# Data Engineering

| | | | |
|---|---|---|---|
| **Course Code:** | | **Course Credits:** | |
| **Teaching Hours / Week (L: T: P):** | | **CA Marks:** | |
| **Total Number of Teaching Hours:** | | **END-SEM Marks:** | |

**Course Pre-requisites: Basic understanding of programming concepts. Familiarity with SQL and relational databases. Knowledge of foundational data concepts, such as structured vs. unstructured data. Introductory knowledge of cloud computing and data storage concepts.**

**Course Description: This course provides an in-depth journey into data engineering, covering essential tools, platforms, and techniques to work with large datasets and design effective data pipelines. Students will explore distributed computing with Hadoop and Spark, database management with MongoDB, and cloud-based data services from Azure and AWS. Through hands-on labs and real-world examples, students will develop skills in data integration, pipeline optimization, and data security, preparing them to manage complex data workflows in enterprise environments.**

Course Objectives: (3 to 5)

The main objectives of the course are to:

1. Explain the differences between structured and unstructured data, understand the data lifecycle, and identify use cases for NoSQL databases.
2. Perform CRUD operations, work with JSON data formats, and demonstrate proficiency in using MongoDB for data management.
3. Evaluate the key components of Hadoop and Spark, comparing their functionalities and identifying optimal use cases.
4. Configure and manage data pipelines using AWS Glue and Azure Data Factory, optimizing data flow and storage across cloud environments.
5. Design data pipelines using Airflow, Kafka, and Informatica, applying best practices for data security, encryption, and compliance with regulations.

Course Outcomes: (Five) – Use Blooms Taxonomy

On completion of the course, learner will be able to–

1. Understand the fundamental concepts of data engineering, including data types, lifecycles, and NoSQL databases.
2. Apply MongoDB operations and SQL-like queries to retrieve, transform, and analyze data in a NoSQL environment.
3. Analyze big data processing frameworks, such as Hadoop and Spark, to select suitable tools for distributed computing tasks.
4. Evaluate cloud-based data services from AWS and Azure, implementing ETL processes and data storage solutions.
5. Create optimized and secure data pipelines using advanced data engineering tools and best practices in data security and compliance.

| | | |
|---|---|---|
| **Unit I** | **Foundations of Data Engineering and Data Lifecycle** | **(10 Hours)** |

| | | |
|---|---|---|
| Introduction to Data and Opportunities, What is data? (Structured, Semi-structured, Unstructured), The Data Lifecycle (Capture, Store, Process, Analyze, Visualize), Big Data and its characteristics (Volume, Variety, Velocity), Real-world use cases of Data Engineering, Overview of NoSQL Databases, Understanding NoSQL Databases (Key-Value, Document, Column-Family, Graph Databases) | | |
| **Pedagogy** | **ICT Teaching / Power Point Presentation and Videos** | |
| | **Self-study / Do it yourself:** Research and familiarize yourself with different types of data sources (files, relational databases, NoSQL, online services, etc.) that Power BI can connect to | |
| | **Experiential Learning Topics:** | |
| | **Case Study / PBL - Project Based Learning:** | |
| | | |
| **Unit II** | **MongoDB for Data Engineering** | **(10 Hours)** |
| Introduction to MongoDB, JSON data format and working with documents, Introduction to MongoDB Query Language, MongoDB Operations, CRUD operations (Create, Read, Update, Delete) in MongoDB, MongoDB Compass | | |
| **Pedagogy** | **ICT Teaching / Power Point Presentation and Videos** | |
| | **Self-study / Do it yourself:** Practice using the Power BI model framework by creating and managing tables, relationships, and dimensions. | |
| | **Experiential Learning Topics:** | |
| | **Case Study / PBL - Project Based Learning:** | |
| **Unit III** | **Big Data Technologies and Distributed Computing** | **(10 Hours)** |
| Big Data Technologies, Introduction to Big Data Processing, The need for distributed computing frameworks, Apache Hadoop Ecosystem (HDFS, YARN, MapReduce) - High-Level Overview, Apache Spark Basics, Introduction to Apache Spark for large-scale data processing, Spark Basics (RDDs, DataFrames, Transformations, and Actions), Introduction to Cloud Platforms, Benefits of using Cloud Platforms for Data Engineering Overview of Microsoft Azure and Amazon Web Services (AWS) | | |
| **Pedagogy** | **ICT Teaching / Power Point Presentation and Videos** | |
| | **Self-study / Do it yourself:** | |
| | **Experiential Learning Topics:** Experiment with DAX variables in complex calculations to simplify formulas and improve performance. | |
| | **Case Study / PBL - Project Based Learning:** | |
| **Unit IV** | **Data Services on Azure and AWS** | **(8 Hours)** |

