# CSE587 Project: End-to-End Big Data Pipeline using Hadoop and Spark (Phase1)

Release Date: Friday, February 14, 2025
Delivery Date: Saturday, March 22, 2025 by 11:59PM

February 14, 2025

## 1 Introduction

This course project of CSE587 is divided into two parts. This document is all about phase1 of the project (referred to below as a **project** or the **phase1**). The project requires you to design and implement an end-to-end big data pipeline on a Hadoop/Spark cluster. The project encompasses:

- Formulation of data problems to be solved using machine learning algorithms
- Date ingestion
- Data cleaning
- Exploratory Data Analysis (EDA)

This project is a group project with **2-3** student members (from either section of the course). You must register your group by filling out the ”group registration” form by the end of next Friday (Feb 21, 2025). Only one entry is required per group. Failing this would mean that you have not started working on your project.

**Your group should discuss and choose one of the datasets from the following selection:**

1. **Yelp Open Dataset** *URL:* `https://www.yelp.com/dataset`
2. **eCommerce behavior data from multi category store**

   *URL:* `https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store`
3. **NYC Taxi Trip Data (Yellow Taxi)** *URL:* `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`

4. **1.3M Linkedin Jobs & Skills** *URL:* `https://www.kaggle.com/datasets/asaniczka/1-3m-linkedin-jobs-and-skills-2024`

5. **Amazon Books Reviews** *URL:* `https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews`

# 2 Problem Statement

In addition to setting up the data pipeline, you are required to develop your own project focus. Explain what you are proposing to solve / find / achieve through these items on your dataset. In this section, you must:

1. **Formulate three machine learning problem statements:** Describe three problems that can be addressed using machine learning techniques on your selected dataset. For example, you might propose tasks such as classification, regression, or clustering tasks that are relevant to the dataset.

2. **Outline five data analysis objectives:** Clearly articulate five analytical goals or insights that you wish to derive from the data.

Your written problem statements and analysis objectives should be detailed enough to guide the work in Phase II.

# 3 Tasks (Step-by-Step)

Below are the tasks you need to complete for Phase I:

1. **Local EDA using Pandas**
   Perform exploratory data analysis on your chosen dataset using Pandas in a Jupyter Notebook or Python script. Generate summary statistics and preliminary visualizations.

2. **Hadoop Cluster Setup**
   Use the provided `docker-compose.yml` file to launch a local Hadoop cluster. Following the `Readme.md`, verify that the NameNode, DataNode, ResourceManager, and NodeManagers are running properly.

3. **Data Ingestion Script**
   Write a script to import your dataset from the local filesystem into HDFS. Ensure that the data is successfully stored in HDFS by verifying with HDFS commands.

# 4 Grading Criteria (Phase I: 100 Points)

- Local EDA with Pandas: **20 Points**
- Hadoop Cluster Setup (docker-compose): **20 Points**
- Data Ingestion Script into HDFS: **20 Points**
- Problem Statements (3 ML problems): **20 Points**
- Data Analysis Objectives (5 goals): **20 Points**

## 5 Submission Requirements

You should submit a zip file contains:

- Source code for your local EDA (Jupyter Notebook or Python scripts).
- The script used to import data into HDFS.
- A written report (in PDF format) that includes your problem statements, data analysis objectives and all the result you have for previous tasks. You are required to submit your report in IEEE/ACM format. `https://www.ieee.org/conferences/publishing/templates.html`
- Any additional visualizations and supporting materials.

## 6 References

For guidance, you may refer to:

- `https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html`
- `https://spark.apache.org/docs/latest/`
- Relevant research papers and online tutorials on big data pipelines and machine learning.

**Note1:** Only Phase I is released at this time. After Phase I submission, a separate document for Phase II will be provided.
**Note2:** Register your project within 1 week of the release of this document. See details sent by email. Only one entry is required. Failure to do so will be penalized.
**Note3:** A workshop on the phase1 will be conducted next week. Information will be sent by email.
**Note4:** No late submissions are accepted, so start working on the project early on and submit as per guidelines by the due date/time.
**Note5:** Phase2 of the project will continue on the phase1, so you would not be able to change the group or the project (dataset & problem).