Final Project:

Overview:

Perform a large-scale data analysis using PySpark and Hive. Tailor the analysis based on your interests and the characteristics of the chosen dataset.

Steps:

1. **Data Ingestion:**
    - Obtain a diverse and large dataset. (Use Public data sets)
    - Upload the dataset to HDFS.
2. **Data Exploration with Hive:**
    - Create a Hive table to read the dataset.
    - Explore the structure of the data using Hive queries.
    - Identify any missing or inconsistent data.
3. **Data Preprocessing and Cleaning:**
    - Use Hive to clean and preprocess the data.
    - Handle missing values, outliers, or any data quality issues.
4. **Data Analysis with PySpark:**
    - Utilize PySpark SQL and DataFrame API for analysis.
    - Calculate descriptive statistics, aggregations, or any meaningful insights.
5. **Machine Learning Exploration:**
    - Explore machine learning tasks using PySpark's MLlib.
    - Experiment with classification, regression, or clustering based on the data characteristics.
6. **Data Visualization and Reporting:**
    - Use data visualization libraries to create visualizations.
    - Generate reports summarizing key findings from the analysis.

Tools and Technologies:

- Apache Hadoop (HDFS)
- Apache Hive
- Apache Spark (PySpark)
- Data visualization libraries
- Development environment (Jupyter Notebooks or others)

Business use cases:

- Analyzing customer behaviour and preferences.
- Detecting fraudulent activities in financial transactions.
- Improving efficiency and reducing costs in the supply chain.
- Analyzing electronic health records for insights and decision-making.
- Enhancing the customer shopping experience through personalized recommendations.
- Predicting and optimizing energy consumption in a smart city.

Each student needs to present for 5 minutes about their project during the 13th(April 13th) or 14th (April 20th – 10:am – noon) week. Please book the time in advance.

Deliverables:

- Hive SQL scripts for data exploration and preprocessing.
- PySpark script for data analysis.
- Data visualization notebook or script.
- Final report PowerPoint presentation.
- Demo video showcasing the project.

Notes:

- Choose a dataset aligned with your interests or domain expertise.
- Document your code and analysis steps