RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based On AICTE Flexible Curricula

CSE- Data Science/Data Science, VII-Semester

CD701 Data Engineering

UNIT – I

Data Driven Organizations & Elements of Data:

Data-driven decisions, data pipeline infrastructure for data-driven decisions, role of the data engineer in data-driven organizations, Modern data strategies, Introduction to elements of Data, the five Vs of data – volume, velocity, variety, veracity, and value, Variety – data types & data sources, Activities to improve veracity and value.

UNIT – II

Design Principles and Patterns for Data Pipelines

The evolution of data architectures, Modern data architecture on various cloud platforms, Modern data architecture pipeline - Ingestion and storage, Modern data architecture pipeline - Processing and Consumption, Streaming analytics pipeline

Securing and Scaling the Data Pipeline:

Cloud security, Security of analytics workloads, ML security, Scaling Data Pipeline, creating a scalable infrastructure, Creating scalable components.

UNIT – III

Ingesting and Preparing Data:

ETL and ELT comparison, Data wrangling, Data Discovery, Data structuring, Data Cleaning, Data enriching, Data validating, Data publishing

Ingesting by Batch or by Stream

Comparing batch and stream ingestion, Batch ingestion processing, Purpose-built data ingestion tools, Scaling considerations for batch processing, stream processing, Scaling considerations for stream processing, Ingesting IoT data by stream

UNIT – IV

Storing and Organizing Data

Storage in the modern data architecture, Data Lake storage, Data warehouse storage, Purpose-built databases, Storage in support of the pipeline, Securing storage.

Processing Big Data

Big data processing concepts, Apache Hadoop, Apache Spark, Amazon EMR

UNIT – V

Processing Data for ML & Automating the Pipeline:

ML Concepts, ML Lifecycle, Framing the ML problem to meet the business goal, Collecting data, Applying labels to training data with known targets, Pre-processing data, Feature engineering, Developing a model, Deploying a model, ML infrastructure on AWS, AWS SageMaker, Automating the Pipeline, Automating infrastructure deployment, CI/CD, Automating with Step Functions.

List of Experiments:

- 7 - 10 experiments to be framed as per the syllabus.

Recommended Books:

1. Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, by Martin Kleppmann
2. T-SQL Querying (Developer Reference) by Itzik Ben-Gan
3. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling by Margy Ross
4. Spark: The Definitive Guide: Big Data Processing Made Simple by Bill Chambers
5. Data Pipelines with Apache Airflow by Bas P. Harenslak
6. Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing by Tyler Akidau
7. Kubernetes in Action by Marko Luksa