BYTEWISE LIMITED

Data Engineering Track

Task: Week – 1 (First Month)

Task No: 1

Task Date: 14-03-2023

Internee Name: Umer Farooq

Mentor Name: Ahtisham

Task Details:

This task includes the following:

1. Big Data
2. Data Lake
3. Database
4. Data Warehouse

Each terminology will be briefly explained in document that is also available at GitHub repository.

1.BIG DATA:

- Big data refers to **extremely large and complex data sets** that are too big to be processed using traditional data processing tools and techniques.
- Big data can come from various sources such as:
    - social media,
    - sensors,
    - machines, and
    - other data-generating devices.
- Processing and analysing big data require specialized technologies such as Hadoop, Spark, and NoSQL databases.

For example, a healthcare provider might collect patient data from various sources, including electronic health records, wearables, and medical devices. Analysing this data using big data technologies can reveal patterns, trends, and insights that could improve patient outcomes, reduce costs, and enhance the overall quality of care.

Q Big Data is basically 3V's.

A - In this world, everyone leaves trace, from our travel habits to our workouts & entertainement, the increasing no. of internet connected devices that we interact with on a daily basis record vast amounts of data about us.

A Big Data refer to the dynamic, large, & disparate volumes of data being created by people, tools, & machines.

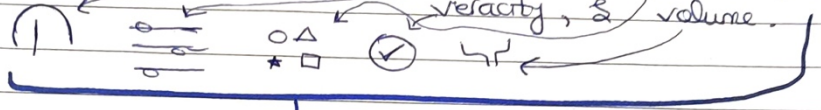
It requires new innovative, scalable technology to collect, host, & analytically process the vast amount of data gathered in order to derive real-time business insights that relate to customers, risk, profit, performance, productivity management

NADEEM

Date: _____ M T W T F S S

& enhance value."

- There are no accurate defination of Big Data, but every defintion has certain element that are same : velocity, volume, variety, veracity, & volume.



↳ The V's of Big Data

Inconsistent data flows with periodic data peakes

① : Velocity :- → Velocity is the speed at which data accumulates.

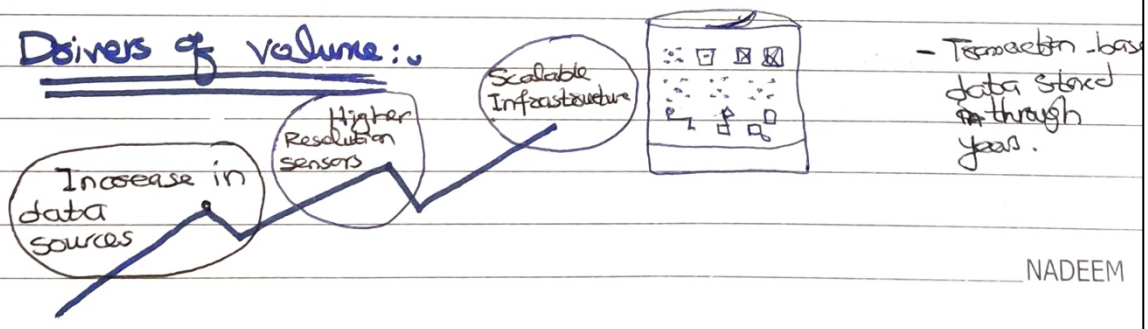
→ Data generated extremely fast, in process that never stops.

→ Near or real-time streaming, local, & cloud-based technologies can process info very quickly.

② : Volume :

- Volume is the scale of the data, or the increase in the amount of data stored.

- Drivers of volume :-



- Transaction-based data stored through years.

NADEEM

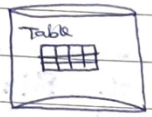
Date: _____ M T W T F S S

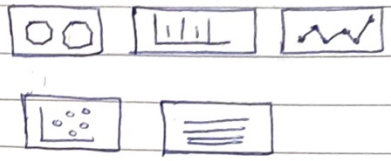
Semi-structured data XML & JSON

③ : Variety :

- Variety is the diversity of the data.

- Structured data fits neatly into rows & cols, in relational databases.





- While Unstructured data is not organized in a pre-defined way like Tweets, blog posts, pictures, numbers, & videos.

- Variety also reflects that data comes from different sources, machines, people, & processes, both internal & external to organization.

- **Drivers:** (variety sources)
  - Mobile Technologies
  - Social Media
  - Wearable Technologies
  - Geo Technologies
  - Video

NADEEM

Date: _____ M T W T F S S

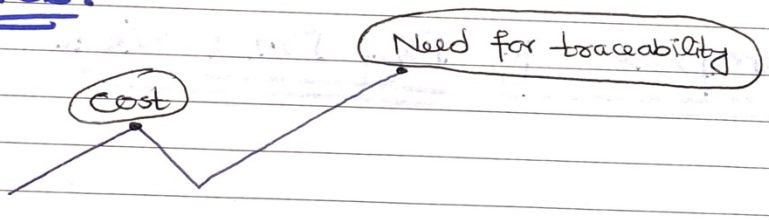
(4): Veracity :v ✓

- Veracity is the quality & origin of data, and its conformity to facts & accuracy.

