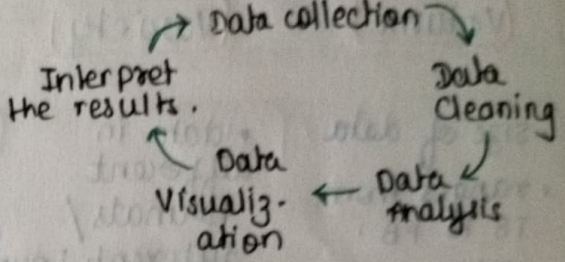
Unit 1 DSBDA :

Introduction to Data Science & Big Data.

Data Science :- It is an interdisciplinary field that uses scientific methods, algorithms & systems to extract knowledge and insights from structured and unstructured data.

↳ Data Science provides tools & techniques to analyze raw data.



Big Data :-

↳ Refers to extremly large & complex data sets that are difficult to process using traditional data processing applications.

↳ Relationship ? Big Data provides the raw material (large, complex datasets), while Data Science provides tools & techniques to analyze & make sense of material.

↳ Big Data is the oil, Data Science is the refinery.

Introduction to Big Data :-

Definations :-

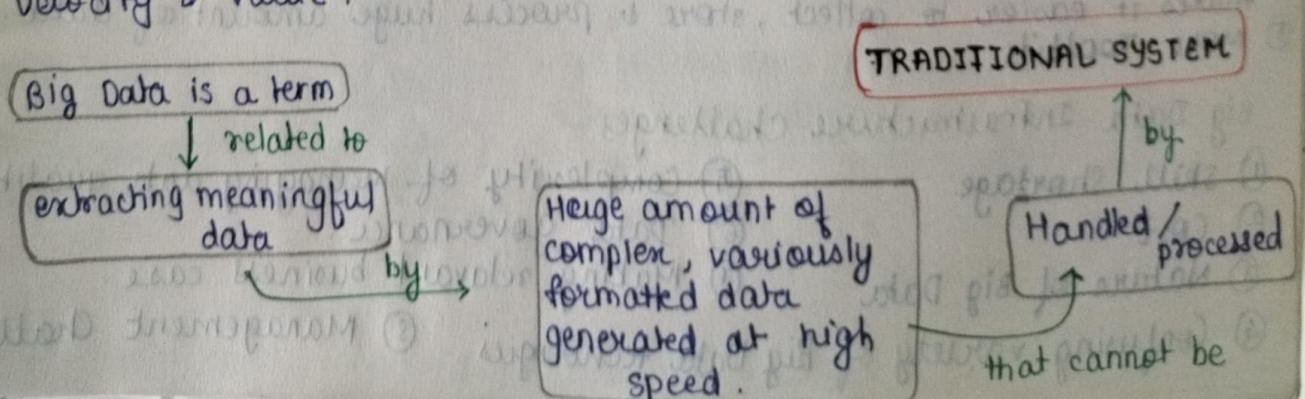
↳ Big data consist of extensive datasets that require a scalable architecture for efficient storage, manipulation & analysis.

↳ Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your db architectures. To gain value from this data, you must choose an alternative way to process it.
- O'Reilly

↳ Big data is the data characterized by 3 attributes: vol^m, variety & velocity.
- IBM

↳ Big data is the data characterized by 4 attributes: vol^m, variety, velocity & value.
- Oracle

↳



Examples of Big Data Applications :

- Fraud Detection
- IT log analytics
- medical advancement
- Call Centre analysis
- Social Media analysis
- E-commerce & marketing

Data Explosion:- Rapid growth of data

Sources of BD:
- Stock exchange
- Social media data
- Video sharing portals
- Search engine data
- Transport data
- Banking data

9] Characteristics of BD / Data Explosion
V's of BD

Volume
- size of data.
- ranges from TB to PB.
- it's a LOT of data.
- emphasis scalability challenges.
- eg. Facebook generates PB of data daily.

Variety
- data in different formats / different forms that data takes
- eg. structured, semi-structured, unstructured data.
- eg. Hospital patient data

Velocity
- describes rate at which data is generated & needs tobe processed.
- eg. financial markets require real-time analysis of stock prices.