| | | |
|---|---|---|
| Azure Data Services, Azure Data Factory (ADF) for ETL/ELT orchestration, Creating and scheduling data pipelines with ADF, Azure Synapse Analytics for data warehousing and big data analytics, Azure Blob Storage for scalable data storage, Azure Databricks for distributed data processing with Apache Spark, AWS Data Services, Introduction to AWS Data Services, Amazon S3 for object storage, Amazon Redshift for data warehousing, AWS Glue for ETL/ELT jobs | | |
| **Pedagogy** | **ICT Teaching / Power Point Presentation and Videos** | |
| | **Self-study / Do it yourself:** | |
| | **Experiential Learning Topics:** Monitor and evaluate the performance of reports and dashboards, analyzing user interactions and improving report load times. | |
| | **Case Study / PBL - Project Based Learning:** | |
| **Unit V** | **Advanced Tools, Security, and Optimization** | **(10 Hours)** |
| Advanced Data Engineering Tools, Apache Kafka: A distributed streaming platform for real-time data ingestion, Apache Airflow: A workflow orchestration tool for managing data pipelines, Snowflake: A cloud-based data warehouse solution, Informatica: A data integration platform for ETL/ELT processes, Hive: A data warehouse framework for distributed data management, Data Pipeline Optimization and Security, Best practices for designing and optimizing data pipelines, Managing data latency and reliability, Data Engineering Security and Compliance, Data privacy regulations (GDPR, HIPAA), Implementing data access controls and data encryption on cloud platforms | | |
| **Pedagogy** | **ICT Teaching / Power Point Presentation and Videos** | |
| | **Self-study / Do it yourself:** | |
| | **Experiential Learning Topics:** | |
| | **Case Study / PBL - Project Based Learning:** Design and implement a secure reporting solution in Power BI for a multinational organization, applying both row-level and object-level security, real-time reporting, and mobile optimization. Document the steps and provide a final report showcasing the security and interactivity features configured. | |

**Learning Resources**

**Text Books:**

1. Reis, J., & Housley, M. (2022). Fundamentals of data engineering. " O'Reilly Media, Inc.".
2. Kumar, K. P., Unal, A., Pillai, V. J., Murthy, H., & Niranjanamurthy, M. (Eds.). (2023). Data engineering and data science: concepts and applications.

**Reference Books:**

1. Atzmueller, M. (Ed.). (2016). Enterprise big data engineering, analytics, and management. IGI Global.
2. Crickard, P. (2020). Data Engineering with Python: Work with massive datasets to design data models and automate data pipelines using Python. Packt Publishing Ltd.

**MOOC/ Video Lectures available at: https://learn.microsoft.com/en-us/credentials/certifications/data-analyst-associate/?practice-assessment-type=certification**

ASSESSMENT AND EVALUATION PATTERN

| WEIGHTAGE | Continuous Assessment (CA) | End Semester Assessment (ESA) |
|---|---|---|

CA - Assignments (PBL / Case Study / Presentation / Seminar / Group Discussions / Quiz / Test

ESA - Bloom’s Taxonomy Levels: Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating

Experiential Learning - Internship (Summer / Winter) / Industry Visit / Site Visit / Field Trips

Project Based Learning – Mini (Minor) Project / Major Project

**Course Articulation Matrix (CO-PO Mapping) [**
**Prompt Engineering for Analytics and Visualization**
**– Course Code]**

| COs | PO-1 | PO-2 | PO-3 | PO-4 | PO-5 | PO-6 | PO-7 | PO-8 | PO-9 | PO-10 | PO-11 | PO-12 | PSO-1 | PSO-2 | Relevant to National/ Global/Regional |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CO-1 | 3 | 3 | | | 2 | | | | | | | | 2 | 2 | Global |
| CO-2 | | | 3 | | 3 | | | | | | | 2 | | 2 | Global |
| CO-3 | | | | 3 | 3 | 2 | | | | | | | 2 | | Global |
| CO-4 | | | 3 | 3 | | | | | | 2 | | | | 2 | Global |
| CO-5 | | | | | | 3 | | 2 | | | 2 | | 2 | | Global |

**3** – HIGH, **2** – MEDIUM, **1** – LOW