

Data Engineering and Analytics

Code: CS32213CS	Course Type: Core	Semester: II
Evolution Scheme (TA-MSE-ESE): 20-30-100	Total Marks: 150	L-T-P (Credits): 3-0-0 (3)

Pre-Requisites: Database Management Systems, Programming in Python/R, Data Structures & Algorithms, Probability & Statistics, Linear Algebra

Course Objective: To enable students to understand, design, and implement end-to-end data engineering pipelines and perform large-scale analytics using modern tools and distributed computing frameworks. The course integrates data lifecycle management, data storage, batch and stream processing, enabling data-driven decision making in AI and Data Science applications.

Course Outcomes: At the end of the course, the student will be able to

CO-1	Understand the fundamentals of data engineering, big data characteristics, data lifecycle, and the architecture of modern data platforms.
CO-2	Design and implement scalable ETL/ELT pipelines, data storage (SQL/NoSQL/Distributed), and data warehousing solutions.
CO-3	Apply distributed data processing frameworks for large-scale batch and real-time analytics.
CO-4	Demonstrate proficiency in data quality, governance, visualization, and integration of analytics with ML/AI pipelines for real-world applications.

Course Articulation Matrix:

PO	PEO-1	PEO-2	PEO-3	PEO-4	PEO-5
CO-1	3	2	1	-	-
CO-2	3	2	1	-	1
CO-3	3	2	-	-	3
CO-4	3	2	1	1	-

1 - Slightly;

2 - Moderately;

3 - Substantially

Syllabus:

Unit-1: Introduction to Data Engineering, Big Data & Data Lifecycle

Big Data Overview: Definition, Definition, 5 Vs (Volume, Velocity, Variety, Veracity, Value), challenges of conventional systems, data lifecycle. Data Engineering Fundamentals: Role of Data Engineer, Data Engineering vs Data Science vs Data Analysis, modern data stack architecture. ETL/ELT Paradigms: Lambda and Kappa architectures, CRISP-DM lifecycle. Cloud Platforms: AWS, GCP, Azure. Hadoop Ecosystem: HDFS, YARN, Hive, HBase; Apache Airflow for pipeline orchestration.

Unit-2: Data Storage, Management & Pipeline Design

Relational Databases: Schemas, ER diagrams, indexing, query optimization. NoSQL Databases: Key-value, document, column-family, graph models (MongoDB, Cassandra), CAP theorem, ACID vs BASE. Data Warehousing: OLTP vs. OLAP, star/snowflake schema, SCD Types 1/2/3, cloud warehouses (BigQuery, Redshift, Snowflake). Lakehouse Architecture: Delta Lake, Apache Iceberg, Parquet, ORC, Avro formats. Pipeline Design: Batch/streaming transformations, dbt, Fivetran, Airbyte.

Unit-3: Distributed Processing & Real-Time Stream Analytics

Hadoop MapReduce: MapReduce anatomy, job scheduling, shuffle and sort, PIG, HIVE. Apache Spark: RDDs, DataFrames, Spark SQL, Catalyst Optimizer, transformations and actions, performance tuning. Spark Structured Streaming: Micro-batch processing,

watermarks, event-time processing. Apache Kafka: Topics, partitions, producers, consumers, Kafka Streams, KSQL. Stream Analytics: Sampling, decaying windows, Apache Flink overview.

Unit-4: Data Quality, Analytics & Advanced Applications

Data Preprocessing & EDA: Cleaning, normalization, PCA, feature selection, Pandas, NumPy, Matplotlib. Data Mining: Association rules, classification, clustering (K-means), regression, time-series analytics. Analytics Engineering: dbt, MLflow, ML pipeline integration. Data Governance: GDPR, India PDPB, RBAC, data masking, lineage tracking. Applications: Recommendation systems, fraud detection, anomaly detection, ethical considerations.

Learning Resources:

Text Books:

1. Joe Reis and Matt Housley, **Fundamentals of Data Engineering**, O'Reilly Media, 2022.
2. Tom White, **Hadoop: The Definitive Guide**, 4th Edition, O'Reilly Media, 2015.
3. Jiawei Han, Micheline Kamber and Jian Pei, **Data Mining: Concepts and Techniques**, 3rd Edition, Elsevier, 2011.
4. Raj Kamal & Preeti Saxena, **Big Data Analytics: Introduction to Hadoop, Spark, and Machine Learning**, McGraw Hill.
5. Seema Acharya & Subhashini Chellappan, **Big Data and Analytics**, Wiley Publications.

Reference Books:

1. Holden Karau and Rachel Warren, **High Performance Spark**, O'Reilly Media, 2017.
2. Max Beauchemin, **Data Pipelines Pocket Reference**, O'Reilly Media, 2021.
3. Bill Inmon, **Building the Data Warehouse**, 4th Edition, Wiley Publications, 2005.
4. Narges Norouzi, **Modern Data Engineering with Apache Spark**, Apress, 2022.
5. Ananth Packkildurai, **Data Engineering with dbt**, Packt Publishing, 2023.
6. Silberschatz, Korth & Sudarshan, **Database System Concepts**, McGraw Hill.
7. Jure Leskovec, Anand Rajaraman & Jeffrey Ullman, **Mining of Massive Datasets**, Cambridge University Press.