**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**WORK INTEGRATED LEARNING PROGRAMMES**

**Part A: Course Design**

| **Course Title** | Data Management for Machine Learning |
| --- | --- |
| **Course No(s)** | DSE\* ZG529 / AIML\* ZG529 |
| **Credit Units** | 4 |
| **Content Authors** | Pravin Y Pawar |
| **Version** | 1.0 |

**Course Description**

| Data Models and Query Languages: Relational, Object-Relational, NoSQL data models; Declarative (SQL) and Imperative (MapReduce) Querying; Data Encoding: Evolution, Formats, Models of dataflow; Machine learning workflow; Data management challenges in ML workflow; Data Pipelines and patterns; Data Pipeline Stages: Data extraction, ingestion, cleaning, wrangling, versioning, transformation, exploration, feature management; Modern Data Infrastructure: Diverse data sources, Cloud data warehouses and lakes, Data Ingestion tools, Data transformation and modelling tools, Workflow orchestration platforms; ML model metadata and Registry, ML Observability, Data privacy and anonymity. |
| --- |

**Course Objectives**

| **The course aims at providing:** | |
| --- | --- |
| **CO1** | Introduction to the data models, storages and querying languages used in data management emphasizing on machine learning aspects |
| **CO2** | Required guidance on architecture of modern data platform, usage and types of data pipelines |
| **CO3** | Hands-on exposure to the common techniques, and tools used by data engineers to support build, test, deploy and automate the machine learning pipelines |
| **CO4** | Exposure to the industry best practices essential to deal with data privacy, metadata and observability |

**Text Book(s)**

| **T1** | Fundamentals of Data Engineering: Plan and Build Robust Data Systems by Reis and Housley |
| --- | --- |
| **T2** | Reliable Machine Learning By Cathy Chen, Niall Richard Murphy, Kranti Parisa, D. Sculley, Todd Underwood |

**Reference Book(s) & other resources**

| **R1** | Designing Data-Intensive Applications by Martin Kleppmann |
| --- | --- |
| **R2** | Data Pipelines Pocket Reference by Densore |
| **R3** | Building Machine Learning Pipelines by Hapke, Nelson |

**Learning Outcomes:**

| **Students will be able to :** | |
| --- | --- |
| **LO1** | Understand the necessity, position and role of data management components appearing in the modern data stacks |
| **LO2** | Acknowledge the patterns, challenges and possible solutions associated with the data ingestion, flow, storage and processing on data platforms |
| **LO3** | Gain experience in designing and handling the dataflow during machine learning pipeline by means of state-of-art tools |
| **LO4** | Apply the acquired conceptual data management knowledge and practices over a real-world machine learning workflow addressing the model metadata, privacy and monitoring aspects |

**Part B: Course Handout**

| **Academic Term** | II Semester 2022-2023 |
| --- | --- |
| **Course Title** | Data Management for Machine Learning |
| **Course No** | DSE\* ZG529 / AIML\* ZG529 |
| **Lead Instructor** | Pravin Y Pawar |

**Glossary of Terms**

| **Module** | **M** | Module is a standalone quantum of designed content. A typical course is delivered using a string of modules. M2 means module 2. |
| --- | --- | --- |
| **Contact Hour** | **CH** | Contact Hour (CH) stands for an hour long live session with students conducted either in a physical classroom or enabled through technology. In this model of instruction, instructor led sessions will be for 32 CH. |
| **Recorded Lecture** | **RL** | RL stands for Recorded Lecture or Recorded Lesson. It is presented to the student through an online portal. A given RL unfolds as a sequences of video segments interleaved with exercises. |
| **Lab Exercises** | **LE** | Lab exercises associated with various modules |
| **Self-Study** | **SS** | Specific content assigned for self-study |
| **Homework** | **HW** | Specific problems/design/lab exercises assigned as homework |

**Modular Structure**

**Module Summary**

| **No.** | **Content of the Module** |
| --- | --- |
| M1 | Foundations of data management |
| M2 | Modern Data Platform |
| M3 | Data Management in ML Workflow |
| M4 | Advanced Topic in Data Management |

**Detailed Structure**

**M1: Foundations of data management**

**Contact Session 1-2**

| Session | Type | Description/Plan | Reference |
| --- | --- | --- | --- |
| 1 | CH1 | * Data Management Principles * Data Management Components | T2 |
|  | CH2 |
| 2 | CH3 | * Data Models and Query Languages * Data Encoding | R1  T1 |
| CH4 |
| Post CS | LE | * Lab 1 |  |

**M2: Modern Data Platform**

**Contact Session 3-4**

| Session | Type | Description/Plan | Reference |
| --- | --- | --- | --- |
| 3 | CH5 | * Data Architectures * Modern Data Stack * Data Pipelines and patterns | T1 |
|  | CH6 |
| 4 | CH7 | * Data Storage * Data Science Infrastructure * Serving Data for Analytics and ML | T1 |
| CH8 |
| Post CS | LE | * Lab 2 |  |

