Date:

Intro to Big data

Big_data with characteristics & types

Big Data:

Massive amount of data which cannot be stored, processed and analyzed using traditional tools is known as Big Data.

2* Big data issues:
- Collection issue
- Storage issue
- Processing issue
- Analyzing issue

1* Big data is characterized by it's Volume, Velocity, Variety & Veracity. (known as 4 V's of Big data)

Note:

Volume: Volume is the increase in amount of data stored, over time.

Velocity: Velocity is the speed of data, at which data arrives.

Variety: Variety is the diversity of data formats and types.

Veracity: Varacity is the certainty, accuracy, quality or reliability of data.

Date:

3 * On the basis of data structure, Data can be categorized into:

① Structured data
② Unstructured data
③ Semi-structured data

Further explanations are below:

* **Structured data:** Structured data has a well-defined structure or organized and formatted in a specific structured manner.
  - In many cases has tabular manner with rows and columns
  - We can typically store structured data in relational databases.
  - Structured data is objective facts and numbers that can be collected, exported, stored, and organized in typical databases.
  - Some sources of structured data

Date:

could include: SQL databases, Online Transaction processing (OLTP), Spreadsheets, Sensors, Online forms etc

* **Unstructured data:** Unstructured data does not have an easily identifiable structure, or data that can not be organized in a predefined or structured manner.
  - It can not be stored or organized in a mainstream relational databases in the form of rows and columns.
  - It does not follow any particular format, sequence, semantics, schemas or rules.
  - Unstructured data can deals with the heterogeneity of sources and has a variety of business intelligence and analytical applications.
  - Some sources of unstrutured data

Date:

could include: Web pages, Social media feeds, images, videos, audios, PDF files and power point presentations etc.

* **Semi-Structured data:** It has some organizational properties but lacks a fixed or rigid schema or organizational structure. Or, Semi-structured data is a type of data that is partially structured and partially unstructured. It contains elements of both structured data and Unstructured data data.

- It has some organizational structure or schema, but not enough to fit neatly into a relational database, as it does not have fixed or rigid schema.

Date: _____

the data more easily and make informed decisions.

Diff stages from which data passed

Stage#1 Data Acquisition (Collection/Ingestion)

This stage involves collection of data from various data sources. *Some important data sources include:

Data Sources

① Databases:

Databases are important source for data, such as SQL, NoSQL and Hadoop file.

Note

Database: A database is a structured collection of data that is stored and organized in a way that allows for efficient retrieval and manipulation of data.

↳ There're two types of database

Relational (/SQL): Use table based-structure to store data.

Non-relational (/noSQL): Does not use the table-based struct to store data.

② Third-party data providers:

Purchasing

Date:

data from third-party data providers who specialize in collecting data, is also a source of data.

* Such data sets are typically made available as flat files, spreadsheets files or XML documents, JSON file etc

Note

Flat files: Flat files store data in plain text format, with one record or row per line, and each value is separated by delimiters such as commas, semi-colons, or tabs.

↳ Data in flat file map to a single table, unlike relation database that contain multiple tables.

↳ CSV (Comma-separated Values), TSV Tab-separated files, ~~JSON (Javascript object Notation)~~ are famous examples

XML (Extensible markup language) file: is ~~a markup language used to store data~~ and exchange data in a structured format with the help of tags that define the structure and content of the data.

JSON (Javascript Object Notation) file: is a type of file in which information is stored in a way that's similar to dictionary, where each piece of the data is represented by key-value pairs

↳ It's lightweight data-interchange format that is easy to read and write.

Date:

③ APIs:

Accessing data through APIs (application programming interfaces) provided by different software applications is also a source of data.

Note:

API: It is a set of protocols, routines, and tools that allow software applications (/components) to communicate with each other
↳ including what types of data can be sent, and recieved, and what actions can be performed.

↳ Web APIs: These're APIs that are accessed over web, typically through HTTP requests. Web APIs are commonly used for accessing web services and other online resources

Operating System APIs: These're the API's provided by operating systems to access system resources such as file system, network devices, & hardware components.

