Muranakie Andrew Nzama
Research Assignment

Section A: Database Fundamentals

1) What are the main types of databases?
→ The main types of databases are relational (SQL) and No SQL, which are often considered the two primary categories. Other key types include hierarchical, network, object-oriented and cloud databases. Relational databases store data in structured tables with defined relationships, while no SQL databases offer more flexible models for unstructured or large scale data.

(2) What is a Relational Database Management System
→ A software system that stores and manages data using tables with relationships. Supports SQL, ensures data integrity and follows ACID rules: MySQL, Oracle, SQL Server

(3) What is a primary key and foreign key in a database?
→ Primary Key - Uniquely identifies each record e.g. customer_id
→ Foreign Key - A field in one table that links to a primary key in another table. Used to maintain relationships.

(4) What is database normalization and why is it important?

-> Normalization is the process of organizing data to eliminate redundancy and ensure data consistency.
Benefits: Saves storage, avoids duplication

(4) What is database normalization and why is it important?

(5) What is a database schema?
A blueprint or structure of the database - defines tables, columns, relationships, data types, constraints.

(6) Differentiate between structured, semi-structured and unstructured data
Semi-structured - contains some organizational properties, like tags or markers, but doesn't follow a rigid, fixed schema like structured data.
Structured - Highly organized data with a fixed predefined schema that fits neatly into rows and columns
Unstructured - Data that has no predefined format or schema, making it challenging to search and analyze directly.

(7) What is the difference between a fact table and a dimension table in a data warehouse?
- Fact table - stores measurable business data
- Dimension table - stores descriptive attributes

(8) What is a data model, and why is it important in database design?
→ A visual or logical representation of how data is organized and related.
It ensures clarity, structure, and efficient database design before building the database.

(9) Explain the difference between a database, a data warehouse, and a data lake
- Database → For daily operations (OLTP)
Stores current transactional data

- Data Warehouse → for analytics / reporting (OLAP)
Stores historical + structured data

- Data Lake → Stores raw data (All types: structured, semi, unstructured) for AI / Big data.

(10) What is a data mart, and how does it differ from a data warehouse?
- Data Mart → A smaller, department-specific subset of a data warehouse e.g. only Sales data

- Data Warehouse → Enterprise-wide data for all departments.

Musawakhe Anocen Nzama
Research Assignment

Section B:

(11) What is a query language and why is SQL the most commonly used?
→ A query language is used to retrieve and manipulate data in databases. SQL is the most used because it is standardized, powerful, easy to learn, and works with all major RDBMS.

(12) What are indexes in databases, and how do they improve performance?
→ An index is like a search shortcut. It helps speed up data retrieval (especially Select queries) without scanning the whole table. But too many indexes slow down insert/update operations.

(13) What are ~~indexes~~ transaction in databases, and what are the ACID properties? A transaction is a group of database, ~~and what are the~~ operations that must succeed or fail together.
ACID properties ensure reliability:

- Atomicity - all or nothing
- Consistency - maintain valid data state
- Isolation - no interference between transactions
- Durability - data survives crashes.

(14) What is a database engine, and how does it impact performance? The database engine is the core service that stores, processes, and secure data e.g. Inno DB

(15) What are views, stored procedures, and triggers in SQL?

- View – Virtual table based on a query (for security and simplified access).
- Stored Procedure – Pre-saved SQL code that runs when called (used for automation).
- Trigger → Auto – execute when an INSERT update, DELETE happens (used for auditing and validation).

(16) Differentiate between ETL and ELT?

ETL → transforms data before it's loaded into a data warehouse, while ELT loads raw data first and transforms it later within the warehouse itself. ETL is a more traditional approach that requires upfront planning and is often used for smaller, structured datasets, whereas ELT is a modern approach suited for large, diverse datasets that leverages the scalability of cloud-based warehouses.

(17) Differentiate between batch processing and stream processing in data pipelines.

* Batch Processing – processes data in large chunks at intervals
* Stream Processing – Real time continuous processing

and how does
abase engine is
esses, and

dures, and
query (for
QL code that
omation].
an INSERT
r auditing

ELT?
aded into a
raw data first
e warehouse
pproach that
often used for
ELT is a
iverse datasets
cloud-based

g and stream

ge chunks of

processing

(18) Explain what a join is in SQL and list different types of joins with examples. A Join combines data from multiple tables using related columns.

- Inner Join – Only matching rows
- Left Join – All records from left and matches from right.
- Right Join – All records from right and matches from left.
- Full Outer Join – All records from both tables.
- Cross Join – Every combination of both tables

(19) What is referential integrity, and why is it important in relational databases? It ensures foreign keys correctly reference valid primary keys. Prevent orphan records and maintains accurate relationships

(20) How does data redundancy affect database performance and storage too much duplicate data causes:

- wasted storage space
- Slower performance during updates
- Higher risk inconsistencies

Section C: Data Management and Analytics Concepts

(21) How does cloud database management differ from on-premise databases? Cloud databases are hosted and managed by a third-party provider, offering high scalability, pay-as-you-go pricing, and

, and accessibility from anywhere with an internet connection. On-premise databases are hosted locally, providing greater control and potentially faster performance for localize applications, but require significant upfront investment, manual scaling, and internal managem-ent.

(22) What is data governance, and why is it important in data management? It is the framework for managing data policies, security, access and compliance. Ensures trust, security, and regulatory compliance.

(23) What is data integrity, and how can it be maintained? Data integrity means data is accurate, consistent, and complete. Maintained using validation rules, foreign keys, constraints, and audit controls.

(24) What is data quality, and why is it critical for analytics? Good data quality means clean, accurate, complete, consistent, and reliable data. Bad data = wrong business decisions.

(25) Explain the role of a Data Analyst in managing and analyzing database information?
A Data Analyst extracts, cleans, analyzes, and interprets data to support decision-making, often using SQL, Excel, PowerBI, Python.

Section C: Data Management & Analytics Concepts.

(26) What are the key responsibilities of a DBA?
- Install & Maintain database
- Back up & recovery
- Security & Access Control
- Performance monitoring & tuning
- Ensure uptime & Integrity

(27) What are the main steps involved in designing a data pipeline?
1. Source identification
2. Data extraction
3. Transformation / Cleaning
4. Load to target (Warehouse / Lake)
5. Monitoring & Automation.

(28) What are some common challenges in managing large-scale databases?
- Storage and Scaling issue
- Slow query performance
- Data Security & Privacy
- Backup & disaster recovery
- Maintaining Real time access

(29) What are some popular database platforms & their use cases?
- MySQL
- PostgreSQL
- Oracle
- Snowflake
- Mongo DB

(30) What are the main data storage formats used in analytics?
- CSV - Simple, human-readable
- JSON - Semi-structured
- Parquet - Columnar
- Avro - Compact, good for streaming pipelines