

Semester: 4				
Programme: Data Science				
Course: Data Engineering - II				
Paper code: S2DS250411P			Credits: 3	
Hours/week: 4				
Category: Core/MDC/SEC/VAC: Skill				
Theory / Practical / Composite: Practical				
No of Module: 1				
Course Outcome:				
1. Remember the fundamental definitions of data engineering, the various dimensions of data quality, and the core components of cloud service models and data architecture.				
2. Understand the stages of the data engineering lifecycle, the conceptual framework of "tidy data," and the structural differences between batch and real-time data integration.				
3. Apply Python’s pandas library for web scraping, data collection, and advanced munging tasks such as string formatting, datetime handling, and vectorization.				
4. Analyze the results of data profiling and group-by operations to identify patterns, relationships, and inconsistencies across multiple structured and semi-structured datasets.				
5. Evaluate the significance of data quality dimensions and the impact of missing values to determine the most effective strategies for data assembly, merging, and cleaning.				
6. Create integrated ETL pipelines and automated data workflows using cloud-native tools to move and transform data from source systems to analytical platforms for decision-making.				
SYLLABUS				
UNIT	CONTENT	HOURS or NUMBER OF CLASSES	CO Mappin g	COGNI TIVE LEVEL
1.	Introduction: Data Engineering defined. Data Generation in Source Systems. Data Engineering lifecycle. Data architecture.	4	CO1, CO2	K1, K2
2.	Data collection: Data Quality (dimensions of data quality; importance; consequences of poor quality). Data Profiling (benefits; types of profiling: structure, content, relationship). Series and DataFrame in pandas library of Python. Creating our own data. Importing data. Exporting data. Web scraping	8	CO1, CO3, CO4	K1, K3, K4
3.	Data Manipulation: Data assembly (concatenation; merging multiple datasets). Missing data handling (significance of NaN value;	15	CO2, CO5	K2, K5

	source of missing values; working with missing values). Tidy data (understanding the relationships among rows, columns, variables and values).			
4.	Data Munging: Data types (converting types; categorical data). Strings and text data (slicing; built-in string methods; string formatting; regular expressions). Apply (apply over a Series; apply over a DataFrame; vectorization). Group by operations (aggregate; transform; filter). The datetime data type (handling data that includes dates).	15	CO3, CO4	K3, K4
5.	Cloud Platforms: Relevance of the Cloud in data engineering; overview of major cloud service providers; cloud service models; cloud-based data storage and databases; data ingestion and processing using cloud-native tools; basics of serverless computing and workflow orchestration; security, identity management, and monitoring in cloud environments.	5	CO1, CO6	K1, K6
6.	Data Integration for Data Engineering: Importance of data integration; sources of structured, semi-structured, and unstructured data; ETL (Extract, Transform, Load); data integration tools and platforms; handling data quality, consistency, and schema mapping; real-time vs. batch integration; data pipelines for analytics and decision-making.	5	CO2, CO5, CO6	K2, K5, K6

Text Books

1. Pandas for Everyone: Python Data Analysis by Daniel Y. Chen. 1st edition. Pearson Publication.
2. Data Engineering with Python: Work with Massive Datasets to Design Data Models and Automate Data Pipelines Using Python by Paul Crickard. 1st edition. Packt Publishing.
3. Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems by Martin Kleppmann. 1st edition. O'Reilly Media.
4. Fundamentals of Data Engineering: Plan and Build Robust Data Systems by Joe Reis and Matt Housley. 1st edition. O'Reilly Media.

References – Online Resources

1. [Microsoft Learn – Get Started with Data Engineering on Azure: guided learning path introducing data engineering concepts, data architecture and Azure Data Lake Storage.](#)
2. [Microsoft Learn – Training for Data Engineers: structured self-paced learning paths covering data engineering roles, pipelines, storage and cloud data platforms.](#)
3. [pandas Documentation – Getting Started Tutorials: official tutorials for reading/writing data, DataFrames, combining datasets, time series and text manipulation.](#)

4.	<u>IBM – Data Engineering Professional Certificate (Coursera): multi-course online program with hands-on work in Python, ETL, data pipelines, databases, Spark and related data engineering tools.</u>
5.	<u>AWS Training – Data Analytics Learning Plan: AWS-authored self-paced training covering cloud data analytics services and concepts relevant to data engineering workflows.</u>
6.	Google Cloud Data Engineer Learning Path
7.	https://onlinecourses.nptel.ac.in/e-learning/preview/noc26_cs86
Evaluation	Continuous Assessment

Course outcomes (COs) and Cognitive Level Mapping

COs	CO Description	Cognitive levels
CO1	Remember the fundamental definitions of data engineering, the various dimensions of data quality, and the core components of cloud service models and data architecture.	K1
CO2	Understand the stages of the data engineering lifecycle, the conceptual framework of "tidy data," and the structural differences between batch and real-time data integration.	K2
CO3	Apply Python's pandas library for web scraping, data collection, and advanced munging tasks such as string formatting, datetime handling, and vectorization.	K3
CO4	Analyze the results of data profiling and group-by operations to identify patterns, relationships, and inconsistencies across multiple structured and semi-structured datasets.	K4
CO5	Evaluate the significance of data quality dimensions and the impact of missing values to determine the most effective strategies for data assembly, merging, and cleaning.	K5
CO6	Create integrated ETL pipelines and automated data workflows using cloud-native tools to move and transform data from source systems to analytical platforms for decision-making.	K6