Database APIs: These are the APIs that allow applications to access and manipulate data stored in database.

Library APIs: These are the APIs that ~~allow applications~~ are provided by libraries of pre-written code, allowing developers to access the functionality provided by those libraries.

④ Web Scraping:

This is a power-full tool for collecting data

Date:

from websites, often in large volumes.

Note

Web Scraping: Web scraping is the process of automatically extracting data from websites. It involves using softwares tools or scripts to access and extract data from web-pages.

↳ The process typically involves accessing the HTML code of a web page and using regular expressions or other techniques to extract specific data points.

↳ Some websites may prohibits web scraping, and legal issues can arise if copyrighted or proprietary information is scrapted without premission.

⑤ IoT (Internet of Things):

IoT devices such as sensors, smart appliances, and wearable generate a massive amount of data, including Temprature, humidity, location, and data usage etc.

* Hence, IoT devices are also the source of

Date:

data.

Note

IoT: IoT refers to the interconnectivity of physical devices (such as sensors, cameras, and other machines) with each other and with the internet, allowing them to share data and performed automated tasks.

Sensor devices: Sensor devices are a key component of IoT, as they are responsible information about the physical world and transmitting data over the internet to other devices.

⑥ Customer Relationship Mangement (CRM):

Collecting data from CRM systems, which contain customers data such as contact information, purchase history, transactional detail and preferences etc.

⑦ Publicaly Available data:

It include goverrament databases, public records and online repositories.

⑧ Social Networks:

Collecting data

from social networks such as facebook, Instagram, and TikTok etc, including user behavior data, engagement metrics, and sentiment analysis.

Step#3 Data Storage

In this stage collected and transformed (/processed) data is stored in data repository.

Some important Data Repositories:

Data Warehouse: A data warehouse is a centralized repository that store data from various sources in a structured and organized manner for business intelligence and decision making purpose.

Data Lake: A data lake is a large centeralized repository that stores raw, unstructured, semi-structured

Date:

data from various sources. Unlike a data warehouse, a data lake does not require structured data models or predefined schemas.

Lake House: A lake house is a hybrid architecture that combines the benifits of both data warehouse and data lakes. It allows for both structured and unstructured data to coexists in a centralized repository.

Data Mart: A data mart is a subset of a data warehouse that contain specific subset of data that is relevant to a particular department or business unit.

Data Hub: A data hub is a centralized repository that aggrigates and integrates data from multiple resources, such as databases, applications, and data lakes. It is typically used for data integration, master

Date:

data management, and data governance and provide a central point for data sharing and collaboration.

Note:

Data Governance: Data governance refer to the process of managing and ensuring quality, consistency, security, privacy, integrity, usability, availability and compliances of data. It involves defining policies, standards and procedures for management as well as implementing tools and technologies to enforce them.

OR

Data Governence is a collection of principles, practices, and processes to maintain the security, privacy and integrity of data.

Data Mesh: Data Mesh is a paradigm ~~approshift~~ shift in data architecture that promotes a decentralised (distributive) approach to managing and integrating data. It is based on the principles of domain-driven design, where each domain is responsible for it's own data and provide data products that can be shared and consumed by other domain.

Date:

The data mesh architecture is based on a mesh of interconnected services that collaborate to provide data products to other services in the organization.

* Distribud & Centralized Storage:

~~For big data, we also use~~

distributede storage: This approach provides better scalability, fault tolarance, and performance than traditional centralized storage systems. In distributed storage, data is distributed across multiple nodes, and each node may store a copy of data, ensuring redundancy and availability.

~~They are typically built using commodity hardware and software.~~

and

Distributed storage: It refers to the storage of data across multiple machines instead storing it on single machine. Vs
Centralized storage: It refer to the storage of data on single machine. It is more traditional approach for storing data.

- Commodity hardware is often

Date:

used in distributed storage system, as they are inexpensive and off-the-shelf hardware components. Using commodity hardware allows for the creation of large clusters of machines without the need of expensive specialized hardware.

> **Note** **Cluster:** A cluster is a group of computers that work together as a single system.
> They can range from small cluster of a few machines to large clusters with thousands of machines

Distribution Model: A distribution model refers to the way in which data is distributed across multiple machines or nodes in a cluster or a distributed storage system.