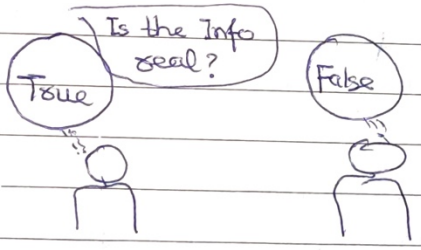
Attributes of Veracity:

1. Consistency
2. Completness
3. Integrity
4. Ambiguity

- **Drivers:**



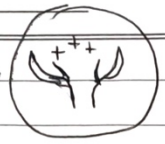
- With large amount of data available, the debate on veracity is there.



NADEEM

Date: M T W T F S S

⑤ Value :-



- Value is our (devs) ability & need "to turn data into value".
- It may have medical or social benefits, as well as customer, employee, or personal satisfaction. So, its not only about profit.
- The main reason that people invest time to understand Big Data is to derive value from it.

Examples of Big Data Vs :-

- Velocity: every 60 seconds, how: footages uploaded to Youtube is generating data.

  → So think of how quickly data accumulates over hours, day, week, moths etc.

- Volume: 7 billion population of earth, and vast majority of that is now using Mobiles, desktop & Laptop comps, wearable devices etc

NADEEM

Date: _____ M T W T F S S

- These devices generate, capture, & store data -- approx. 2.5 Quintillion bytes every day.

- **Variety:-** Let's think about the different types of data; text, pictures, film, sound, health data from wearable devices & many different types of data from devices connected to the interet-of-things (IOT).

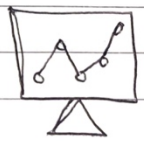
- **Veracity:-** 80% of data is considered to be unstructured & we must devise ways to produce reliable & accurate insights.
  - The data must be categorized, analyzed, and visualized.

- Data Scientists today derive insights from Big Data & cope with the challenges that those massive data sets presented.
  - The scale of the data being collected means that its not feasible to use conventional data analysis tools.

