**COMP 263 – Mid Term**

**Total Questions:** 20  
**Format:** 18 Multiple Choice + 1 Descriptive + 1 Pipeline Drawing

1. What is the main difference between **SQL** and a **Relational Database System (RDBMS)?**  
   A) SQL stores data, RDBMS only queries it  
   B) SQL is a query language; RDBMS is the software that runs and manages the data   
   C) SQL enforces constraints, RDBMS does not  
   D) RDBMS defines tables using JSON
2. Which SQL keyword belongs to **Data Definition Language (DDL)?**  
   A) INSERT  
   B) UPDATE  
   C) CREATE  
   D) SELECT
3. Which SQL keyword is used to **retrieve** data from a database table?  
   A) EXTRACT  
   B) SELECT   
   C) RETURN  
   D) GET
4. What does the **ACID** property Consistency ensure?  
   A) All transactions execute simultaneously  
   B) The database remains valid before and after a transaction   
   C) Data is synchronized across multiple systems  
   D) Transactions can skip validation rules
5. According to the **CAP Theorem**, a distributed system cannot guarantee all three of the following at the same time:  
   A) Concurrency, Availability, Partition Tolerance  
   B) Consistency, Accuracy, Performance  
   C) Consistency, Availability, Partition Tolerance   
   D) Consistency, Atomicity, Durability
6. What makes **SQLite** unique compared to other RDBMS systems like Oracle or PostgreSQL?  
   A) It uses JSON format for data  
   B) It requires a dedicated server  
   C) It stores all data in a single local file   
   D) It doesn’t support SQL
7. According to the NoSQL article by Strozzi, one reason to not rely only on SQL is:  
   A) SQL systems are too fast  
   B) SQL requires complex licenses and lacks flexibility   
   C) SQL supports too many formats  
   D) SQL cannot be automated
8. What is the **main purpose of metadata** (author, timestamp, UUID) in big data synchronization?  
   A) It increases query speed  
   B) It tracks data ownership and freshness   
   C) It validates JSON structure  
   D) It replaces raw data
9. Which of the following is **not** a core dimension of data quality?  
   A) Completeness  
   B) Accuracy  
   C) Speed   
   D) Consistency
10. A dataset has mismatched units (kg vs lbs). This violates which dimension of data quality?  
    A) Timeliness  
    B) Validity   
    C) Accuracy  
    D) Completeness
11. If a sensor value shows 2000°C instead of 20°C, which data quality issue does it represent?  
    A) Completeness  
    B) Accuracy   
    C) Consistency  
    D) Integrity
12. In Lab 2, why were timestamps converted to UTC before uploading to MongoDB Atlas?  
    A) To make sorting faster  
    B) To ensure all records use a consistent time zone   
    C) To hide time zone data  
    D) To minimize storage size
13. How does filtering data before analysis improve quality?  
    A) Removes irrelevant or incorrect records   
    B) Adds redundancy  
    C) Increases dataset size  
    D) Randomizes data for better variance
14. What is the primary purpose of the **map()** function?  
    A) Summarize all data into one value  
    B) Transform each element of an array and return a new one   
    C) Filter invalid data  
    D) Modify a database schema
15. The **filter()** function is best used to:  
    A) Add new elements to a dataset  
    B) Select only elements meeting certain conditions   
    C) Rename columns  
    D) Remove duplicates automatically
16. The **reduce()** function:  
    A) Combines all elements into one summarized result   
    B) Filters invalid records  
    C) Maps one dataset to another  
    D) Converts JSON to SQL
17. What is the main goal of an **ETL (Extract, Transform, Load)** pipeline?  
    A) Load raw data directly into production  
    B) Extract, clean, and load data into a warehouse   
    C) Replicate data from one schema to another  
    D) Transform data after it is stored
18. In an **ELT** workflow, transformation occurs:  
    A) Before data is loaded into the target system  
    B) After data is loaded into the target system   
    C) During extraction  
    D) Only in real-time streaming
19. **Explain how metadata improves data quality during synchronization.**  
    In your answer, mention at least two metadata fields (e.g., author, last\_sync, timestamp) and describe how they help ensure consistency, traceability, and accuracy in big data workflows.
20. Draw a **simple data pipeline** that shows how data moves from where it is created to where it is used.

Your diagram should include **five main stages**:

1. **Data Sources** – where the data comes from (e.g., sensors, apps, files).
2. **Ingestion** – how the data enters the system.
3. **Processing / Transformation** – where the data is cleaned or modified.
4. **Storage** – where the data is saved (e.g., database or data lake).
5. **Visualization / Output** – where the data is shown or used (e.g., dashboard or report).