**M3: Data Management in ML Workflow**

**Contact Session 5-12**

| Session | Type | Description/Plan | Reference |
| --- | --- | --- | --- |
| 5 | CH9 | ML Workflow/lifecycle   * Data Pipeline vs ML Pipeline * ML Pipeline Stages * Training / Serving pipeline * Data management challenges in ML workflow | T2  R3 |
|  | CH10 |
| 6 | CH11 | Data Collection / Ingestion   * Diverse data sources * Data generation in source systems * Batch Ingestion * Message and Stream Ingestion * Ingestion strategies | T1 |
| CH12 |
| 7 | CH13 | Data Validation   * Common problems with data * Data skew and drift * Bias and Fairness * Data leakage * Data validation approaches | R3 |
| CH14 |
| 8 | CH15 | Data Analysis   * Types of Analytics * Data Exploration and Visualizations * Data Cubes and OLAP * Data Cube Operations * Data Cubes and ML | Instructor-supplied material |
| CH16 |
| 9 | CH17  CH18 | Analytics Engineering   * Data Integration * Data Transformation * Data Partitioning * Data Versioning * Test data management challenges | Instructor-supplied material |
| 10 | CH19  CH20 | Distributed Data Processing   * Big Data Analytics * Technologies for big data processing * Distributed and Parallel data processing * In-memory data processing * Hadoop, Spark, Kafka as exemplar architecture | Instructor-supplied material |
| 11 | CH21 | Feature Preparations   * Feature life cycle * Data Annotation / labeling * Data augmentation and Data Synthesis * Common Feature Engineering Operations * ~~Interaction data~~ * Feature Importance * Feature Generalization * Feature Stores * ~~Dataset Preparations~~ | T2 |
| CH22 |
| 12 | CH23 | ML Experimentation & Metadata   * Model training & experimentation * Model Analysis & Validation * ML Metadata Store * Dataset, Feature, Label, Pipeline metadata * ML Experiment Tracking data * ML model metadata and Registry | Instructor-supplied material |
| CH24 |
| 13 | CH25 | Pipeline Orchestration   * Pipeline Stages and DAGs * Apache Beam * Apache Airflow * Metaflow | Instructor-supplied material |
| CH26 |
| Post CS | LE | * Lab 3, ~~4,~~ 5, 6 |  |

**M4: Advanced Topic in Data Management**

**Contact Session 14-16**

| Session | Type | Description/Plan | Reference |
| --- | --- | --- | --- |
| 14 | CH27 | Data Privacy and anonymity   * Data privacy issues * Differential privacy * Anonymization * Methods to preserve privacy * Federated learning * Encrypted ML | T2  Instructor-supplied material |
| CH28 |
|  |
| 15 | CH29 | Data Observability   * Data Observability * Data downtime * Five pillars * Tools selection | T2  Instructor-supplied material |
| CH30 |
| 16 | CH31 | ML Monitoring & Observability   * Causes of ML System failure * Data Distribution Shifts * Problems with ML Production Monitoring * ML-specific metric * Monitoring Toolbox * Monitoring vs Observability | T2  Instructor-supplied material |
| CH32 |
| Post CS | SS | * To be identified |  |

**Experiential Leaning Component**

| **Lab** | **Topic** |  |
| --- | --- | --- |
| 1 | Design and implement the simple data flows involving various data formats  Modes of data flows   1. Through Databases – use SQL / Custom Program to read/write into databases 2. Through REST/RPC – Synchronous mechanism for data exchange 3. Through Message Brokers / Queues – Asynchronous mechanism for data exchange | * Virtual Labs |
| 2 | Build a Modern Data Stack  Components   1. a fully managed ELT data pipeline 2. a cloud-based columnar warehouse or data lake as a destination 3. a data transformation tool 4. A business intelligence or data visualization platform. | * Virtual Labs |
| 3 | ~~Implement and orchestrate a data pipeline~~  ~~Stages / Steps~~   1. ~~Extracting data using REST / file download~~ 2. ~~Data pre-processing~~ 3. ~~Integrating cleaned data obtained from multiple sources~~ 4. ~~Common data transformations on the data~~ 5. ~~Storing data for serving~~ 6. ~~Orchestration of pipeline stages~~ | * Virtual Labs |
| 4 | ~~Construct an end-to-end Machine Learning Pipeline~~  ~~Stages~~   1. ~~Problem understanding (aka business understanding)~~ 2. ~~Data collection~~ 3. ~~Data annotation~~ 4. ~~Data wrangling~~ 5. ~~Model development, training and evaluation~~ 6. ~~Model Validation~~ 7. ~~Local Model deployment~~ 8. ~~Prediction~~ | * Virtual Labs |
| 5 | Manage Machine Learning Model Metadata using MLFlow / Neptune  Components   1. Projects 2. Experiments 3. Model metadata 4. Model tracking / logging 5. Model Registry | * Virtual Labs |
| 6 | Construct a Machine Learning Pipeline with Data Versioning Tool ~~and Feature store~~  Components   1. Data Pipeline 2. Data Versioning Tool 3. Feature Store 4. ML Pipeline 5. Prediction Service | * Virtual Labs |

**Evaluation Scheme**:

Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

| **No** | **Name** | **Type** | **Duration** | **Weight** | **Day, Date, Session, Time** |
| --- | --- | --- | --- | --- | --- |
| **EC-1** | Experiential learning Assignment-I | Take Home | 15 days | 15% | TBA |
| Experiential learning Assignment-II | Take Home | 15 days | 15% | TBA |
| **EC-2** | Mid-Semester Test | Closed Book | 2 hours | 30% | Per programme schedule |
| **EC-3** | Comprehensive Exam | Open Book | 3 hours | 40% | Per programme schedule |

Syllabus for Mid-Semester Test (Closed Book): Topics in Session Nos. 1 to 8

Syllabus for Comprehensive Exam (Open Book): All topics (Session Nos. 1 to 16)

**Important links and information:**

Elearn portal: https://elearn.bits-pilani.ac.in

Students are expected to visit the Elearn portal on a regular basis and stay up to date with the latest announcements and deadlines.

Contact sessions: Students should attend the online lectures as per the schedule provided on the Elearn portal.

Evaluation Guidelines:

1. EC1 consists of two assignments. Announcements will be made available on the portal, in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: Use of books and any printed / written reference material (filed or bound) is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam which will be made available on the Elearn portal. The Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self-study schedule as given in the course handout, attend the online lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout.