# Cover Page

# 23CD22T1 - DATA ENGINEERING

(CSE(DS))

|  |  |  |  |
| --- | --- | --- | --- |
| Course Category: | Professional Core | Credits: | 3 |
| Course Type: | Theory | Lecture-Tutorial-Practical: | 3-0-0 |
| Prerequisite: | Knowledge of relational database management systems (RDBMS) like MySQL, PostgreSQL, or Oracle and Familiarity with NoSQL databases. | Sessional Evaluation:  Univ. Exam Evaluation:  Total Marks: | 30  70  100 |
| Objectives: | **Upon successful completion of the course, the students will be able to:** | | |
| * Explain basic concepts of Data Engineering * Discuss bout Data Engineering Life Cycle * How to design Good Data Architecture. | | |

|  |  |  |
| --- | --- | --- |
| Course Outcomes | **Upon successful completion of the course, the students will be able to:** | |
| CO1 | Understand the definition and evolution of data engineering and differentiate between data engineering and data science (L2 & L4) |
| CO2 | Analyze the data engineering life cycle in comparison to the data life cycle. Identify and evaluate the major undercurrents across the data engineering life cycle, such as security. (L4) |
| CO3 | Apply the principles of good data architecture, differentiate between enterprise and data architecture, and explain major architecture concepts. (L3) |
| CO4 | Differentiate between various data storage systems and abstractions, including data warehouses, data lakes, and data lake houses. (L4) |
| CO5 | Design, optimize, and process queries on both static and streaming data, develop data models, and perform data transformations. (L5) |
| Course Content | UNIT-I  **Introduction to Data Engineering:** Definition, Data Engineering Life Cycle, Evolution of Data Engineer, Data Engineering Versus Data Science, Data Engineering Skills and Activities, Data Maturity, Data Maturity Model, Skills of a Data Engineer, Business Responsibilities, Technical Responsibilities, Data Engineers and Other Technical Roles.  UNIT-II  **Data Engineering Life Cycle:** Data Life Cycle Versus Data Engineering Life Cycle, Generation: Source System, Storage, Ingestion, Transformation, Serving Data.  **Major undercurrents across the Data Engineering Life Cycle:** Security, Data Management, DataOps, Data Architecture, Orchestration, Software Engineering.  UNIT-III  **Designing Good Data Architecture:** Enterprise Architecture, Data Architecture**,** Principles of Good Data Architecture, Major Architecture Concepts.  **Data Generation in Source Systems:** Sources of Data, Files and Unstructured Data, APIs, Application Databases (OLTP), OLAP, Change Data Capture, Logs, Database Logs, CRUD, Source System Practical Details.  UNIT-IV  **Storage:** Raw Ingredients of Data Storage, Data Storage Systems, Data Engineering Storage Abstractions, Data warehouse, Data Lake, Data Lakehouse.  **Ingestion:** Data Ingestion, Key Engineering considerations for the Ingestion Phase, Batch Ingestion Considerations, Message and Stream Ingestion Considerations, Ways to Ingest Data  UNIT-V  Queries, Modeling and Transformation: Queries, Life of a Query, Query Optimizer, Queries on Streaming Data, Data Modelling, Modeling Streaming Data, Transformations, Streaming Transformations and Processing.  Serving Data for Analytics, Machine Learning and Reverse ETL: General considerations for serving Data, Business Analytics, Operational Analytics, Embedded Analytics, and Ways to serve data for analytics and ML, Reverse ETL. | |
| Text Books &  References  Books | **TEXTBOOKS:**   1. Joe Reis, Matt Housley, Fundamentals of Data Engineering, O'Reilly Media, Inc.,June 2022,ISBN: 9781098108304   **REFERENCE BOOKS:**   1. Paul Crickard , Data Engineering with Python,Packt Publishing, October 2020. 2. Ralph Kimball, Margy Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Wiley, 3rd Edition, 2013 3. James Densmore, Data Pipelines Pocket Reference: Moving and Processing Data for Analytics, O'Reilly Media, 1st Edition, 2021 | |
| E-Resources | 1. <https://nptel.ac.in/courses/> | |