University at Buffalo
Department of Computer Science & Engineering
CSE 4/587 - Data Intensive Computing

Fall-2025

Project Phase I

Release Date: September 13, 2025
Submission Date: October 18, 2025, by 11:59 PM

1 Introduction

The CSE 4/587 course project is a two-phase assignment. This document outlines the requirements for Phase I, which involves designing and implementing an end-to-end big data pipeline on a Hadoop cluster. This phase encompasses the following key tasks:

1. **Problem Formulation:** Defining data problems suitable for big-data solutions using machine learning algorithms.
2. **Data Ingestion:** Acquiring and loading data from the source into the distributed data pipeline.
3. **Data Cleaning:** Preprocessing data to handle missing values, inconsistencies, and errors.
4. **Exploratory Data Analysis (EDA):** Analyzing the dataset to understand its key characteristics and identify patterns.

This project is a group assignment for teams of 3-4 students. Nonetheless, we recommend UG students make their own group and Graduate students make their own groups. The reason is that some of the requirements in Phase II will be different for CSE487 and CSE587. Making your groups for the project was already discussed by instructor.

So, the first thing you want to do is your project group registration via this link:
https://forms.office.com/r/YjMQvwBeTu?origin=lprLink
Please note that only one entry for a group is required. Group registration needs to be completed in the first week. Even if you are not fully decided on the dataset, you should still submit the form with prospective datasets that you are considering.

2 Dataset Selection

Start by exploring the following datasets, and as a group discuss and select one of the datasets to be used for your project.

1. eCommerce behavior data from multi category store.
   `https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store`

2. NYC Taxi Trip Data (Yellow Taxi).
   https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
3. 1.3M Linkedin Jobs & Skills.
   https://www.kaggle.com/datasets/asaniczka/1-3m-linkedin-jobs-and-skills-2024
4. Amazon Books Reviews.
   https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews
5. Buffalo Trees dataset.
   https://data.buffalony.gov/Quality-of-Life/Tree-Inventory/n4ni-uuec/data_preview
6. Yahoo Finance Dataset.
   https://www.kaggle.com/datasets/iveeaten3223times/massive-yahoo-finance-dataset
7. Car Sales Dataset.
   https://www.kaggle.com/datasets/msnbehdani/mock-dataset-of-second-hand-car-sales.

3 General Guidelines

1. You will document everything when you are writing / compiling the report for submission, so it is better that you start documenting things as they happen (e.g. discussion, or refinements).
2. We also recommend that you use some collaborative environment such that everything is visible to all group members, and they can access and edit it at any time. It could be ONEDrive, UBBOX, or even GitHub.
3. You should set up weekly group meeting time (in-person or via zoom) when all group members can attend. You should also maintain the division of work and contributions by each member. In case of conflict, production of these items (attendance/work division/contribution) will let us resolve issues.
4. This is your project, own it and show passion doing this project. One of group members can take a leading role and should maintain attendance, division and participation etc.
5. Start working early on the project. Those groups who delay working on a project tend to suffer at the end, as there are multiple assignments by that time that they need to complete.
6. If there are any kinds of conflicts, then let us know early on. Resolution of such issues at an earlier stage is important.

4 Problem Statement

In addition to setting up the big data pipeline (next section), you are required to develop your own project focus. Explain what you are proposing to solve / find / achieve through big data processing of your chosen dataset in the cluster. After group discussions, you need to:

1. **Formulate machine learning problem statements:** Think, discuss, formulate and describe *N* **problems** that can be addressed using machine learning techniques on your selected dataset, where *N=number of students in your group*. So each member should think and contribute towards one of the problems. **For example**, you might propose tasks such as *classification, regression, clustering, or other suitable ML tasks* that are relevant to your dataset.
2. **Outline data analysis objectives:** Clearly articulate *2N* analytical goals or insights that you wish to derive from the data. So probably each member can contribute 2 goals / values /insights that they want to pursue by carrying out the project on their dataset.

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

5 Tasks(step-by-step)

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

1. **Data cleaning:** Clean and provision the data for downstream explorations and analytics.
2. **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.
3. **Hadoop Cluster Setup** Use the provided docker-compose.yml file to launch a local Hadoop cluster. Refer Hadoop Installation
4. **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.

6 Grading Criterion

1. Data Cleaning + Local EDA with Pandas: [**20 Points**]
2. Hadoop Cluster Setup: [**20 Points**]
3. Data Ingestion Script into HDFS: [**20 Points**]
4. Problem Statements (N ML problems): [**20 Points**]
5. Data Analysis Objectives (2N goals): [**20 Points**]

Report Guidelines:

You should submit a `.zip` file with the following:

```
<ubitname1_ubitname2_ubitname3>_phase1/
├── scripts/
│   ├── eda.ipynb
│   └── data_ingestion.sh
└── <ubitname1_ubitname2_ubitname3>_report.pdf
```

1. