institutions must get insights to the following
  - Identifying high credit rating business groups and individuals.
  - Identifying risk involved before lending money.
  - Identifying individuals sectors with greater risk.
  - Identifying types of employee with greater risk
  - Anticipating liquidity issues over the year.

- Big Data ~~And~~ Analytics monitors social media, interaction data, contact addresses, mobile numbers, websites, financial status, activities and job changes to find the credit risk that may affect a customer loan returning capacity.

- Digital Footprint provide alternate data source for risk management

- Friends on Facebook and their credit rating, comments posted also help in determining the risks

- Three benifits of Data analytics over Credit risk management
  → Minimize non-payment and fraud
  → Identify new credits, new customers, new credit opportunities thereby by building high credit rating customer base.
  → Marketing to low risk businesses and households.

(9) Big Data and Healthcare:

Data sources used by BDA:
i) Clinical records
ii) Pharmacy records
iii) Electronic medical records
iv) Diagnosis logs & notes
v) Additional Data like ads deviations from personal usual activities medical leaves from jobs, social interactions.

Health care analytics can facilitate the following.
- giving value based & customer centric healthcare
- using IoT for health care.
- Preventing, fraud, waste, abuse

- Improving outcomes
- monitoring patients in real time.

Value based and customer centric health care.

- Must provide cost effective patient care by improving healthcare quality and using latest knowledge.

- Usage of electronic and medical records

- Improving coordination among healthcare providing agencies.

Healthcare IOT

Create unstructured data, which enable monitoring patient parameter such as glucose, BP, ECG and necessities of visiting physician.

(10) Analytics Scalability to Big Data.

Vertical scalability means scaling up the given system's resources & increasing the system's analytics, reporting & visualization capabilities. Scaling up means designing the algorithm according to the architecture that uses resources efficiently.

Horizontal scalability means increasing

the number of systems working in coherena and scaling out the workload.
Processing different datasets of a large dataset deploys horizontal scalability.
Scaling out means using more resources and distributing the processing & storage tasks in parallel.

The easiest way to scale up ~~or~~ and scale out execution of analytics software is to implement it ~~to~~ a bigger machine with more CPUs for greater volume, velocity, variety and complexity of data.

Alternate ways for scaling up and out processing of analytics software and Big Data analytics deploy Massively Parallel Processing Platforms (MPPs), cloud, grid, clusters and distribution softwares.

Massively Parallel Processing Platforms

- It is required to enhance (scale up) the computer system or use MPPs
- Parallelization of tasks can be done at several levels.

i) distributing separate tasks onto separate threads on the same CPU

ii) distributing separate tasks onto separate CPUs on the same computer

iii) distributing separate tasks onto separate computers.

- Multiple computer resources are used in parallel processing systems.
- The computational problem is broken into discrete pieces of sub-tasks that can be processed simultaneously.
- The system executes multiple program instructions or sub-tasks at any moment of time.
- Total time taken will be much less than with a single compute resource.

(11) Big Data Sources ~~and~~

- Social networks & web data like Facebook, Twitter emails, blogs etc
- Transactions data & Business Processes data, such as credit card transactions, flight bookings etc
- Customer master data, such as data for facial recognition, name, DoB etc.

- Machine generated data such as machine to machine or IoT data from sensors, trackers, web logs and computer system logs.

- Human-generated data such as biometrics data, human-machine interactions data, e-mail records with mail servers and MySQL databases.

(12) Big Data Storage:

Big Data store uses NoSQL.
It stands for Not only SQL.
The stores do not integrate with applications using SQL.
NoSQL is also used in cloud data store.

Features of NoSQL.
- It is a class of non relational data storage systems and its flexible data models and multiple schema.
- The class consists of uninterrupted key/value or a big hash table.
- class consists of unordered keys using JSON
- May not used fixed table schema.
- Do not use JOINS
- Data written in one node can replicate to multiple nodes, therefore fault tolerant.

- Relaxed on the ACID properties for transactions.
- It follows the CAP theorem where at least two of the three must be followed.

(13) Berkeley Data Analytics Stack (BDAS)

Big Data should be able to help in obtaining the following findings:
i) cost reduction
ii) time reduction
iii) new product planning & development
iv) smart decision making using predictive analysis
v) knowledge discovery.

BDAS is an open source data analytics stack for complex computations on Big Data.
It supports large-scale, in memory data processing & thus enables user applications achieving three fundamental processing requirements
- accuracy - time - cost

It has three important layers -
→ Data Processing, applications, AMP-Genomics and Carat run at BDAS.
Data processing software component provides in-memory processing which process the data effectively

across the framework.
AMP stands for Berkeley's Algorithms, Machines and Peoples Laboratory.
→ Data processing combines batch, streaming & interactive computation.

