Mikateko Mpapele

DAT: Assignment1

Section A: Database Fundamentals

1. Main Types of Databases: Relational Database and NoSQL Database.

. Relational databases comprise of structured, tabular schema, Fixed data types. No SQL is schema-less, accommodater a variety of data types.

2 Relational Database Management System is a program used to create, update, and manage relational databases; which organizes data into tables with relationships (commonality) between them.

3. Primary Key: a primary key is a unique identifier for each record in a table.

Foreign Key: A column that highlights a relationship between tables by referring to a primary key in another table.

4. Database Normalization - is the process of organizing a ~~relationsh~~ relational database to reduce data redundancy and improve data integrity by structuring tables and linking them with relationships. This is done by breaking down a larger table into smaller, more manageable ones that are ~~connected~~ connected by primary and foreign keys.

- Normalization is an important in database design as it helps improve the databases' integrity, efficiency, consistency and accuracy.

Database schema: is the logical structure of a database and it defines how data is organized, structured and related.

6. Structured Data	Semi- Structured	Unstructured
* Information that is organized and stored in rows and columns, making it easy for both human and machine to access, search and analyze	· Information that contains organizational properties but lacks the rigid structure of a relational database.	· information without a pre-defined data model. · Lacks structured like a spreadsheet with rows and columns. ex: Word Document text, audio/video Files

7. Fact Table - contains measurable quantitative data about business processes.

Dimension Table - contains descriptive information which describes the Facts table.

8. Data model - is a visual representation that organizes data and shows the relationship between different data elements. in a database.

- it ~~allows~~ helps to clarify data requirements, ensure consistency.
- shows how different pieces of data are connected, making them earier to ~~use~~ find and use.

9. Database - an organized collection of structured data stored electronically, typically managed by a database management system (DBMS) for efficient access, organization and modification.

Datawarehouse : central warehouse/storage of multiple databases.

- They store both current and historical data from multiple sources.
- It stores clean, structured data for optimized reporting, business intelligence and analytics.

Data Lake : store massive amounts of raw, unstructured data For ~~explorable~~ data science and machine learning.

10. DataMart : is focused, smaller subset of a data warehouse - departmental or subject specific.

Datawarehouse - the large, central repository of enterprise-wide data from multiple sources.

Section B : SQL and Data Processing

11. Query Language : A query language is a computer programming language used to retrieve data and manipulate data from databases.

SQL is the most commonly used ~~query~~ programming language for storing and processing information ~~in~~ a relational database. SQL is designed to handle large volumes of data and can scale to meet business demands/needs.

- It is a standardized language
- Efficient Database Management - it provides structured, relational tables with well-defined schemas.

12. Indexes in databases : an index is a data structure that speeds up data retrieval ~~of data~~ operations by creating a sorted list with pointers to the location of data ~~in~~ a table.

13. Transactions in databases ; A single unit of work that groups multiple operations on database.

ACID properties :

Atomicity - each transaction is treated as a single, indivisible unit.

Consistency - Guarantees that a transaction brings the database from one valid state to another.

Isolation - Concurrent transactions do not interfere with each other.

Durability - One a transaction is committed, changes a permanently stored and will survive subsequent system failures.

14. Database Engine is a core software that handles data storage, retheval, and manipulation for a database management system. It performs CRUD operations (Create, Read, Update, Delete), and manages the physical storage of data on disk and ensures data integrity, security and efficient access by executing SQL commands.

15. Views, Stored Procedures and Triggers in SQL:

* Views : virtual tables whose content is defined by a SQL query.

* Stored Procedure: A precompiled collection of one or more SQL statements and optional control-of-Flow statements.

* Triggers - a special typeof stored procedure that automatically executes in response to certain events occurring in a database.

16. ETL : process data by transforming it on a separate server before loading it into a target system.

ELT : extracts raw data, loads it directly into a data warehouse and target system like a cloud data warehouse and then performs transformations on-demand, making it Faster, more Flexible and better for handling large volumes of diverse, including unstructured data.