Veracity
- quality & accuracy of data.
- e-commerce site needs to ensure the customer addresses are accurate to avoid shipping errors & maintain customer satisfaction.

Value
- potential for insights & benefits that can be derive from analysing the data.
- measures the degree of usefullness of data.
- eg. Netflix analyzes the viewing patterns to understand preferences.

Variability
- Inconsistency / fluctuations in data flow.
- Weather forecasting models must account for consistent changing atmospheric conditions.

Sheer / Big Size → Diverse/Many Types → fast speed → Data Truth → Good Use → Data Change / Changing Nature.

Factors responsible for data Explosion
1. IOT (Everything is connected) : so tons of new data all the time.
2. Social Media Platforms (Twitter, Insta, Fb) : adds to new data mountain.
3. Conversational AI (Chatbots) : adding data each time you interact with it.
4. Technological Advancements : cloud storage & better, faster computers makes it easier to collect, store & process huge amount of data.
5. More Toolkits

Big Data Infrastructure Challenges
1. Skill Shortage
2. Cost
3. Nature of Big Data
4. Confusing variety of Big Data techologies
5. Complexity of managing data quality
6. Data governance.
7. Making relavant business case.
8. Management Challenge

Big Data Processing Architectures.

Data Warehouse :- central place to store tons of data for analysis and decision-making.

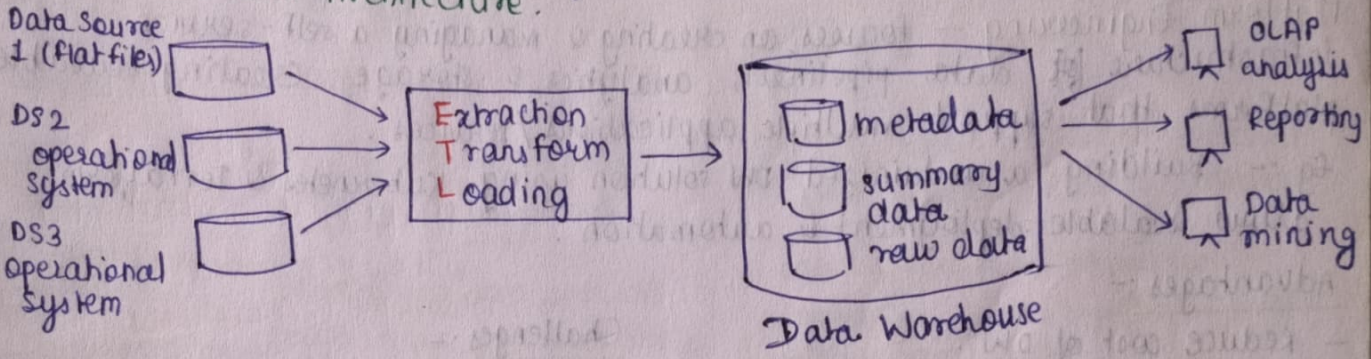
A DW is a central repository of info that can be analyzed to make more informed decisions.

Purpose :- TURN RAW DATA → USEFUL INSIGHTS.

DW is RDB(Relational Database) designed for analysis.
DW can store both current & historical data in one place.
Key Features :- Subject Oriented, Integrated (combines data from different sources), Time Varient (Stores historical data), Non-Volatile (Data is stable & read only).

How it Works :-
1. ETL - get data, clean it and load in right format.
2. Store - data is stored in warehouse.
3. Analyze - Use tools (OLAP engines, customer analysis tools) to find trends & answer questions.

Data Warehouse Architecture.



Bottom Tier (Data Storage) → collect & store raw data. DW, ETL
Middle Tier (Processing) → Transform & analyze data. OLAP engine
Top Tier (Reporting) → Visualize & present insights. UI, PowerBI, Tableau

Steps in Data Warehousing :-
1. Extraction of Data.
2. Cleaning of Data.
3. conversion of data.
4. Storing in a warehouse.

Benefits :-
- Better data quality.
- Faster, business insights.
- Smarter decision making.
- Gaining & growing competative advantage.
- Most cost-effective decision making.
- Enhance customer service.
- Saves time.
- Potential high returns on investment.

