Nolwazi Ndlovení - Research assignment 1

1. main types of databases are
   (a) Relational databases
   (b) No SQL databases
   (c) Object-oriented databases
   (d) Time-series databases

2. A relational database management system is a software that manages data stored in tables.

3. A primary key is a unique identifier for each record in a table
   A foreign key is a field in one table that refers to the primary key in another table.

4. Database normalisation is the process of organising data in a database to reduce redundancy and improve data integrity. It is important because
   ① Reduce duplication
   ② Makes data easier to maintain
   ③ Improves data accuracy.

5. A database schema is the structure of a database. It defines how data is organised.

6. Structured data - is organised in rows and columns.

Semi structured data - Has some structure but not in a fixed format

Unstructured data - No predefined format or organisation.

7. Fact table stores quantitative data and is used for analysis and reporting.

Dimension table stores descriptive data and is used to provide context to facts.

8. A data model is a conceptual framework that defines how data is structured, stored and accessed in a database. It is important because:

1. Organizes data clearly
2. Improves data quality
3. Enable efficient queries
4. Ensures data integrity

9. Database stores current structured data for day to day operations

Data warehouse - stores large volumes of structured data for analytics and reporting

Data Lake - stores raw data in any format

10. Data Mart is a small part of a larger data warehouse. It contains data focused on one specific area.

Difference between Data Mart and Data Warehouse:
Data warehouse covers all data from the whole company.
Data Mart only covers data for one department

## Section B: SQL and data processing

11. A query language is a special type of computer language used to ask questions and get data from databases.

SQL is most commonly used because

1. Standardized
2. Easy to learn
3. Versatile
4. Widely supported.

12. Index helps the database find data faster without scanning the whole table.

How indexes improve performance:

- Faster search
- Quicker filtering
- Improves sorting

13. A transaction is a group of one or more SQL operations that are executed as a single unit.

ACID stands for
Atomicity - All or nothing
Consistency - Valid state only
Isolation - no interference
Durability - permanent changes

14. A database engine is the core software component that handles how data is stored, retrieved, and managed in a database system.

How does it impacts performance
1. speed of queries
2. Concurrency handling
3. Support for transactions

15. A view is a virtual table based on the result of a SQL query.

Trigger is a set of instructions that automatically runs in response to specific events on a table.

Stored procedure is a saved block of SQL code that you can run anytime.

16 ETL - used when you have a traditional data warehouse

ELT - used when you have a modern data warehouse.

17. Batch processing processes large volumes of data collected over a period

Stream processing processes data in real-time as it arrives.

18 A join in SQL is used to combine rows from two or more tables based on a related column

Types of joins

1. INNER JOIN

Example

```
SELECT customer.name,
       orders.order_id
From customers
INNER JOIN orders
ON customer.customer_id = orders.customer_id;
```

2. Right Join
3. LEFT JOIN
4. Full outer JOIN.

19. Referential Integrity is a rule in relational databases that ensures relationships between tables. Its important because

1. Maintains data consistency
2. Enables safe updates and delete

20. How data redundancy affects performance

1. Slow queries
2. Inconsistent data
3. More complex management.

How it affects storage
1. Wasted disk space
2. Higher costs.

21. Cloud databases are hosted and managed on remote servers by cloud providers eg AWS, Azure

On premise databases are installed and managed locally within an organisation's infrastructure.

22. Data governance is the framework of policies, standards and procedures that ensures data is accurate.

Importance
- Ensures data consistency and reliability
- Improves data security and privacy compliance
- Supports better decision-making through trustworthy data.

23 Data integrity means maintaining the accuracy, consistency and reliability of data throughout its lifecycle.

ways to maintain it:

- Use data validation rules
- Control user access and permissions
- Regularly back up data

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

importance:

- High quality data leads to reliable analytics and insights
- Poor quality data can cause wrong decisions
- Ensures trust in data-driven business strategies

25. A data analyst collects, organises and interprets data to support business decisions

26- A DBA manages and maintains database systems to ensure performance, security and availability.

Main responsibilities:

- Installing and configuring database software
- Managing user access and security
- Performing backups and recovery
- Monitoring database health

27. Main steps involved in designing a data pipeline

1. Data source identification
2. Data ingestion
3. Data transformation
4. Data storage
5. Data orchestration
6. Monitoring and Maintenance

25. Challenges in managing large-scale database

1. Performance and scalability issues
2. Data security and privacy risks
3. Backup and recovery complexity
4. Data inconsistency across distributed systems

| | Database platform | Use case |
|---|---|---|
| 29. | MySQL | Web applications, small to medium systems |
| | PostgreSQL | Advanced analytics, complex queries |
| | Oracle | Enterprise level business systems |
| | Snowflake | cloud data warehousing and analytics |

30. Main data storage formats used in analytics

| Format | Type | Description/use case |
|---|---|---|
| CSV | Text | simple, readable |
| JSON | Text | stores nested data |
| Parquet | Binary | optimized for big data processing |