# MASTER OF SCIENCE
# DATA SCIENCE AND ANALYTICS
# (MSCDSA)

**MSCDSA/ASSIGN/SEMESTER-I**

## ASSIGNMENTS

**(January – 2026 & July – 2026)**

**MCS-061, MCS-062, MCS-063, MCS-207, MCSL-064,
MCSL-065**

**SCHOOL OF COMPUTER AND INFORMATION SCIENCES
INDIRA GANDHI NATIONAL OPEN UNIVERSITY
MAIDAN GARHI, NEW DELHI – 110 068**

# CONTENTS

**Important Notes**

1. Submit your assignments to the Coordinator of your Study Centre on or before the due date.
2. Assignment submission before due dates is compulsory to become eligible for appearing in corresponding Term End Examinations. For further details, please refer to MSCDSA Programme Guide.
3. To become eligible for appearing the Term End Practical Examination for the lab courses, it is essential to fulfill the minimum attendance requirements as well as submission of assignments (on or before the due date). For further details, please refer to the MSCDSA Programme Guide.
4. The viva voce is compulsory for the assignments. For any course, if a student submitted the assignment and not attended the viva-voce, then the assignment is treated as not successfully completed and would be marked as ZERO.

| | | |
|---|---|---|
| **Course Code** | **:** | **MCS-207** |
| **Course Title** | **:** | **Database Management Systems** |
| **Assignment Number** | **:** | **MSCDSA (I)/207/Assignment/2026** |
| **Maximum Marks** | **:** | **100** |
| **Weightage** | **:** | **25%** |
| **Last Dates for Submission** | **:** | **30th April, 2026 (for January session)** |
| | | **31st October, 2026 (for July session)** |

**There are four questions in this assignment, which carries 80 marks. Rest 20 marks are for viva voce. You may use illustrations and diagrams to enhance the explanations. Please go through the guidelines regarding assignments given in the Programme Guide for the format of the presentation. The answer to each part of the question should be confined to about 300 words. Make suitable assumption, if any.**

**Q1: Foundations of Data-Centric Database Systems (5 × 4 = 20 Marks)**

a. Data scientists often work with large, evolving datasets. Explain the **limitations of traditional file-based systems** in the context of analytics and machine learning workflows. How does a DBMS address these limitations?

b. Explain the following concepts with respect to the **relational data model**, giving one data-science–oriented example for each:

- Candidate Key
- Functional Dependency
- Referential Integrity
- Selection Operation
- Projection Operation

c. A **health analytics platform** maintains data about patients, diagnostic tests, doctors, and test results. Analysts want to query trends across diseases, age groups, and regions. Design an **ER diagram** for this system. Clearly identify entities, relationships, key attributes, and constraints. State assumptions made.

d. Convert the ER diagram designed in part (c) into **normalized relations up to 3NF**, clearly indicating primary and foreign keys.

e. Explain the importance of **indexes** in analytical databases. Differentiate between primary index, secondary index, and clustering index with suitable examples.

**Q2: Data Normalization, Dependencies, and SQL for Analytics (4+4+4 = 20 Marks)**

a. Consider the relation:
Dataset(DatasetID, DatasetName, Source, CollectionDate, Domain, OwnerName, OwnerEmail, UpdateFrequency)

- Identify the **primary key**
- List meaningful **functional dependencies**

b. Populate the relation with **8–10 sample records** and highlight potential **data redundancy and anomalies**.
c. Decompose the above relation into **2NF and 3NF**, explaining each step clearly.
d. Consider the following relations used in a **research publication analytics system**:

- Researcher(ResearcherID, Name, Email, Affiliation)
- Publication(PubID, Title, Year, Venue)
- Authorship(ResearcherID, PubID, AuthorOrder)

Perform the following using SQL:

- Create tables with appropriate primary and foreign keys
- Insert sample data (minimum 5 researchers, 6 publications)
- List publications by a given researcher
- Find researchers who have not authored any publication
- Find the publication with the highest number of authors
- List venues that have more than two publications

**Q3: Transactions, Concurrency, and Consistency in Data Systems (4+8+8 = 20 Marks)**

a. Explain the **ACID properties** of transactions with a **data science pipeline example**, such as feature store updates or model versioning.
b. Consider the following schedule involving two transactions T1 and T2 operating on a dataset table storing aggregate metrics:

| **Time** | **T1** | **T2** |
|---|---|---|
| t1 | READ(X) | |
| t2 | X = X + 50 | |
| t3 | | READ(X) |
| t4 | | X = X * 1.2 |
| t5 | | WRITE(X) |
| **Time** | **T1** | **T2** |
| t6 | WRITE(X) | |

i. Compute the **final value of X** (assume initial X = 100).
ii. Determine whether the schedule is **serializable**.
iii. Identify the concurrency issue involved and explain its impact on analytics accuracy.

c. Explain the **Two-Phase Locking (2PL) protocol**. Discuss how it ensures consistency in multi-user analytical environments. Can deadlocks still occur? Explain with a suitable example.

**Q4: Advanced Topics and Case-Based Understanding (5 × 4 = 20 Marks)**

Write short notes on the following. Support your answers with examples relevant to data science or analytics.

a. Centralized vs Distributed Databases in Large-Scale Analytics
b. Star Schema and Snowflake Schema in Data Warehousing
c. NoSQL Databases for Data Science (Explain one type with use cases)
d. Query Optimization Techniques for Analytical Queries
e. Log-Based Recovery and Checkpointing in Data-Intensive Systems