**Final Project: Big Data Analytics & MLOps using Hive/Spark**

**Objective**

In this project, your will leverage **big data tools** (Apache Hive and Apache Spark) to perform **data ingestion, transformation, and predictive analysis** on a dataset of your choice. The goal is to **process large-scale data efficiently, extract meaningful insights, and build predictive models** while incorporating MLOps best practices.

**Project Deliverables**

1. **Project Report (PDF/Markdown)**
2. **Code Notebook (Databricks, Jupyter, or Script)**
3. **Presentation Slides (10–15 minutes)**
4. **Project Repository (GitHub or Databricks Workspace export)**

**Project Scope**

You should follow a structured workflow to showcase your skills in **Big Data processing and Machine Learning**.

**Phase 1: Data Selection & Problem Definition**

* Identify a dataset (**open-source, industry-relevant, or self-collected**) that contains **structured or semi-structured data**.
* Define a **business problem or research question** that requires big data processing and predictive analytics.
* Following are some example problem domains:
  + Customer segmentation for a retail business
  + Predicting loan defaults using financial data
  + Analyzing social media trends for sentiment analysis
  + Fraud detection using transaction data
  + Predicting equipment failure in IoT sensor data

(The examples above are only to give you an idea of the level of complexity, you may choose a problem outside of the list mentioned above).

**Phase 2: Data Ingestion & Preprocessing**

* **Ingest the dataset** into a big data platform (use databricks dbfs for the purposes of this project).
* Use **Hive** or **Spark SQL** for **data exploration and cleaning**.
* Handle missing values, duplicates, and perform transformations using **Spark DataFrames** or **Hive queries**.
* Apply **categorical encoding** techniques like one-hot encoding (if needed).
* Optimize data storage using **Parquet/ORC formats**.

**Phase 3: Exploratory Data Analysis (EDA)**

* Generate summary statistics using **Spark/Hive queries**.
* Visualize trends, distributions, and correlations using **Matplotlib/Seaborn**.
* Identify key insights that can impact business decisions.

**Phase 4: Predictive Modeling using Spark MLlib**

* Choose a **predictive modeling approach**:
  + **Regression:** Forecasting sales, price prediction, demand forecasting.
  + **Classification:** Fraud detection, sentiment analysis, churn prediction.
  + **Clustering:** Customer segmentation, anomaly detection.
* Implement the model using **Spark MLlib**.
* Tune hyperparameters using **cross-validation**.
* Evaluate model performance using **RMSE, AUC, Precision-Recall, etc.**.

**Phase 5: MLOps Best Practices**

* Automate data processing using **Spark pipelines**.
* Use **MLflow** or **Databricks experiments** to track different model runs.

**Phase 6: Insights & Business Recommendations**

* Summarize key findings from the predictive model.
* Provide actionable business insights based on model outcomes.
* Discuss any challenges and possible improvements.

**Evaluation Criteria**

| **Criteria** | **Description** | **Weightage (%)** |
| --- | --- | --- |
| **Problem Definition** | Clear articulation of dataset choice and business problem | 10% |
| **Big Data Processing** | Use of Hive/Spark for data transformation and ingestion | 20% |
| **EDA & Insights** | Quality of exploratory analysis and visualization | 25% |
| **Predictive Modeling** | Model selection, training, and evaluation | 25% |
| **MLOps Practices** | Reproducibility, automation, and tracking experiments | 10% |
| **Presentation & Report** | Clarity, completeness, and storytelling | 10% |