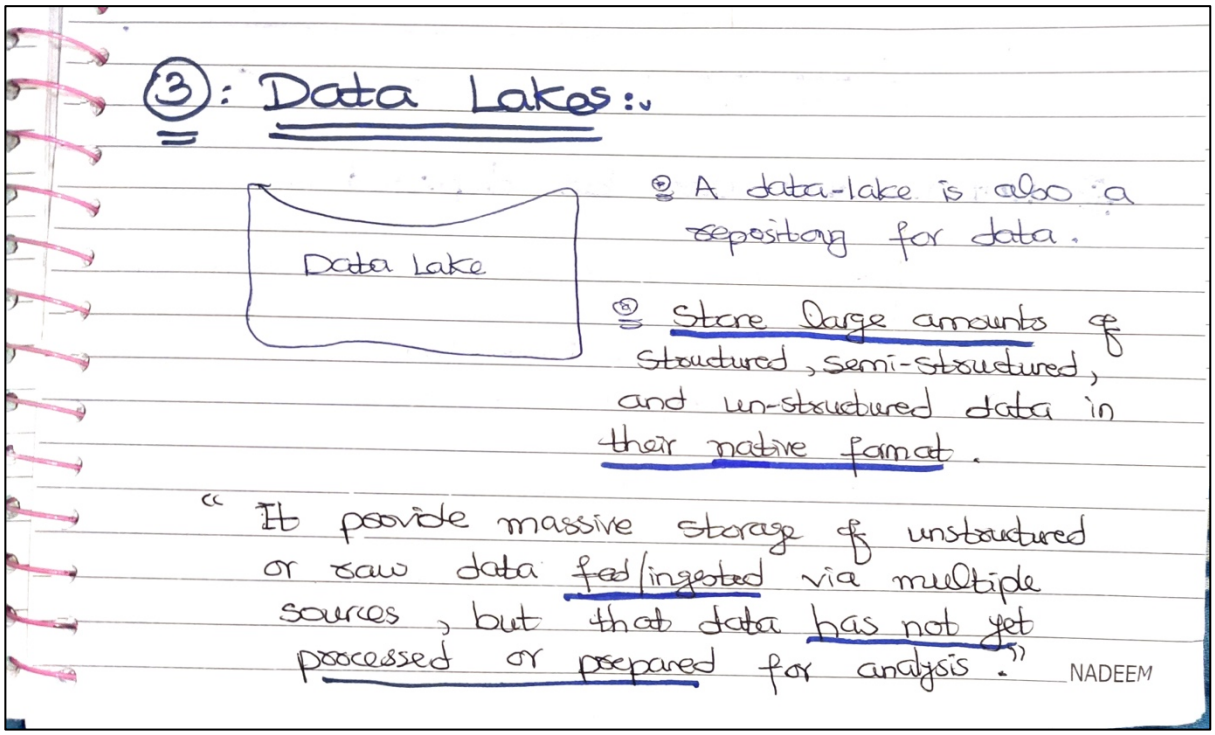
NADEEM

Date: _____ M T W T F S S

- However alternative tools that take advantage of (leveraged) distributed computing power can overcome this problem.
  - Tools such as Apache Spark, Hadoop & its ecosystem provide ways:
    - to extract, load, analyze, & process data across distributed compute resources, providing new insights & knowledge.

Apache Spark  Apache Hadoop

2.DATA LAKE:


③: Data Lakes:-
Data Lake
A data-lake is also a repository for data.
Store large amounts of structured, semi-structured, and un-structured data in their native format.
" It provide massive storage of unstructured or raw data fed/ingested via multiple sources, but that data has not yet processed or prepared for analysis. "
NADEEM

e-g of data lake: Grab uses data such as [illegible] behavior, power usage, waterways, & store it in data lake, while they figure out using the data to create "smart cities".

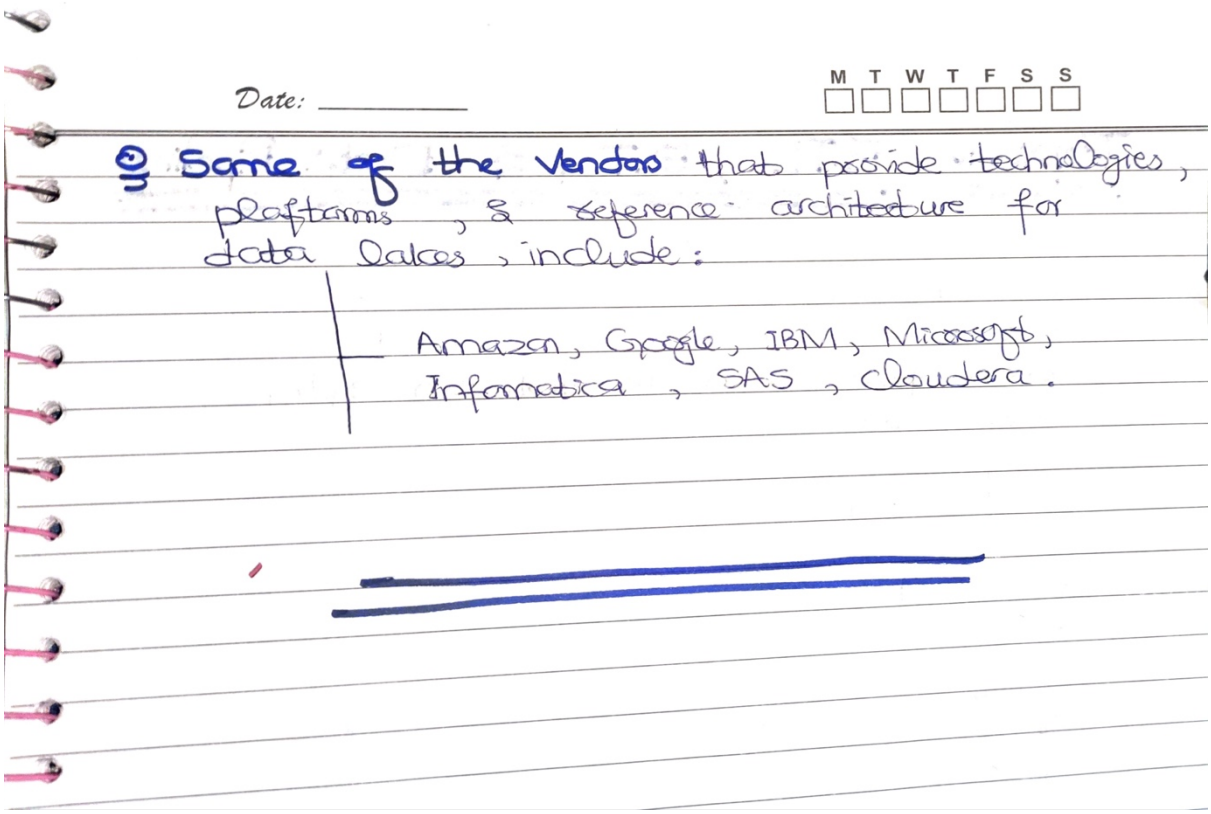
Date: M T W T F S S

- As a result of being able to store data in a raw format, data lakes are more accessible & cost-effective than data warehouse.
- Data can be loaded without defining the structure & schema of data.
- Data is transformed based on use-case.
- In Data Lake, Data is classified, protected, & governed.

