

Nirma University

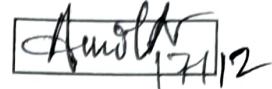
Institute of Technology

Semester End Examination (IR), December - 2025

M. Tech. in Computer Science and Engineering (Data Science), Semester-I
6CS302CC25 Data-Science System Design

Roll /
Exam No. 25MC0005

Supervisor's Initial
with Date


Amith
17/12

Time: 3 Hours

Max Marks :100

Instructions: 1. Attempt all questions.

2. Figure to right indicate full marks
3. Use section-wise separate answer book.
4. Draw neat sketches wherever necessary.
5. Assume necessary data wherever required, and indicate clearly.

SECTION-I

Q.1 Do as directed

[18]

CO1

- (A) Describe the phases of the Data Science project life cycle as proposed by (6)
BL1 the Cross-Industry Standard Process for Data Mining. Draw a neat, labeled diagram of the life cycle.
- (B) Why is 'reliability' critical system requirement of a data science system? (6)
BL2 Explain any three strategies to improve reliability in the data science system with help of suitable example.
- (C) Explain difference between Data lake and Data warehouse with a suitable (6)
BL1 case study.

Q.2 Do as directed

[16]

CO2

- (A) Draw and explain design of LSM tree based storage engine. Why it is more (6)
BL3 suitable for IoT sensor ingestion?
- (B) What is role of rate limiter in large scale system design? Explain any one (6)
BL2 rate limiter algorithm.
- (C) What is significance of service discovery in large scale system design? (4)
BL2 Explain using suitable example.

Q.3 Do as directed

[16]

CO4

- (A) Choose the most appropriate communication interface (REST, GraphQL, (6)
BL4 or gRPC) for a food delivery system and justify your choice with at least two technical reasons. A food delivery platform that provides the following functionalities:
- Fetch restaurant listings with menus.
 - Retrieve user order history.
 - Submit new orders and receive real-time status updates.
 - Provide real-time recommendations for restaurants and dishes.

- (B) BL6 Design the real-time Chatbot system for web and mobile users considering (10) following requirements:
- Users can send message and receive text responses
 - Throughput: Each Chatbot API server can handle 15k messages/sec, total system must support 50k messages/sec.
 - Availability : Each server need 1+2 replica to achieve 99.99% uptime
 - Storage : All messages must be stored for audit and future training
 - Use cache server/CDN to reduce latency and improve throughput
- Draw neat and labeled diagram of the system design. Justify selection of components (CDN/cache at server, Load Balancer, Message Queues, Database(SQL/NoSQL) for message and logs)

SECTION -II

- Q.4 CO3 Do as directed [18]**
- (A) BL2 Explain following partitioning techniques with respect to scalability: (6)
i) hash-based Sharding ii)range-based Sharding iii) consistent-hashing
- (B) BL1 Explain difference between Linearizable, strong consistency, and eventual (6) consistency with suitable example.
- (C) BL1 How does grid index support following query patterns efficiently? (6)
i) Range query on multiple attribute
ii) Partial Match
iii) Nearest Neighbour
- Q.5 CO3 Do as directed [16]**
- (A) BL2 Explain read process and write process on a storage system adopting (6) leaderless replication among 3 replicas.
- (B) BL4 Why is binary encoding (e.g., Avro, Protocol Buffers) preferred over JSON (6) for large-scale data processing systems? Explain in terms of storage efficiency, network usage, serialization/deserialization speed, and schema evolution.
- (C) BL2 What is significance of clock synchronization in a distributed system (4) design? Explain using a suitable example
- Q.6 CO4 Do as directed [16]**
- (A) BL6 Design a high-level data pipeline architecture of Food Delivery Time (10) Prediction system. Your answer should describe the key components: data ingestion, storage, preprocessing, feature engineering, model training, and prediction serving. A food delivery company wants to predict estimated delivery time (ETA) for customer orders. Data comes from: Order details (items, restaurant, order timestamp), Rider GPS data (location updates every few seconds), Restaurant preparation time logs, Traffic & weather data, and Historical delivery performance. The company wants the system to process GPS data continuously, compute real-time features (rider distance, restaurant load), and update ETA predictions during every stage of delivery. (Draw block diagram for your proposed real-time data pipeline)

- (B) BL3 There is a need to design a distributed ML inference service that must remain highly available. Each server has a MTTR of 2 hours, and MTBF of 40 hours. How many minimum additional replica servers required to achieve at least 99.9% system availability? (6)

