BrightLight Data Analytics

Research Assignment

Section A: Database Fundamentals

1. What are the main types of databases?

>Relational databases (SQL) - store data in tables (e.g., MySQL, PostgreSQL).

>NoSQL databases - designed for unstructured/large-scale data (e.g., MongoDB).

>Cloud databases - hosted on cloud services (e.g., BigQuery).

>Graph database - store relationships (e.g., Neo4j).

>Time-series databases - store time-based data (e.g., InfluxDB).

2. What is a Relational Database Management System (RDBMS)?

>A Relational Database Management System stores data in tables with rows & columns and supports SQL (Structured Query Language).Examples: MySQL, SQL Server, PostgreSQL.

3. What is a primary key and a foreign key in a database?

>Primary Key - a unique identifier for each row in a table.
>Foreign Key - a field that links one table to another table's primary key.

4. What is database normalization and why is it important?

Normalization is the process of organizing data to reduce redundancy (repeated data). It improves:

data consistency, storage efficiency, And database performance

5. What is a database schema?

A schema is the structure of a database-tables, fields, relationships, constraints.

It’s like a blueprint of the database.

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

Type	Example	Description
Structured	SQL tables	Organized in rows/columns
Semi-structured	JSON, XML	Has structure but more flexible
Unstructured	Images, text, videos	No predefined structure

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

Fact Table, contains numerical values (sales amount, quantity).
Dimension Table, contains descriptive attributes (product, customer, date).

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

A data model defines how data is stored, related, and accessed.
It helps, design the database, ensure accuracy, reduce redundancy

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

Type	Purpose
Database	Stores day-to-day transactional data
Data Warehouse	Stores historical, structured data for reporting
Data Lake	Stores raw, structured + unstructured data

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

A data mart is a smaller, topic-specific section of a data warehouse (e.g., sales mart, HR mart).

Section B: SQL and Data Processing

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

A query language allows you to retrieve, insert, update, delete data.
SQL is most used because it is standardized, easy to learn, powerful, and works for all major databases

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

An index is a special structure that makes data retrieval faster—similar to an index in a book.

>Improves, search speed, filtering and sorting

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

A transaction is a group of operations executed together.

ACID:

>Atomicity - all or nothing

>Consistency - rules are followed

>Isolation - transactions don’t interfere

>Durability - results are permanent

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

The **engine** is the program inside the database that:

>reads/writes data

>manages indexes

>handles queries

Examples: InnoDB, MyISAM.

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

>View - a virtual table based on a query.
>Stored Procedure - saved SQL code you can run anytime.
>Trigger - auto-runs when an event happens (INSERT, UPDATE).

16. Differentiate between ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform).

ETL	ELT
Extract → Transform → Load	Extract → Load → Transform
Used in traditional warehouses	Used in modern cloud systems (BigQuery, Snowflake)

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

>Batch - processes data in large groups at once (daily, hourly).
>Stream - processes data continuously in real time.

18. Explain what a join is in SQL and list different types of joins with examples.

A join connects two tables using a common column.

Types:

>INNER JOIN – matching rows

>LEFT JOIN – all from left, matches from right

>RIGHT JOIN – all from right

>FULL JOIN – all rows from both

>CROSS JOIN – combinations

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

>Rules that ensure relationships between tables remain valid.
>Example: a foreign key must match a primary key.

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

Section C: Data Management and Analytics Concepts

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

Cloud Databases

>Hosted on cloud platforms (e.g., BigQuery, Snowflake, AWS RDS).

>No need to manage hardware — the cloud provider handles servers, scaling, backups, and maintenance.

>Automatically scalable (can handle more users or data instantly).

>Pay-as-you-go pricing (you only pay for what you use).

>Accessible from anywhere with internet

>Faster setup and deployment.

On-Premise Databases

>Installed and run on an organization’s own servers.

>Requires physical hardware, electricity, and IT support.

>Scaling requires buying new hardware → slow and expensive.

>Company is responsible for backups, updates, and security.

>Limited remote access unless special configurations are used.

22. What is data governance, and why is it important in data management?23. What is data integrity, and how can it be maintained?

Data governance is the set of policies, standards, and processes that ensure an organization's data is:

>accurate

>secure

>consistent

>accessible

>properly used

Why important:

>Protects sensitive data,

>Ensures data accuracy and trust,

>Helps with compliance (GDPR, POPIA)

>Improves decision-making

>Reduces data misuse or errors

24. What is data quality, and why is it critical for analytics?

Data quality refers to how accurate, complete, consistent, timely, and relevant the data is.

Why important:

>Bad data = incorrect analysis

>Better insights & dashboards

>Reliable business decisions

>Helps build trust in analytics

>Reduces costly errors

25. Explain the role of a Data Analyst in managing and analyzing database information.

database information

A **Data Analyst**:

>Extracts and cleans data from databases

>Uses SQL to query data

>Creates dashboards and reports

>Analyzes trends and patterns

>Ensures data accuracy during analysis

>Communicates insights to stakeholders

>Helps improve business decisions using data

26. What are the key responsibilities of a Database Administrator (DBA)?

A **DBA** manages the database infrastructure.

Responsibilities include:

>Installing, configuring, and upgrading databases

>Performance tuning

>Creating backups & restoring data

>Ensuring security & access control

>Managing users and permissions

>Monitoring system health

>Ensuring high availability and reliability

27. What are the main steps involved in designing a data pipeline?

Source identification (where the data comes from)
Extraction (pulling data from sources)
Transformation (cleaning, formatting, enriching)
Loading (moving data to destination) (warehouse, lake)
Orchestration (scheduling and automating tasks)
Monitoring (checking for failures, performance, quality)
Maintenance (fixing issues, updating pipeline logic)

28. What are some common challenges in managing large-scale databases?

>>Slow query performance
>>Storage and scaling issues
>>High cost of infrastructure
>>Maintaining data quality
>>Ensuring data security

>>Managing millions/billions of rows
>>Backup and disaster recovery
>>Complexity in data integration
>>Schema changes over time

29. What are some popular database platforms (e.g., MySQL, Snowflake, PostgreSQL, Oracle) and their use cases?

MySQL

>>Open-source SQL database

>>Best for web applications, e-commerce, startups

PostgreSQL

>>Advanced open-source relational database

>>Strong for analytics, complex queries, GIS data

Oracle

>>Enterprise-level SQL database

>>Used in banking, telecom, large companies

Snowflake

>>Cloud data warehouse

>>Excellent for big data, BI, ELT processing

Microsoft SQL Server

>>Enterprise SQL database

>>Used in corporate applications and reporting systems

MongoDB

>>NoSQL document-based database

>>Great for flexible, unstructured data

30. What are the main data storage formats used in analytics (e.g., CSV, Parquet, JSON, Avro

CSV

>>Simple, easy to read

>>Used for exports, spreadsheets

>>Not compressed → large file size

JSON

>>Semi-structured format

>>Used in APIs, web applications

Parquet

>>Columnar storage

>>Highly compressed → great for big data

>>Used in Spark, BigQuery, Snowflake

Avro

>>Row-based binary format

>>Good for data streaming (Kafka)

>>Schema evolution support