Limitations :-
- High Maintenance.
- Long-duration projects.
- Increased end user demands.
- Complexity of integration.
- High demand for resources.
- Data ownership.
- Data homo-geni-zation.

Re-Engineering the Data Warehouse :- (Modernizing existing DW)
Transforming & optimizing its architecture to improve data management, analytics & efficiency. The choices depends on business goals, technological advancements and infrastructure needs.

Re-platforming :- involves migrating DW to new platform including all hardware & infrastructure, such as changing the database engine, improving ETL pipelines, etc.

Eg:- Moving from Hadoop-based DW to Google Bigquery, optimizing schema for faster queries.

Advantages :-
- Low cost of ownership.
- Improve speed & performance.
- More secure data.
- Better disaster recovery.
- Leveraged cloud flexibility & agility.

Challenges :-
- Requires query & workflow adjustments.
- Some legacy features may not be supported.

Platform Engineering :- focuses on creating & managing a self-service infrastructure for data pipelines, analytics & storage. Creating reusable platforms that support multiple applications / services.

Eg :- Building a containized DW solution using Kubernetes & Terraform, allow scalable deployment & automation.

↳ pizza shop without PE :- make pizza from stratch
with PE - has already pre-made dough & ingredients

Advantages :-
- Reduce cost of DW.
- Increase efficiencies of processing.
- Reduce Redundancies.
- Minimise costomisation.
- Isolate complexity into manageable modular environments.

Challenges :-
- Requires specialized skills (DevOps & Cloud)
- High setup complexity & maintenance effort.

Data Engineering :- focuses on building efficient data pipelines, transforming data & optimize storage formats for better processing. ie Data structures are re-engineered to create better performance.

Eg :- Replacing Batch ETL jobs with real-time streaming pipelines.

Advantages :-
- Improves data freshness & reduce processing time.
- Enhances data quality & consistency.
- Enables AI/ML-ready structure datasets.

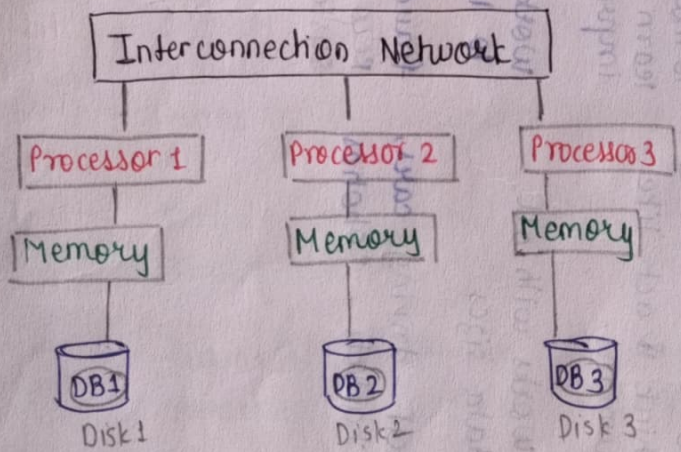
Challenges :-
- Requires expertise in big data frameworks.
- Maintainance is complex.

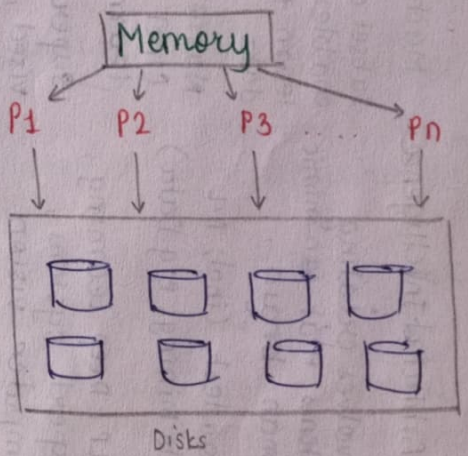
Re-building :- Building new DW using modern architectures.
eg. Replacing Batch DW with real-time Data Lakehouse.

Re-architecturing :- Redesigning DW architecture.
eg. Converting normalized schema to denormalized Star Schema to speed up query.