→ A reference architecture (that is independant of technology) that combines multiple technologies, to facilitate agile data exploration for analyst & data scientists.

- **Data Lakes can be deployed** using:
  - → Cloud object storage, such as Amazon S3
  - → Large-Scale Distributed System such as Apache Hadoop.
  - → Relational DBMS & NOSQL data repositories that can store large amount of data.

NADEEM



Benefits of Data Lake:

I. Scalability: Data lakes can store vast amounts of data, making it easy to scale as your data needs grow.

II. Flexibility: Data lakes allow for the storage of raw, unprocessed data in its native format, making it easier to store and manage data from multiple sources.

III. Cost-Effective: Data lakes are typically less expensive than traditional data warehouses, as they use commodity hardware and open-source software.

IV. Faster Data Processing: With data lakes, data can be processed in real-time, allowing for faster decision-making and better business outcomes.

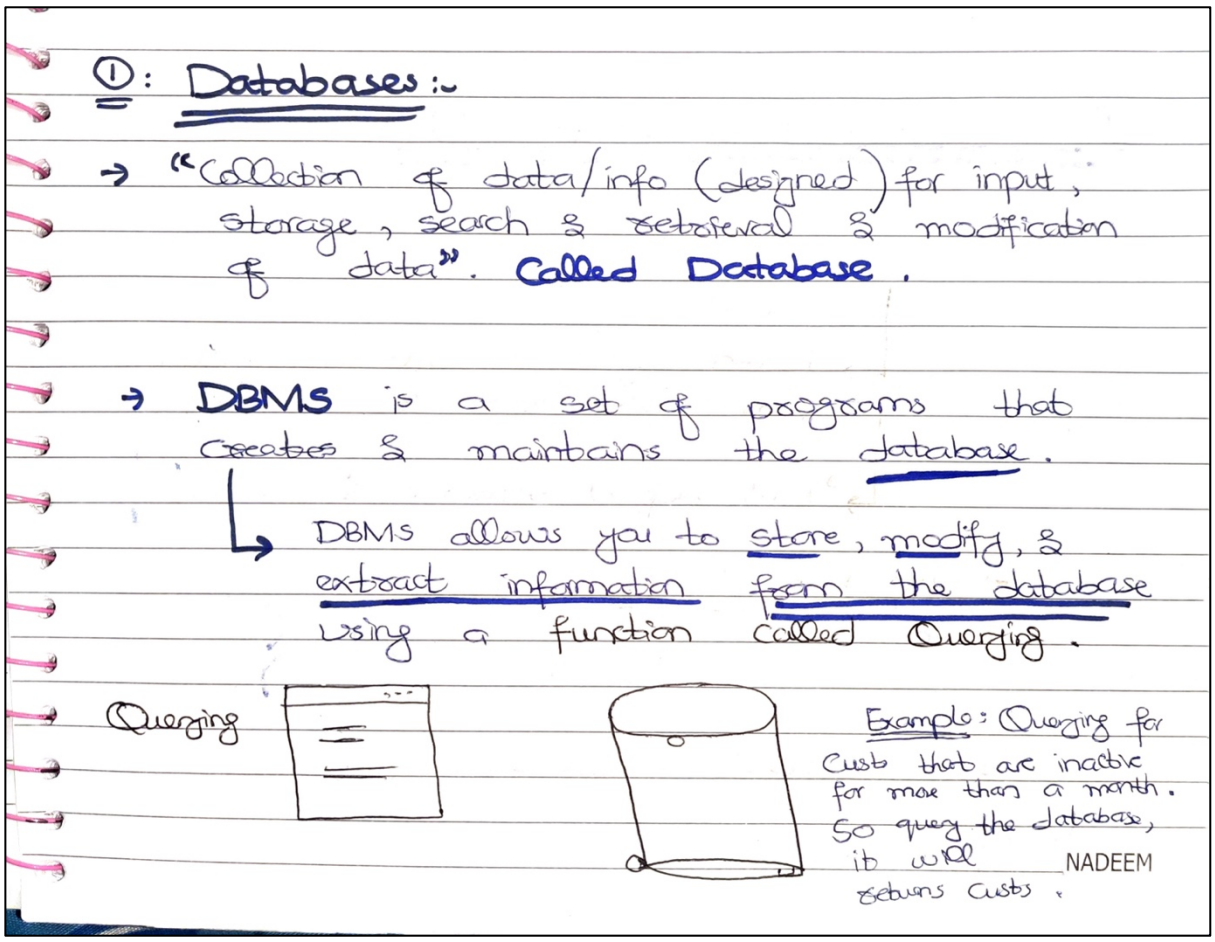
V. Improved Analytics: Data lakes allow for the use of advanced analytics and machine learning algorithms to uncover insights that might otherwise be hidden in large and complex data sets.