→ Resource management software component provides for sharing the infrastructure across various frameworks.

(14) Big Data in Marketing and Sales

- Customer Value depends on three factors
  - quality, service & price.

A definition of marketing is creation, communication and delivery of value to customer.
CVA means Customer Value Analysis.

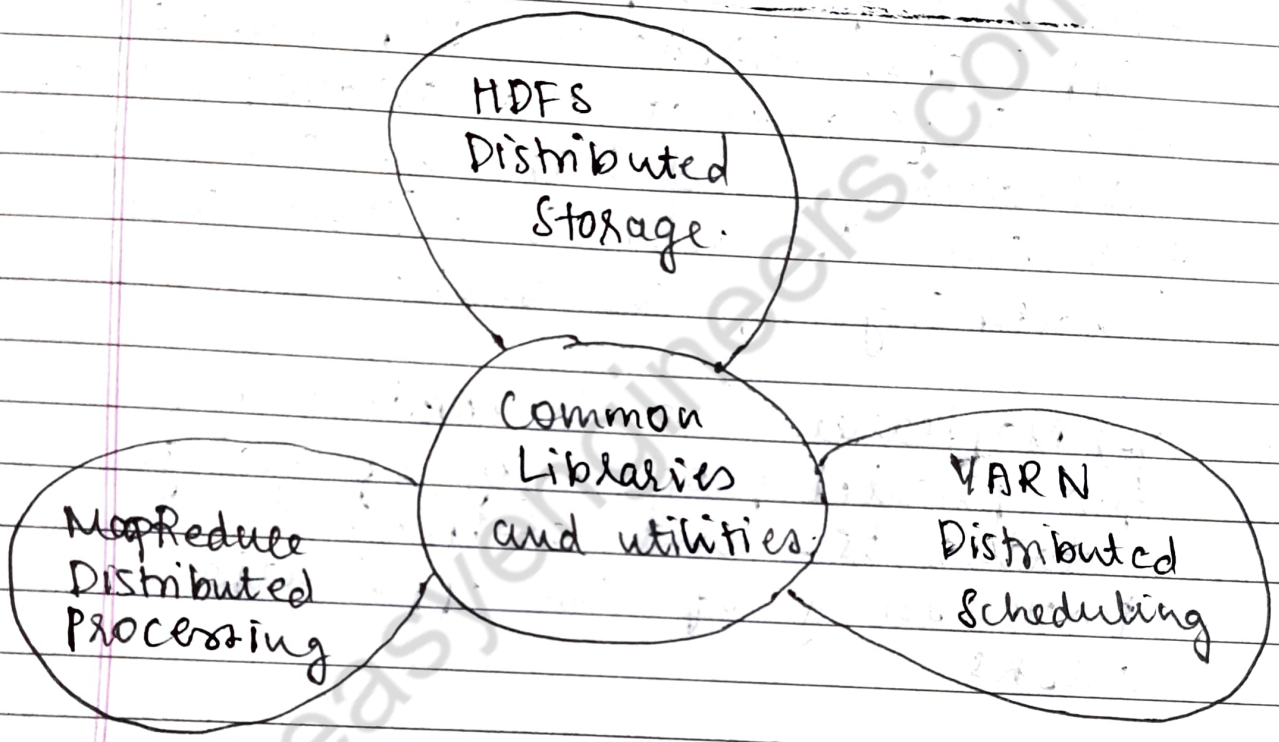
These are the five application areas in order of the popularity of Big Data uses ~~case~~ cases.

i) CVA using the inputs of evaluated purchases patterns, preferences of quality, price and post sales servicing requirements.
ii) Operational analytics for optimizing company operations.
iii) Detection of frauds and compliances

iv) New products and innovation in service
v) Enterprise data warehouse optimization

Module-2

(1) Hadoop Core components



The hadoop core components of the framework are:

i) Hadoop Common: It contains libraries and utilities that are required by the other modules of Hadoop.
Ex: Java RPC

ii) Hadoop Distributed File System (HDFS)
Java-based distributed file system which can store all kinds of data on the disk at a cluster

V1 iii) MapReduce - Software programming model in Hadoop 1 using Mapper and Reducer. It will processes large sets of data in parallel and in batches

iv) YARN - Software for managing resources for computing. The user application tasks or sub-tasks run in parallel at the Hadoop, uses scheduling and handles the requests for the resources in distributed running of the tasks.

V2 v.) MapReduce - It is YARN based for parallel processing of large datasets and distributed processing of the application tasks

(2) Hadoop ecosystem components.