- **Replication:** In this model, data is replicated across multiple machines in cluster, with each replica serving as a backup in case

Date:

of a failure. This approach is commonly used in distributed file system such as Hadoop Distributed file system (HDFS).

- Partitioning: In this model data is divided into partitions based on some key or attribute, and each partition is stored on different machine. This approach is commonly used in distributed databases such as Cassandra.
- Sharding: This model is similar to partitioning, but instead of storing partitions on different machines, they are distributed across multiple clusters or nodes. This approach is commonly used in distributed NoSQL databases.
- Hybrid: A combination of replication, partitioning and sharding models can be used to achieve high availability, scability and fault tolerance

* **Shared Everything Architecture:** In shared everything architecture, all nodes in a cluster has its own dedicated set of physical resources, such as CPU, memory and storage etc.

* **Shared-Nothing Architecture:** In shared-nothing architecture, each node in a cluster share access to same physical resources while in shared-nothing architecture, each node has its own dedicated set of physical resources, such as CPU, memory and storage etc.

Date:

Stage #2 Data ~~storage~~ cleaning/ Processing/ transformation

In this stage ~~stored~~ collected data is processed (/transformed) in a more refined and structured format, to make ready it for analysis.

1 *This may include data cleaning, data normalization, data enrichment and data validation and data transformation, the primary goal of this stage to prepare data for analysis tools or downstream applications.

4 * **Batch & Stream Processing:**

Batch Processing: Batch processing refers to processing of large amount of data all at once, in a batch.

- In this method, data is collected over a period of time, and then processed in a batch at a specific time interval or when a certain threshold is reached.

Date:

- Batch processing is often used for large-scale data analysis, i.e; running complex analytics on a historical data, generating reports or performing computation on large volume of data.

Stream Processing: Stream processing refers to the processing of data in real-time as it is generated.

- In this method, data continuously following and is processing as it arrives, often in small, incremental updates.
- Stream processing is often used for monitoring and analyzing data in real-time i.e; detecting anomalies, making immediate decisions or taking actions based on real-time data feeds, and perform real-time analytics.

Date:

* OLTP & OLAP:

OLTP: Online Transaction Processing (OLTP) is used for transactional processing of large volumes of data in real-time. It is designed for processing a large ~~amount of data~~ number of small transactions, i.e customers' orders, payments, and inventory updates.

- The focus of OLTP is on transaction processing, which involves adding, updating, and deleting data in database.
- OLTP databases are optimized for fast read/write operations and typically use a normalized data model.

OLAP: Online Analytical Process (OLAP) is used for analytical processing of large volumes of data. It is designed for complex queries that involves

Date: _____

aggregating and summarizing data across multiple dimensions such as time, geography, and product categories.

- The focus of OLAP is on analysis of data which involves querying and aggregating data to support decision-making.
- OLAP databases are optimized for fast read operations and typically use a denormalized data model, which allows for faster data retrieval.

2 *

5 do stages in data Processing stage:

(i) **Data Cleaning:** In data cleaning, the data is cleaned to remove any inconsistencies or errors, such as missing or duplicate values, invalid characters, or formatting issues.

(iii) **Data Enrichment:** In data enrichement, additional data is added to the raw data to provide more content and insights. This could include adding geographic data, demographic data, or other external data sources

(ii) **Data Transformation:** This involves transforming data into a format that is more suitable for analysis or consumption by downstream applications. This could include aggregating data and converting data types

(iv) **Data Normalization:** This involves standardizing data so that it conforms to a specific set of standards. i.e converting data to common unit of measurement, standaring data formats

(v) **data validation** It involves checking process data for accuracy, completeness, consistency, and confirmity to specific rules or standards. It ensure that the data is of high quality and ready of analysis.

ensuring data is in certain range.

Note

Date:

3 * **Parallel and Distributed Computing:**

Parallel Computing: Parallel computing refers to the use of multiple processors or cores within a single machine to process a task simultaneously.

