# Final Project: Integrated Big Data, Data Science and Machine Learning Pipeline

## Objective:

Design and implement an end-to-end big data, data science and machine learning pipeline that leverages Python libraries and frameworks learned throughout the course. The project should incorporate data manipulation, visualization, machine learning modeling, and natural language processing.

## Project Requirements:

### 1. Data Ingestion and Storage:

* Use TinyDB or a simple file-based system to ingest and store data.
* Data should be sourced either from publicly available datasets or simulated using any data generation library.

### 2. Data Processing and Analysis with Pandas:

* Perform data cleaning, transformation, and feature engineering using Pandas.
* Ensure the dataset is ready for modeling and visualization tasks.

### 3. Data Visualization:

* Create static visualizations with Matplotlib and Seaborn.
* Develop interactive visualizations using Plotly to allow dynamic exploration of the dataset.

### 4. Statistical Analysis:

* Use Statsmodels to conduct statistical tests or build statistical models to understand relationships within the data.

### 5. Machine Learning with PySpark and MLlib:

* Utilize PySpark to handle large datasets and perform data manipulation, aggregation, and summarization at scale.
* Apply machine learning algorithms from Spark MLlib on the dataset to predict outcomes or uncover patterns.

### 6. Deep Learning with TensorFlow and Keras:

* Build and train a deep learning model using TensorFlow and Keras to tackle a prediction or classification problem relevant to the dataset.

### 7. Natural Language Processing with NLTK and Transformers:

* If the dataset includes text data, use NLTK for text preprocessing and feature extraction.
* Leverage the Transformers library to apply a pre-trained model or fine-tune it for a specific NLP task.

### 8. Model Deployment (Optional):

* Deploy the machine learning or deep learning model using a simple Flask or Streamlit app that can make predictions based on user input.

## Deliverables:

* A Jupyter Notebook that contains the code, commentary, and visualization outputs, detailing each step of the project.
* A final report in a Word document summarizing the project's methodology, key findings, challenges encountered, and potential future work.
* Presentation slides for a brief presentation of the project, outlining the problem statement, approach, results, and insights.

## Evaluation Criteria:

* Completeness: All aspects of the project requirements should be addressed.
* Code Quality: Code should be clean, well-documented, and easy to understand.
* Analysis Quality: Visualizations and statistical/machine learning models should provide meaningful insights into the dataset.
* Innovation and Creativity: Extra credit for creative use of technologies and novel approaches to solving data science problems.