Programme Code: TU856/TU857/TU858
Module Code: CMPU 4003
CRNs: 22415, 30088, 31081

TECHNOLOGICAL UNIVERSITY DUBLIN

CITY CAMPUS - GRANGEGORMAN

TU856 – BSc (Hons) in Computer Science
TU857 – BSc (Hons) in Computer Science (Infrastructure)
TU858 – BSc (Hons) in Computer Science (International)

Year 4

SEMESTER 1 EXAMINATIONS 2024/25

CMPU 4003 Advanced Databases

Internal Examiner(s):
Dr. Deirdre Lawless
Dr. Paul Doyle

External Examiner(s):
Ms. Caroline McEnroy
Dr. Colm O'Riordan

Instructions To Candidates: Answer any **THREE (3)** Questions
All questions carry equal marks (33 marks)
1 additional mark will be applied to each student.

Exam Duration: 2 hours

1. A telecommunications customer care and billing company is undergoing a digital transformation and is considering different data storage and management solutions to handle its diverse data needs. The company deals with transactional data, historical data analysis, and unstructured data from various sources such as social media, telecom mast sensors, and log files. You are tasked with evaluating the suitability of data storage solutions for different use cases within the company.

(a) Define a *data lake,* a *data warehouse*, and a *relational database.* For each, highlight TWO (2) key characteristics and provide a typical use case.

(3 x 4 marks)

(b) Compare and contrast a *data lake*, *data warehouse* and a *relational database* as data storage solutions in terms of their *scalability*, *data structure flexibility*, and *query* capabilities.

(3 x 3 marks)

Question One continues on the next page

1. (c) (i) Discuss the suitability of *data lakes, data warehouses*, and traditional *relational databases* for the following scenario considering both the *type of data* and *objectives* in your answer:

The company needs to manage and process real-time transactions involving customer calls, SMS, and data usage. It requires solutions that ensure data integrity, support high transaction throughput, and provide fast query responses for customer service, billing, and fraud detection. Furthermore, the company must analyse historical customer usage patterns and average revenue per user (ARPU) over the past five years to inform strategic marketing and service improvement.

- Type of Data: Historical structured data, including call detail records (CDRs), billing data, and customer profiles.
- Objectives: Analyse data for strategic insights, ensuring the accuracy and reliability of data used in billing and fraud detection.

(3 x 2 marks)

(ii) Discuss the suitability of *data lakes, data warehouses*, and traditional *relational databases* for the following scenario considering both the *type of data* and *objectives* in your answer:

The company also needs to combine social media data, network sensor logs, and call centre transcripts to monitor customer sentiment, identify network performance issues, and gain insights for proactive service improvements.

- Type of Data: Unstructured and semi-structured data, including social media posts, network sensor data, log files from telecom equipment, and text data from call centre interactions.
- Objectives: Analyse social media posts and call centre transcripts to gauge customer sentiment in real time.

(3 x 2 marks)

[Question One Total Marks: 33 marks]

2. (a) (i) Explain the concept of *denormalization* and *briefly discuss its importance* when *migrating* from a relational database to a *NoSQL wide-column datastore* like Apache Cassandra.

(5 marks)

(ii) Identify and briefly discuss THREE (3) challenges that need to be addressed in the *Extract Transform and Load (ETL)* process when migrating from a relational database to a wide column data store.

(3 x 3 marks)

(b) (i) Define the *CAP theorem* and explain each of its THREE (3) components (*Consistency (C), Availability (A), and Partition Tolerance (P)*).

(5 marks)

(ii) For each of the following scenarios discuss the *trade-offs* of adopting a CA, CP or AP approach before *recommending a suitable approach*, justifying your choice:

- A banking application that requires strict consistency in account balances and transaction histories.
- A social media platform that prioritizes high availability and user engagement, where eventual consistency is acceptable.

(2 x 7 marks)

[Question Two Total Marks: 33 marks]

3. You are tasked with designing database replication for a large online educational platform that operates in multiple geographic regions and experiences high traffic. Data consistency and availability are critical for the platform, which provides courses, resources, and forums for students and educators around the world.

(a) **(i)** Compare *master-slave* replication, *master-master* replication, and *masterless* replication in the context of the educational platform.

In your answer you should:

- clearly explain each approach
- briefly discuss for each approach TWO (2) advantages of adopting it for the educational platform
- and briefly discuss for each approach ONE (1) challenge of adopting it for the educational platform.

(3 x 5 marks)

(ii) Recommend one of the approaches you discussed in part (a) (i) for use with the educational platform. Justify your answer.

(3 marks)

(b) Suppose Master-Slave replication was adopted for the educational platform.

Recommend THREE (3) mechanisms that can be employed to *minimize downtime* in case of a master server failure for the educational platform.

In your answer you should clearly explain each mechanism and justify your choice of these mechanisms.

(3 x 3 marks)

(c) Suppose that the educational platform decides to use CouchDB for storing user-generated content, such as forum posts and course comments.

Briefly discuss how CouchDB's *document-based model* and *eventual consistency* could be useful to facilitate this aspect of the educational platform.

(6 marks)

[Question Three Total Marks: 33 marks]

4. Partitioning is an important strategy in database management systems to optimize performance, enhance scalability, and facilitate data management.

(a) Explain the concept of *table partitioning* and explain TWO (2) main *advantages* of using table partitioning in database systems.

(6 marks)

(b) **(i)** Briefly explain the key differences between *range, list and hash* partitioning.

(3 x 2 marks)

(ii) For each of the following scenarios, recommend *a partitioning strategy* (*range, list, or hash*).

Justify your decision by providing *discussing* at least ONE (1) *pro and* ONE (1) *con* of your proposed approach for the scenario:

- A company maintains a database table with millions of financial transactions spanning the last 10 years. Queries are often based on a date range (e.g., extracting transactions for a particular month or quarter).
- An e-commerce company operates in multiple geographical regions, and their database table stores product sales data from different regions. Each region's data needs to be logically separated to allow regional managers to query only their data efficiently. The platform must ensure that queries related to one region do not impact the performance of queries from other regions and allow for regional-specific maintenance.
- A social media platform has a database table for user profile information, where user IDs are uniformly distributed and there is a need for balanced data distribution to prevent hotspots during read and write operations.

(3 x 3 marks)

Question Four continues on the next page

4. (c) Indexing plays a crucial role in optimizing data retrieval in databases.

For each of the following PostgreSQL index types *provide an example* of a query or dataset that benefits from each type of index.

Additionally discuss TWO (2) general *advantages* and ONE (1) general *disadvantage* of each type of index.

(i) BRIN (Block Range INdex)

(ii) GIN (Generalized Inverted Index)

(iii) BTREE (Balanced Tree Index)

(3 x 4 marks)

[Question Four Total Marks: 33 marks]