17. Batch processing: a method of running a group of computer tasks together to be processing automatically, typically in high-volume, repetitive scenarios, at scheduled intervals.

Stream Processing:

* a continuos method of ingesting, analyzing and processing data as it is generated.

18. SQL Join : used to combine rows and columns ~~of~~ from two or more tables, based on a related column between them.

Types: Left Join, Right Join, Full outer Join, Inner Join

~~Left~~ Inner Join :

```
SELECT product_id, product_name, category_name
FROM products
INNER JOIN Categories ON products.categoryID =
                         categories.categoryID;
```

Left Join :

```
SELECT product_id, product_name, category_name
FROM products
Left Join Categories ON products.categoryID =
                        categories.categoryID;
```

Right Join :

```
SELECT product-id, product.name, categoryName
FROM products
  RIGHT JOIN categories ON Product. CategoryID =
                         categories. CategoryID;
```

19. Referential Integrity refers to the relationship between tables tablos. It ensures that relationships between tables remain valid by guaranteeing that a Foreign key value in a child table always refers to a valid primary key value in a parent table.
It ensures data accuracy, reliability and data consistency.

20. Data redundancy - storage of the same information in multiple places - as a strategy for data backup and recovery or ~~unit~~ unintentional poor data managed.

Section C : Data Management and Analytics Concept.

21. * Cloud Database Management is the process of creating, maintaing and securing database on cloud platfor

* On-premise solutions are hosted on a company's own servers and managed internally, providing greater control but requires upfront investment and ongoing maintenance.

22. Data governance and Why it is important in data management :

- Data governance: set of policies, processes and roles that manage and control an organization's data throughout its life cycle to ensure accuracy, security and responsible use.

Why Data Governance matters :

* Ensure data is accurate and consistent
* Protects sensitive information and ensures compliance with regulations and privacy standards.
* Improved decision making

23. Data Integrity - means that data is accurates complete, consistent and reliable throughout its entire lifecycle, from creation to deletion.

Steps to maintain data integrity:

~~24~~
* Data validation: Processes are put in place to ensure data meets specific rules and standards when it is entered.
* Access controls and authentication: Implement measures to ensure only authorized users can access and modify data.
* Databackup + recovery: Procedures for creating copies of data and restoring it in case of loss or corruption.
* Data governance: implement policies and procedures that define how data should be managed. throughout its lifecycle.

24. Data quality: is a measure of how well a dataset is fit For its intended purpose based on criteria like accuracy, completeness, consistency, validity, uniqueness and timeliness.

- it impacts the accuracy and reliability of information used for decision-making.

25. A data analyst manages and analyzes database~~s~~ information by collecting, cleaning and interpreting data to find trends and provide actionable insights for decision-making.

26. A DBA is responsible for maintaing, securing and operating databases and ensures design consistency, quality, security and compliance with rules and regulations.

27. Design a data pipeline:

1) ~~Identify~~ Define objectives and requirements.
2) Identify data sources
3) Choose the ingestion method

4) Plan data transformations. - cleaning, enrichment, formatting and pattern.

5) Select a storage solution (e.g warehouse, data lake)

6) Design and build the pipeline.
   - Choose tools (eg Apache AirFlow)
   - Develop the architecture
   - Develop and test
   - Implement security.

7) Implement monitoring and maintenance.

28. Common challenges in managing large-scale data

①. Performance and scalability
   - databases become slow when data volume increases

- Backup and Recovery of large databases are slow and complex

- Databreaches and unauthorized access of data.

- Failure to adhere to relevant data privacy and security regulations.

- Misconfiguration is a constant risk and can lead to significant performance and security problems.

29. Popular Database Platforms:

- Relational Databases (MySQL and PostgreSQL) for web applications and financial systems.

- NoSQL Databases like MongoDB for high-volume data storage and in-memory database like Redis for caching and real-time applications.

- MS Server SQL and Oracle are used for enterprise-level options

- Elastic Search - for search and analytics.

30.

Row oriented storage formats

*CSV and JSON - text-based formats are often used for ingesting raw data into a platform.

·Avro - raw-based format designed for data serialization and exchange.