- In parallel computing, the workload is divided into smaller sub-tasks that can be executed in parallel on different processors.
- The processors communicate with each other to exchange data and co-ordinate their work. Parallel computing is often used for tasks that can be easily divided into smaller, independent parts.

Distributed Computing: Distributed computing refers to use of multiple machines connected over a network that process task ~~contributly~~ with contribution

Date: _____

- In distributed computing, the workload is divided into smaller sub-tasks that are assigned to different machines,
- And then machines communicate with each other to ~~communicate~~ exchange data and coordinate their work. Distributed computing is often used for tasks that require (processing) large amounts of data, such as data analytics, big data processing data streaming etc

Stage #4 Data Analysis

Once the data has been cleaned and stored, we explore/ Analyze the data to retrieve insight from it, this is called Data Analysis.

* The object of this stage is to understand the data.

Date:

* Data Analysis can be classified as Confirmatory analysis and Exploration analysis.

Confirmatory Analysis: In confirmatory analysis, the cause of a phenomenon is analyzed first (/before). The data is analyzed to approve or disapprove the hypothesis (/assumption).

- This kind of analysis provide definitive answers to some specific questions and confirms whether an assumption was true or not.

Exploratory Analysis: In exploratory Data Analysis (EDA), the data is explored to obtain information, why a phenomenon occured. This analysis answer "why" phenomenon occured.

Date: _____

- This kind of analysis does not provide definitive, meanwhile it provide discovery of pattern.

* Here, Statistical and machine-learning techniques are applied to extract insights and information (/ knowledge) from the processed data.
  - These techniques are applied to identify patterns, trends, relationships and anomalies in the data, and to make predictions and recommendations based on those insights.

Stage #5

Report generation/

Data Visualization

This stage involves creating visual represrepresentations of the data in order to better understand and communicate the insights that have been discovered.

Date:

* Some common types of data visualization include:

Charts and graphs: These are visual representations of numerical data, such as bar charts, line graphs, and pie charts etc. They can be used to show trends over time, compare different data set, or illustrate the proportions of different values.

Maps: Maps can be used to visualize spatial data, such as the locations of customers, stores, or other assets. They can help to identify geographic patterns and relationships in the

Dashboards: Dashboards are a collection of visualizations and other data related

Date:

elements, such as tables and charts, that is presented in single screen.
They can provide a comprehensive overview of key performance indicators and other metrics.

Infographics: Infographics are visual representations of data that combine charts, graphs, and other visuals elements with text and images.
They can be used to present complex data in a simple and engaging way.

* Data visualization is an imp tool for turning data into insights that can be easily understood and acted upon. It helps to communicate complex information in a way that is accessible & and can help to drive better decision-making and business outcomes.

Date:

Stage #6 — Acting to solve Business Case

Now going through the all stages raw data is now in the form of valuable information / insights, that are used to make business decisions. i.e. These decisions can be related to product management, customer engagement, marketing campaigns, or process improvement.

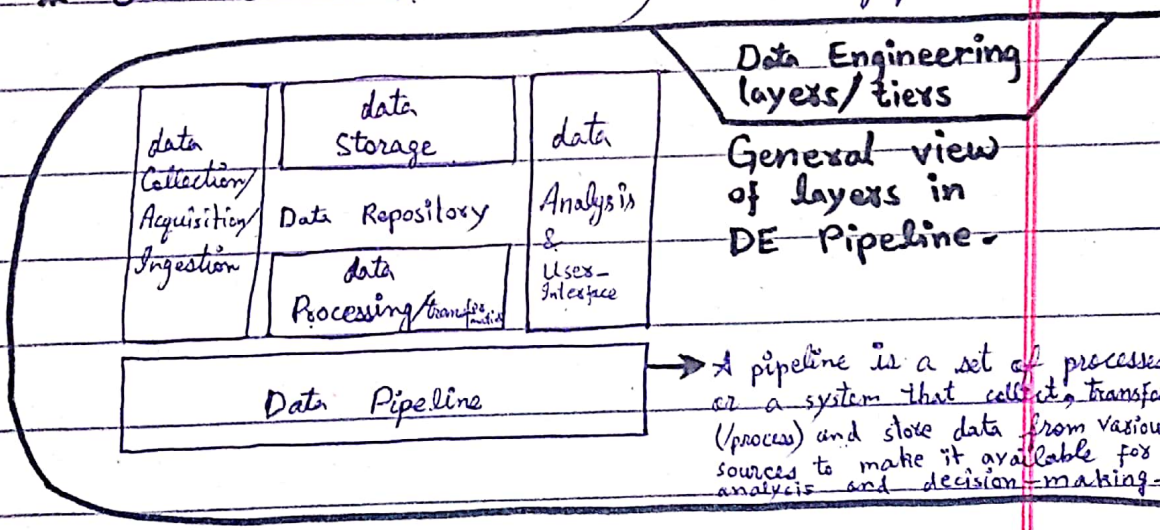
* It involves identifying and prioritizing the use case based on the insight generated through whole process.
* The business case also involves developing a plan to implement to implement the insights and measuring the impact of decisions made.