VI. Collaboration: Data lakes can be accessed by multiple users and teams, allowing for collaboration and the sharing of data and insights across the organization.

3.DATABASE:

- Data repository is a general term used to refer to data that has been collected, organized, and isolated, so that it can be used for:
    - Business Operation, or
    - Mined for reporting & data analysis.
- It can be a small or large database infrastructure with one or more databases, that:
    - Collect,
    - Manage, &
    - Store data sets.

- So, there are different type of repositories, that may your data reside-in. It includes Databases.



Date: 9/7/22

M T W T F S S
☐ ☐ ☐ ☐ ☐ ☑ ☐

- Factors governing choice of database include:
  - Data Type
  - Data structure
  - Querying mechanisms
  - Latency requirements
  - Transaction speeds
  - Intended use of data

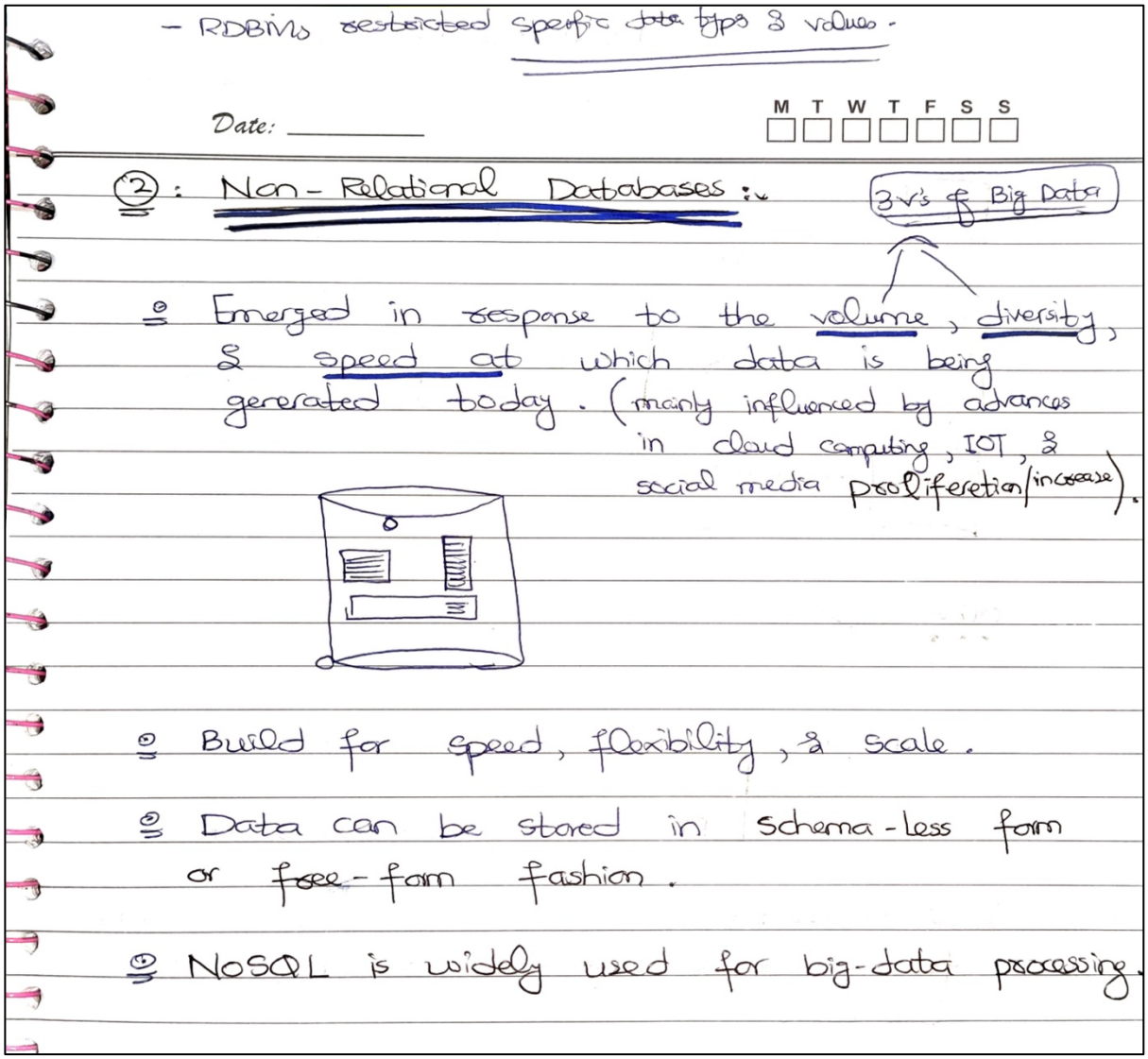
- There are two main types of Databases:

(1) Relational Databases (RDBMS):

- RDBMS builds on Organizational principles of flat files:
  - Similarity b/w RDBMS & flat fils
    - Data organized into a tabular format with rows & columns.
    - well-defined structure & schema.
  - Diff
    - But unlike Flat fils, RDBMS are optimized for data operations & querying (involving many table & larger volumes).
    - SQL used as standard query language.

