BRIGHTLIGHT DATA ANALYTICS RESEARCH ASSIGNMENT 1

Section A: Database Fundamentals

1. Main types of databases
- Relational Databases (RDBMS)
- NoSQL Databases
- In-memory Databases
- Cloud databases
- Hierarchical and Network Databases

2. Relational Database Management System (RDBMS): A system that stores data in tables with rows and columns using relationships between tables.

3. Primary key vs Foreign key:
- Primary key: Uniquely identifies each record in a table
- Foreign key: A field in one table that links to the primary key in another table

4. Database Normalization: The process of organizing data to reduce redundancy and improve integrity. It ensures efficient storage and avoids anomalies

5. Database schema: The structure that defines tables, fields, relationships, views, indexes and other elements in a database

6. Differentiate
- Structured data : organized in rows/columns
- Semi-structured : Has tags or markers
- Unstructured data : No predefined format

7. Difference
- Fact Table : contains measurable, quantitative data
- Dimension Table : contains descriptive attributes

8. Data model : A blueprint for how data is stored and accessed. It's crucial for consistency, scalability and performance

9. Difference
- Database : stores current transactional data
- Data Warehouse : stores historical, structured data for analysis
- Data Lake : stores raw, unstructured / semi-structured data

10. Data Mart : A subset of a data warehouse on a specific business area. It's smaller and more targeted

SECTION B : SQL and Data Processing

11. Query Language : A query language retrives/ manipulates data
SQL is most used due to it's standardization, simplicity, and support accross platforms

12. Indexes : Structures that speed up data retrieval. They reduce the time needed to search for records.

13. Transactions : A transaction is a unit of work
ACID properties ensures:
- Atomicity : All or nothing
- Consistency : valid state maintained
- Isolation : Transactions don't interfere
- Durability : Changes persist after commit

14. Database Engine : The core service that processes queries and manages data storage. It affects speed, scalability and reliability

15. - View : Virtual table from a query
- Stored Procedure : Predefined SQL code block
- Trigger : Automatic action on data changes

16. Differentiate
- ETL (Extract, Transform, Load) - used in traditional systems
- ELT (Extract, Load, Transform) - used in modern cloud systems

17. Differentiate
- Batch : processes large data sets at interval
- Stream : processes data in real-time as it arrives

18. SQL joins : allows us to join two or more tables based on a common column
- INNER JOIN : matches records in both tables
- LEFT JOIN : All from left + matches from right
- RIGHT JOIN : All from right + matches from left
- FULL JOIN : All records from both tables

19. Referential Integrity : Ensures relationships between tables remain consistent. Prevents orphan records and maintains data accuracy

20. Data Redundary : Leads to wasted storage and slower performance. Normalization helps reduce redundancy.

SECTION C: Data Management and Analytics concepts

21. Cloud vs On-premise Database Management:
- Cloud: are hosted on remote servers and accessed via the internet, offering scalability, flexibility and reduced infrastructure costs
- On-premise: are hosted locally within an organization's infrastructure, offering more control and security but requiring higher maintenance

22. Data governance refers to the framework of policies, procedures and standards that ensure data is managed properly. It is important because it ensures data quality, security, compliance and accountability, enabling trustworthy and efficient data usage

23. Data integrity ensures the accuracy, consistency and reliability of data throughout it's life cycle. It can be maintained through:
- use of constraints eg primary keys, foreign keys
- Validation rules
- Access control
- Audit trails and data quality checks

24. Data quality refers to the condition of data based on factors like accuracy, completeness, consistency and timeliness. It is critical for analytics because poor-quality data leads to misleading insights, faulty decisions and

reduced trust in analytical outcomes.

25. Role of Data Analyst in managing and analyzing database information.
- Extracts and clean data from databases
- Perform data analysis to uncover trends and insights
- Creates visualizations and reports
- Collaborates with stakeholders to support data-driven decisions

26. A (DBA) Database Administrator is responsible for:
- installing and configuring database systems
- Monitoring performance
- Backing up and restoring data
- Ensuring security and access control
- Troubleshooting issues
- Maintaining data integrity and availability

27. Main steps involved in designing data pipeline
1. Data Source Identification
2. Data Extraction
3. Data Transformation
4. Data loading
5. Monitoring and Logging
6. Error Handling
7. Optimization and Scaling

28. Common challenges in managing large-scale databases

- Scalability
- Performance optimization
- Data security
- Backup and recovery
- Data consistency
- Cost management
- Compliance with regulations

29. Popular database platforms

- MySQL : open-source, widely used for web applications
- PostgreSQL : Advanced features, ideal for complex queries and analytics
- Oracle : Enterprise-grade, used in large corporations for mission-critical systems
- Snowflake : cloud-native, optimized for big data analytics and data warehousing

30. Main data storage formats used in analytics

- CSV : simple, human-readable, best for small data
- Parquet : columnar format, efficient for big data processing
- JSON : Semi-structured, ideal for APIs and nested data
- Avro : Compact binary format, good for Serialization and schema evolution