Date:

Big Data Engineering

Big data Engineering refers to the process of designing, building and maintaining the infrastructures and tools needed to collect, store, process, and analyze large volumes of data/ big data.

* It involves creating data pipelines.



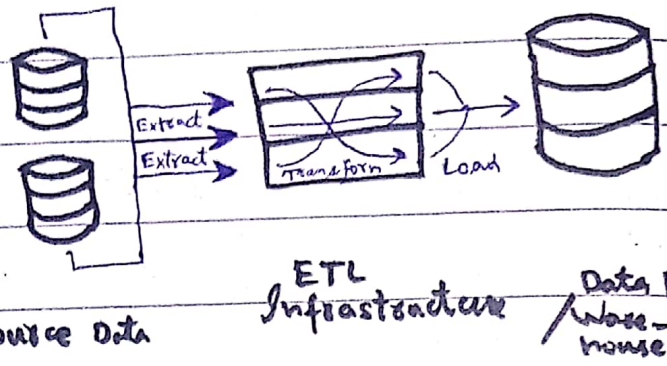
* ETL & ELT are two common approaches to build a data pipeline.

- **ETL (Extract, Transform and Load)**

It is a traditional approach to build a data pipeline, In which data is first extracted from

Date:

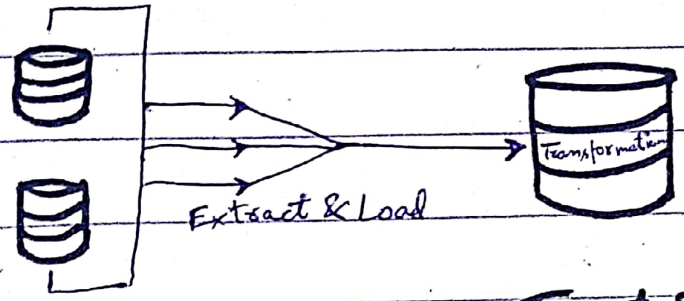
multiple data sources, then transform (or cleaned) and finally loaded into the target system (or data repository).



- **ELT (Extract, Load and Transform)**

In this approach, data is first extract and then loaded as-is. The transformation (or cleaning) of data happens after the data is loaded into target system. Here we typically use the power of cloud data warehouse to process/ transform the raw data. Suitable for scenarios where target system has a high processing power and can handle transformation after data is loaded.

Date: _____



Types of Load:

① **Full Load:** In a full load all data from the source system is extracted and loaded into target system. This is typically done when setting up a new system or when a major change has been done/ made to the source data.

- Historical load is a full load, in which we load historical data into a target system.

② **Incremental load:** In a Incremental load, only the changes that have occurred in the source data since the last load are extracted and loaded into the target system.

- Real-time load: A real time load is used to load data into a target system in real-time. This type of load is typically done when the target system requires up-to-the-minute data for decision-making.
- Batch load: A batch load is a type of load that is scheduled to run at a specific interval. This is typically done when the target system can handle large volumes of data processing in batches rather than real-time.

Date:

* Data Engineer:

A Data Engineer is a type of software engineer who create and maintain data pipelines