- Each table has unique set of rows & cols.
- Relationships defined b/w tables (minize data redundancy).

NADEEM

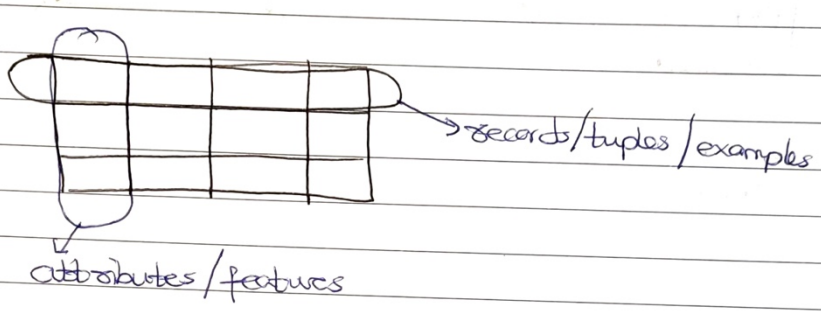


- This was short introduction of relational and non-relational databases. I have explained in details about the Relational and Non-Relational Databases below:

Date: _____ M T W T F S S

Topic :~ RDBMS

A relational database is a collection of data organized into a table structure, where the tables can be linked, or related, based on data common to each.



Example :~

- Let's take example of "Customer Table" that maintains data about each customer in company.

- Attributes in Cust_Table : (CusID, Company Name, ~~Company~~ Cust_Address, Cust_Phone)

CID	CN	CA	CP
1	Ali	---	---
2	Musa	--	--

(each customer record.

⊛: Now what do we meant by linked / related :
So along with "Customer Table", company also maintains "Transaction Table" that describes "multiple individual transactions pertaining to each customer." ~~rel~~ related

NADEEM

Date: _____ M T W T F S S

Trans Date	Cust_ID	Trans Amount	Payment
–	1	–	–
–	2	–	–

→ Customer table & Transaction Table are related to each other, based on common "Cust_ID".

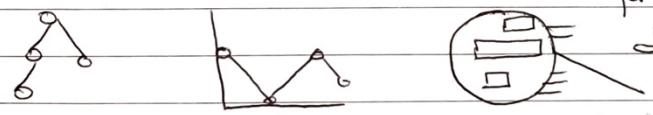
→ We can query the customer table to produce reports such as customer statement that consolidates all transactions in a given period."

(consolidates: to join together into a whole one)

↑

This capability of relating tables based on common data enables you to retrieve an entirely new table from data in one or more tables with a single query.

It allows you understand relationship among all available data & gain new insights for making better decisions.



NADEEM

Date: _____ M T W T F S S

Examples of RDBMs :-

- Open-source with internal support.
- Open-source with commercial support.
- Commercial closed-source.

IBM DB2 SQL Server MySQL Oracle Database Postgres

are some popular relational databases.

- Cloud-Based Relational Databases, or Database-as-Service :

Amazon RDB Google SQL IBM DB2 cloud Oracle Cloud SQL Azure

→ these are gaining popularity because of limitless computing and storage capabilities offered by cloud.

+ : Advantages :

- Create meaningful information by joining tables.
- Flexibility to make changes while the database is in use.

NADEEM

Date: _____ M T W T F S S

- Minimize data redundancy by allowing relationships to be defined b/w tables.

- Offer export & import options that provide ease of backup & disaster recovery. (Exports happening while DB is running, making restore on failure easy).

→ While Cloud-based databases do continuous mirroring, which means the loss of data on restore can be measured in seconds / less.

- Are Acid Compliant, ensuring that the data in the database remains accurate and consistent despite failures, & database transactions are processed reliably.

: Use-Cases :

Relational Databases are well-suited for :

Data Warehouse :

RDBMS can be optimized for OLAP (where historical data is analyzed for business intelligence)

IOT Solutions :

Requires the speed & ability to collect and process data from edge devices.

Date: _____

M T W T F S S

⊕: Limitations :

- Does not work well with semi-structured & unstructured data.
- Cannot done extensive analytics on data.
- Migration possible b/w RDBMS when the source & destination tables have identical schemas & data types.
- Limit on data field & length, and may lead to loss of information.

NADEEM

Schema : how data is organized within a relational database ; It is a logical structures of data.

Date: _____ M T W T F S S

NO SQL

Topic : NOSQL : (not only SQL)

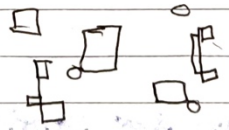
NoSQL is a non-relational database design that provides flexible schemas for the storage & retrieval of data.

