Research Assignment 1:

1. What are the main types of databases?
   - Relational Databases (RDBMS) - structured data in tables
   - NoSQL Databases - non-relational, suitable for semi-structured or unstructured data
   - Time-series Databases - optimised for time-stamped data
   - Object-oriented Databases - store data as objects
   - Graph Databases - represent data in nodes and edges

2. What is a Relational Database Management System (RDBMS)?
   - An RDBMS is software used to create, manage and interact with relational databases.
   - It stores data in rows and columns and uses SQL for querying.

3. What is a primary key and a foreign key in a database
   - primary key: uniquely identifies each row in a table
   - foreign key: is a field in one table that links to the primary key in another table establishing a relationship between tables

4. What is database normalisation
   - normalisation is the process of organising data to reduce redundancy and improve integrity.
   - it's important because it ensures efficient data storage and simplifies maintenance

5. What is a database schema?
   - a database schema is the blueprint of a database structure. It defines tables, columns, relationships and constraints, acting as a framework for data organisation

6. Differentiate between structured, semi-structured and unstructured data

Structured: tabular format with rows and columns eg SQL database

Semi-structured: lacks a rigid schema but has tags or markers eg JSON

Unstructured: No predefined structure eg PDF's, images

7. What is the difference between a Fact Table and a Dimension Table in a data warehouse

- Fact table: contains quantitative data for analysis (eg sales)
- Dimension Table: contains descriptive attributes (eg customer_name)
- Fact tables reference dimension tables via foreign keys

8. What is a data model, and why is it important in database design?

- a data model defines how data is structured, stored and accessed.
- It's crucial for ensuring logical relationships, data consistency, and aligning technical implementation with business needs

9. Explain between a database, a data warehouse and a data lake

Database: stores current, structured operational data

Data warehouse: central repository for structured, historical analytical data

10. What is a data mart and how does it differ from a data warehouse:

- a data mart is a smaller, subject-specific version of a data warehouse tailored to a department (eg HR)
- It differs in scope and scale ~~data warehouses are enterprise-wide~~ ~~data marts are focused~~

Section B : SQL and Data Processing

11. What is a query language ~~retrieves~~ and why is SQL the most commonly used?

- a query language retrieves and manipulates data in databases
- SQL (Structured Query Language) is the standard because its simple, widely supported, and powerful for managing relational data

12. What are indexes in databases, and how do they improve performance?

- indexes are data structures that improve query speed by allowing quick data lookups
- They reduce the need to scan entire tables, similar to how a book index works.

13. What are transactions in databases, and what are the ACID properties

- Transaction: a transaction is a unit of work performed against a database
- ACID ensures reliability
  - A - tomically : All or nothing
  - C - onsistency : Valid state before/after
  - I - solation : Transactions don't interfere
  - D - urability : Changes persists after commit

14. What is a database engine, and how does it impact performance?

- A database engine is the ~~one~~ service for storing, processing, and securing data.
- Performance varies by engine type (e.g. MyISAM in MySQL vs InnoDB), affecting speed, concurrency, and reliability.

15. What are views, stored procedures, and triggers in SQL?
View - a virtual table from a query
Stored procedure: pre compiled SQL code that performs actions.
Trigger: An automatic response to events (e.g. insert/update) on a table.

16 Differentiate between ETL (extract, transform, load) and ELT (extract, load, transform)
ETL - data is transformed before loading into the destination
ELT - data is loaded first then transformed - suited for cloud and big data systems.

17. Differentiate between batch processing and stream processing in data pipelines:
Batch - processes data in chunks at intervals - good for historical data
Stream - Processes real-time data continuously - ideal for time-sensitive use cases.

18. Explain what a join is in SQL and list different types of joins with examples:
A join combines rows from 2 or more tables based on related columns.

INNER join - matches in both tables
LEFT Join - All from left, matches from right
RIGHT Join - All from right, matches from left
FULL Join - All rows from both tables
CROSS Join - Cartesian product

19. What is referential integrity, and why is it important in relational databases?

- It ensures relationships between tables remain consistent
- eg. a foreign key must match a valid primary key or be null. This prevents orphan records

20. How does data redundancy affect database performance and storage?

- redundancy leads to wasted storage and inconsistent data.
- It slows down performance and complicates updates due to duplication.

Section C: Data Management and Analytics Concepts

21. How does cloud database management differ from on-premises databases?

- Cloud: managed by providers (eg. AWS, Azure) - is scalable, cost-effective and accessible
- On-premise: Requires internal hardware, maintenance and limits scalability.

22 What is data governance, and why is it important in data management

- data governance defines policies and processes to ensure data accuracy, security, and compliance
- It supports decision-making, protects sensitive information, and ensures regulatory compliance.

23 What is data integrity, and how can it be maintained?

- Ensures data is accurate, consistent and trustworthy over time.
- Maintained through constraints, validations, backups and access controls.

24. What is data quality, and why is it critical for analytics?
- refers to the reliability, accuracy and completeness of data
- Poor data quality leads to flawed insights and bad business decisions.

25. Explain the role of a Data Analyst in managing and analyzing database information?
- a data analyst cleans, organizes and interprets data to generate insights
- They use tools like SQL, Excel and BI platforms to support decision-making.

26. What are the key responsibilities of a Database Administrator (DBA)?
- Responsibilities include:
  - Database installation and configuration
  - Backup and recovery
  - Performance tuning
  - Security management
  - User access control

27. What are the main steps involved in designing a data pipeline?
1. Identify data sources
2. Extract data
3. Transform (clean / enrich)
4. Load into destination (warehouse / lake)
5. Monitor and maintain pipeline.

28. What are some common challenges in managing large-scale databases?
- Data volume and velocity
- Performance tuning
- Backup and disaster recovery
- Data consistency and integrity
- Security and access control

2a. What are some popular database platforms (eg MySQL, Snowflake, PostgreSQL, Oracle) and their use cases?

MySQL - web apps, open-source, widely used
PostgreSQL - complex queries, strong for analytics
Oracle - Enterprise-scale, robust security
Snowflake - cloud-native, excellent for analytics and data sharing

3a. What are the main data storage formats used in analytics (e.g. CSV, parquet, JSON, Avro)?

CSV: simple, widely supported, but not space-efficient
Parquet: columnar format, optimized for read-heavy analytics
JSON - semi-structured, used for APIs and web data
Avro - row-based, efficient for serialization in big data