Shared-Everything & Shared Nothing Architecture

	Shared Nothing Architecture	Shared Everything Architecture.
Resource Sharing	No sharing, each node has its own CPU, mem, & storage.	Nodes share memory as well as storage.
Disk Usage	Each node has its own local disk, no shared storage.	A single shared storage system for all nodes.
Hardware cost	Cheap since no shared storage required.	Expensive due to shared-memory and high-performance storage.
Data Partitioning	Strictly partitioned	Data is not strictly partitioned.
Load Balancing	fixed	Dynamic
Scalability	High	Limited / Low
failure Handling	If a node fails, only its data is affected.	If one node fails, other can still access shared storage.
Performance	Better performance at scale since no contention.	Can suffer from resource contention when multiple nodes access same data.
Use case	Best for Big Data & Analytics where distributed processing is needed	Best for OLTP systems where consistency is critical.
Examples	Google BigQuery, Apache Cassandra	Oracle RAC, IBM DB2





The Big Picture :- AI → includes ML → uses Statics & Data mining → Applied in Big Data Analytics.

Feature	Artificial Intelligence	Machine learning	Statistical learning	Data Mining	Big Data Analytics.
Definition	Involves creating systems that can mimic human brain.	Subset of AI that enables machines to learn patterns from data.	Uses mathematical & probabilistic models to analyze data.	Extracts useful patterns, relationships and trends from large datasets.	Analyzes massive structured & unstructured datasets for insights.
Scope	Broadest (incl; ML, Data Mining, Big Data)	Narrower than AI (incl. statistical learning)	Theoretical foundation of ML.	A practical application of ML & stats.	Uses AI, ML & DM techniques for analysis.
Techniques Used	NLP, Deep learning, Expert systems, Computer vision	Supervised, Unsupervised, Reinforcement learning.	Regression, Probability, Hypothesis testing.	Clustering, Classification, Association Rules.	Hadoop, Spark, Data Lakes, cloud Computing.
Goal	To make machines think & act like humans.	To make machines learn from data & improve performance.	To use mathematical models for predictions	To find hidden patterns in data	To process & analyze vast amount of data.
Data Size	Works with all data sizes	Works with medium to large datasets	Works with small to medium datasets	Works with structured & semi-struc. data.	Works with huge datasets (TBs to PB)
Example	self-driving cars, Chatbots, Robotics	fraud Detection, Recommendation Systems.	Linear Regression, Bayesian Analysis	Market Basket Analysis, Customer segmentation	Real-time Analytics, Social media trends, Predictive Maintenance.

focus on algo models to predict

exploring data

Recommend product

Tools

Train to make better prediction

understand relationship / trend in data

process

Exploring like book → then buy coffee

Millions of transactions

PYQ's.

Feature	Small Data	Big Data.
Definition	Refers to manageable datasets that fit within standard tools like Excel or SQL databases	Consist of massive datasets that require advanced storage & processing techniques.
Size	MB or GB	TB, PB or EB
Structure	Mostly structured, stored in RDB, spreadsheets, or small csv files.	Can be structured, semi or unstructured.
Processing speed.	quick on single computer with traditional tools.	Requires distributed computing using Hadoop, Spark, etc solutions
Tools used	Excel, SQL, Python, R, simple databases.	Hadoop, Spark, NoSQL db, AI/ML frameworks.
Storage	Stored in local machines, small servers or RDB.	Requires distributed storage systems like HDFS, AWS, Google cloud.
Use cases	Small business reports, local customer surveys, med records.	Social media analysis, real time fraud detection, IOT sensor data processing.
Example	Sales data of a single store in month	Data generated by millions of users on facebook.

Feature	Data Warehouse	Data Mining.
Defn	centralized storage system that integrates data from multiple sources for analysis & reporting.	process of extracting patterns, trends and insights from large datasets.
Purpose	Stores historical & current data for decision-making.	Data analysis & knowledge discovery.
Techniques used.	ETL, OLAP	Clustering, Classification, Association Rules, Regression.
Data Type	Large-scale structured data from different sources.	Works with structured, semi-structured & unstructured data.
Tools used	Oracle, Amazon Redshift, Snowflake, IBM Db2	Python, R, Rapid Miner, WEKA Apache Mahout
Use Cases	Business Intelligence, finance reporting, healthcare analytics	fraud detection, customer segmentation, recommendation systems.
Example	Bank's data warehouse storing customer transactions.	Detecting fraud credit card transactions