- Built for specific data models
- Gained greater popularity due to the emergence of cloud computing, big data, & high-volume web & mobile applications.
- chosen for their attributes around scale, performance & ease of use.
- Has flexible schemas that allow programmers to create & manage modern applications.
- SQL is not support to query data, although some may support it.

(structured, semi-structured, unstructured)

NoSQL allows data to be stored in a schema-less or free-form fashion.

e.g: MongoDB is considered schemaless bcz it does not require a rigid, pre-defined schema.



NADEEM

Date: _____ M T W T F S S

Types of NoSQL Databases :-

- Based on "model being used for storing data", there are four common types of NoSQL databases :

  (i) Key-value Store
  (ii) Document-Based
  (iii) Column Based
  (iv) Graph Based

(i) Key Value Store :-

- Data in Key-value database is stored as a collection of key-value pairs.
- A Key represents an attribute of the data & is a unique identifier.
- Both Keys & values can be anything from simple integers or strings to complex JSON documents.
- **Great for storing** user session data, user preferences, real-time recommendations, targeted advertising, in-memory data caching.

Date: _____ M T W T F S S

⊛: Not a Great fit if you want to:

- Query data on specific data value.
- Need relationships b/w data values.
- Need multiple unique keys.

Redis, Memcached, DynamoDB are key-value based NoSQL.

(ii): **Document-Based** :-

Document databases store each record and its associated data within a single document.

- They enable flexible indexing, powerful ad-hoc queries, & analytics over collections of documents.

= **Preferred for** eCommerce platforms, medical records storage, CRM platforms, & analytics platform.

⊛: Not a great fit if you want to:

- Run complex search queries.
- Perform multi-operation transactions.

NADEEM

Date: _____ elasticsearch ↑ (Document-oriented) M T W T F S S

- MongoDB, DocumentDB, CouchDB, Cloudant are Document-based NOSQL.

(iii): Column-Based :-

Column-based models store data in cells grouped as columns of data instead of rows.

(columns that are usually accessed together)

- A logical grouping of columns is referred to as a column family.

Example:

A customer's name & profile information will most likely to be accessed together (and grouped into column-family).

- All cells corresponding to a column are saved as a continuous disk entry, making access & search easier & faster.

- Great for systems that require heavy write requests, storing time-series data, weather data, & IOT data.

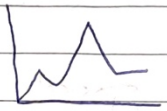
NADEEM

Date: _____ M T W T F S S

- Not a great fit if you want to:
  - Run complex queries
  - Change querying patterns frequently.

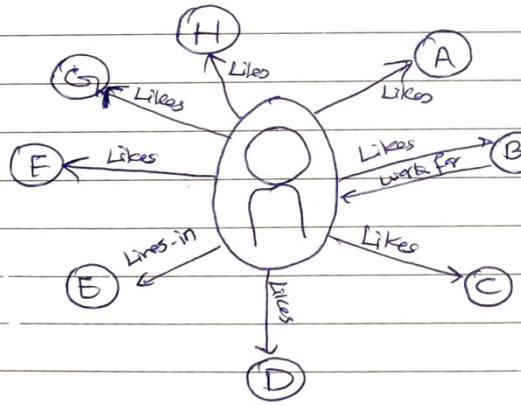
- Cassandra, Apache Hbase are Column-based NOSQL.

(iv): Graph-Based :-



- Graph-based databases use a graphical model to represent & store data.

- Useful for visualizing, analyzing, & finding connections b/w different pieces of data.



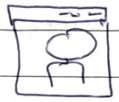
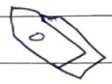
- The circle are nodes, that contain the data.

- The arrows show relationships.

NADEEM

Date: _____ M T W T F S S

- **Great choice** for working with connected data. (which is data that contains lots of interconnected relationships.)

- **Great for**  ,  ,

  Social Networks — Real-time Product Recommendation

  network diagrams, fraud detection, & access management.

- Not a great fit if you want to:

  - Process high volume of transactions

    because graph databases are not optimized for large-volume analytics queries.

  - Neo4j, CosmosDB are Graph-Based NoSQL.

NADEEM

Date: _____

M T W T F S S
☐☐☐☐☐☐☐

Advantages :-

- Ability to handle large volume of data of any structure.
- Ability to run as distributed system scaled across multiple data centers.
- Efficient & cost-effective scale-out architecture that provides additional capacity & performance with the addition of new nodes.
- Simpler design, better control over availability, & improved scalability that makes it agile, flexible, & to iterate more quickly.

