Research Assignment 1

* Section A

(1) The main types of databases are structured and unstructured data

(2) A Relational Database Management System is a type of database management system that stores and manages data in a structured and organized way, using tables, rows and column.
NB It is based on the relational model, which means that data is stored in tables with well-defined relationships between them

(3) A Primary key- is a column or set of columns in a table that uniquely identifies each row in the table. Its a unique identifier for each record, and it cannot be NULL or duplicated.

A foreign key is a column or set of columns in a table that references the primary key of another table. It is used to establish relationships between tables and ensure data consistancy

(4) Data normalization is the process of organising data in a database to minimize data redundancy and dependency

It is important because it improves data quality, reduces data anomalies, supports data

security, improves database performance

(5) A Schema is the overall structure or organization of a database, including the relationships between different table, columns, and data types.

(6) The difference between these 3 types of data sets is that:

* Structured data
- is highly organized
- Fits into a schema
- is easily searchable
- eg Relational database (customer information, financial transactions), Spreadsheets and CSV files

* Semi structured data
- Partially organized
- Flexible schema
- Self-describing
- eg. JSON (Java Script Object Notation) XML (Extensible Markup language), CSV files with variable columns and key-value stores.

* Unstructured data
- lacks organization
- No schema
- Human readable
- eg. Text documents (eg emails, articles, social media posts, images, videos, audio files and sensor data

(7) The difference between a fact table and a dimension table is that a fact table measures business performance, it is also central to the datawarehouse, it is as well typically large. forexample Sales, orders, website interactions, customer transactions.
While a dimension table is a table that provides context, describes attributes and is typically smaller. for example date, time, customer, product, location, product catergory

(8) A data model in a conceptual representation of the structure and relationships of data in a database. Its a blueprint or a map that defines how data is organised, stored, and related
It is important because it:
- improves data intergrity (ensures data consistancy and acuracy)
- Supports scalability (enable easy modification and extension of the database)
- facilitates communication
- Reduces errors.

(9) The difference between the 3 is that a database:
- stores structured data
- Optimized for transactions
- Schema driven
- Eg MySQL, Postgre SQL, Microsoft SQL server.

Datawarehouse:
- stores historical data
- Optimized for analytics
- schema-on-read (datawarehouse often use a schema-on-read approach, where the schema is applied when the data is queried.
  Eg: Amazon Redshift, google BigQuery, Snowflake.

Data lake
- stores raw, unprocessed data like videos, texts and images.
- Schema-on ~~write~~-write
- Flexible and scallable
  e.g hadoop Distributed file System (HDFS) Amazon S3, Azure Data lake Storage

(10) A data mart is a storage solution which:
- subset of data (is a subset of a datawarehouse that focusses on a specific business area or department
- Departmental
- Contains simplified data, making it easier for users to access and analyze
- faster query performance

* The key differences are:
- Scope (datawarehouse (enterprise wide)

vs Data Marts (departmental or fuctional))
- Data complexity (data warehouse (complex data) vs Data Mart (simplified data)
- purpose (Data warehouse (strategic entrepris
- wide analytics) vs Data Mart (tactical departmental analytics)
- Design (Data warehouse (designed for flexibil and scalability) vs Data Mart (designed for fast query performance and ease of use)

Section B (SQL and Data Processing)

(11) A query language is a programming languag designed to manage and manipulate data stored in a database management system (DBMS). its used to retrieve, update, and manipulate data in a database.

SQL is a commonly used because it offers
- standardization
- Portability
- Flexibility
- Easy to use
- Industry adoption
- Data security
- Improved data management
- Data analysis

(12) Indexes in databases are data structures that improve the speed of data retrieval operations, such as SELECT, INSERT, UPDATE and DELETE. An index is a data structure that facilitates quick lookup and access to data in a table.

They improve performance by:
- faster query performance. (reduces number of rows to be scanned).
- Improved data retrieval (quickly locate specific data, reducing the time it takes to retrieve data)
- Reduced I/O operations (minimize the number of disk I/O operations, improving overall system performance)

(13) Transactions in databases are sequences of operations that are executed as a single, all-or-nothing unit of work. They ensure data consistency and integrity by guaranteeing that either all changes are committed or none are, maintaining the database in a consistent state

ACID (Atomicity, Consistency, Isolation, Durability) properties are a set of standards that ensure database transactions are processed reliably

(14) A database engine, also known as a database management system (DBMS), is a software component that manages and interacts with a database. It is responsible for storing, retrieving, and manipulating data in a database.

It impacts performance by:
- Query performance
- Data retrieval speed
- Concurrency
- Scalability

(15) * Views are virtual tables, dynamic data, simplified queries.
* Stored procedures are precompiled SQL, modularity, Improve performance.
* Triggers are automated actions that occurs in response to ~~specific~~ specific database events data integrity and auditing.

(16) key differences
* Tranfomation location: ETL transforms data before loading, while ELT transforms data after loading
* Data processing: ETL processes data in a seperate environment, while ELT leverages the target system's processing capabilities
* Data storage: ETL typically requires a staging area for tranformation, while ELT stores raw data in the target system

(17) *Batch Processing is a type of an approach to processing data it includes:
- Batching data
- Scheduled processing
- high-throughput

While
* Stream processing includes
- real time processing
- continuous processing
- low latency

(18) A join is a clause used to combine rows from two or more tables based on a related column between them.