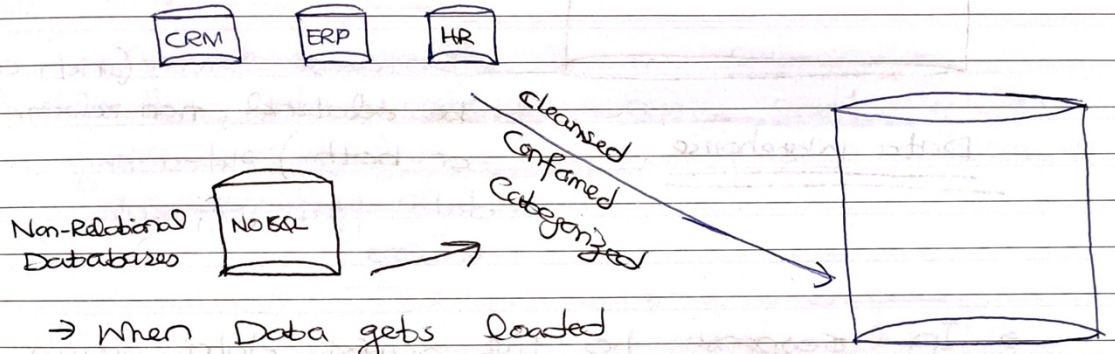
4.DATA WAREHOUSE:

Date: _____ M T W T F S S

① : Data Warehouse :-

- A data warehouse is a central repository of data integrated from multiple sources (consolidated through ETL).

- Data warehouse stores current & historical data that has been cleansed, conformed, & categorized.



→ When Data gets loaded into the data warehouse, it is already modeled & structured for a specific purpose e.g; it is analysis-ready.

- Traditionally, data warehouse are known to store relational data from transactional systems & operation databases (CRM, ERP, HR etc)

- But with emergence of Big Data & NOSQL, Non-Relational data repositories are also used for data warehousing.

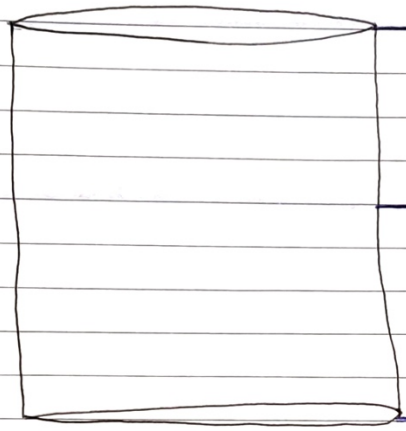
- You may have read the last sentence of the above picture that we can use non-relational databases as Data Warehouse because of the emergence of Big Data and NOSQL data. **So yes, with the emergence of Big Data and NoSQL databases, non-relational data repositories have become popular for data warehousing**. Here are some examples of non-relational databases that can be used as data warehouses:
    - **Hadoop Distributed File System (HDFS):** HDFS is a distributed file system that is used to store large volumes of data across multiple nodes in a cluster. Hadoop also includes tools like MapReduce and Hive, which can be used to process and analyse data stored in HDFS.
    - **Apache Cassandra:** Cassandra is a highly scalable NoSQL database that is designed to handle large volumes of data across multiple data centres. It is often used for real-time data analysis and is popular in the finance, healthcare, and telecommunications industries.
    - **Amazon S3:** Amazon S3 is a cloud-based object storage service that is used to store and retrieve large amounts of data. It is often used as a data lake, where data from various sources is stored for later analysis.

- **MongoDB:** MongoDB is a document-oriented NoSQL database that is designed to store and retrieve unstructured data. It is often used for real-time analytics and is popular in the gaming, healthcare, and media industries.
- **Apache HBase:** HBase is a distributed, column-oriented NoSQL database that is designed to handle large volumes of sparse data. It is often used in combination with Hadoop for real-time data processing and analysis.

Date: 10/7/22

M T W T F S S
☐☐☐☐☐☐☑

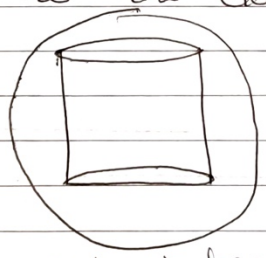
Q: A Data Warehouse has a 3-tier Architecture:



- Client front-end layer (querying, reporting, & analyzing data)
- OLAP Server (process & analyze information coming from database servers)
- Database Servers (which can be relational, non-relational, or both) extracting data from different sources.

Data Warehouse

→ In response to the rapid data growth & today's sophisticated analytics tools, data warehouses that once resided in on-premise data centers are moving to the cloud.



Data Warehouse

Q: Benefits of Cloud-Based Data Warehouse:

- Lower cost
- Limitless storage & compute capabilities.
- Scale on a pay-as-you-go basis.

NADEEM

⊛: On-premise DWH: With this, the organization must purchase, deploy, & maintain all hardware, & software.

Date: _____ M T W T F S S

- Faster disaster recovery.

⊛: Some of on-cloud DWH solutions:

- teradata, Orcale Exadata, IBM DB2 on cloud, Amazon Redshift, Google BigQuery, cloudera, snowflake.