Types of joins

Inner join
Left join
full outer join
cross join
Self join

(19) Referencial intergrity is a database concept that ensures the relationships between tables in a relational database are consist and valid

It is important to databases because
(1) Data consistency

(2) Data accuracy
(3) Prevents ophaned records
(4) Improves data quality

(20) Data redundancy affects database perfomance by
- increased storage requirements
- Data inconsistancy (redundant data can lead to inconsistencies if not properly syncronized causing errors and inaccuracies.
- slower query perfomance
- Update anomalies.

Data redundancy affects database storage by
- Increased storage capacity
- Data duplication
- Data fragmentation.

* Section C

(21)
Cloud and on-site (premise) database management differs in the following ways:

- Location: cloud databases are hosted on cloud servers and accessed via the internet, while on-premise databases are hosted on local servers within an organization's physical premise.

- Scalability: cloud databases offer easy scalability, allowing businesses to quickly adjust resources to meet changing demands. On-site (premise) databases require manual upgrades and hardware purchases to scale

- Cost : cloud databases operate on a pay-as-you-go model, reducing upfront costs. On-premise databases require significant upfront investment in hardware, software, and IT staff

- Accessibility : cloud ~~datab~~ databases enable remote access and global collaboration, where-as on-premise databases are typically limited to local networks or VPNs

- Maintenance : cloud providers manage IT maintenance and updates freeing up IT resources. On-premise databases require in-house IT teams for maintenence, backups, and updates.

(22) Data Governance is the process of establishing and enforcing policies, procedures and standards for the effective and efficient use of data across an organization. It ensures that data is accurate, complete, consistent, and secure

IT is important because it:
- improve data quality
- Reduced Risk
- Increased Efficiency
- Better Decision making
- Regulatory Compliance

(23) Data integrity refers to the accuracy, completeness, and consistency of data throughout its entire ~~life force~~ lifecycle. It ensures that data is reliable, trustworthy and consistent across different systems and applications

It can be maintained by:
- Data validation
- Data normalization
- Data Quality Checks
- Access Controls
- backup and Recovery
- Data standardization
- Data governance

(24) Data quality is the accuracy, completeness, consistency and reliability of data. High-quality data is fit for its intented use and meets the requirements of the organizational

It is critical for analytics because it helps:
- With informed decision making
- with Accurate insights
- with business performance
- with regulatory compliance
- with operational efficienty

(25) A Data Analyst s role in database management and analysis is helping organizations make informed decisions. Is:

- Data collection
- Data cleaning
- Data organization
- Data modeling
- Data visualization
- Insight generation
- Reporting
- Storytelling
- Recomendations

(26) key responsibilities of a DBA are:

- Database design
- Database performence
- Data security
- Data back up and recovery
- Database maintenence
- trouble shooting
- Capacity planning
- Data Migration
- Documentation
- Collaboration

(27) The main steps involved in desrgning a pipeline are:

① Define requirements
② Data ingestion
③ Data processing
④ Data storage
⑤ Data output

*General pipeline steps
① Define pipeline purpose
② Determine pipeline requirements
③ Design pipeline architecture
④ Choose tools and technologies
⑤ Develop pipeline code
⑥ Test and validate
⑦ Deploy and Monitor

(28) Common challenges in Managing large scale data:
① Scalability - handling increasing amounts of data and user traffic.
② Perfomance - ensuring optimal database perfomance, query optimization and indexing
③ Data security - Protecting sensitive data from unauthorized access, breaches, cyber threats
④ Data integrity - ensuring data accuracy, completeness, and consistency across the database
⑤ ~~Bac~~ Data backup and Recovery - ensuring data availability and recoverability in case of failures or disasters

(6) Data Management - managing large amount of data, including data storage, archiving and purging

(7) Query Optimization - optimizing complex queries to improve performance and reduce latency.

(8) Concurrency Control - Managing multiple users and transactions to prevent data inconsistencies

(9) Data intergration - integrating data from multiple sources, formats and systems

(10) Cost Management - Managing costs associated with database infrastructure, maintenance and personnel

(29) Popular Database Platforms

(a) MySQL -> web applications, e-commerce platforms, and content-management systems.

(b) PostgreSQL -> Enterprice applications, data warehousing and Analytics

(c) Microsoft SQL -> Enterprise applications, business intelligence, data analytics

(d) Amazon Aurora -> Cloud native relational databases for MySQL, SQL server, Postgre SQL

(e) Snowflake

(f) Oracle

Use cases

(1) Web Applications → Relational databases like MySQL and PostgreSQL for web applications

(2) Big data Analytics → NOSQL databases like MongoDB and Cassandra for big data storage and analytics

(3) Real-time data processing → In-memory databases like Redis for real-time data processing and ~~cat~~ caching.

(4) ~~IOT~~ IOT Data storage → time series databases like InfluxDB for IOT data storage and analytics

(5) Artificial Intelligence and Machine learning → Graph databases like Neo4j for knowledge graphs and recommendation engines

(6) E-commerce platforms → Relational databases like MySQL for ecommerce platforms

(7) Content Management Systems → relational databases like MySQL and PostgreSQL for content management systems

(30) main data storage formats used in Analytics

(1) ~~CSV~~ Structured data formats

- CSV (comma seperated values): plain text format for tabular data

- JSON (JavaScript Object Notation) - a light weight, human-readable format for data interchage
- Avro - a binary format for big data storage and processing

(2) Columnar Storage formats
- Parquet - a columnar storage format for big data analytics
- ORC - ~~(Optimized)~~ a columnar storage format for hadoop

(3) Specialised data formats
- time-series data formats - formats like influx DB's line protocol and TSDB's data format

(4) Semi structured data format
- XML - a markup language for storing and transporting data

(5) Unstructured data formats.
- text files.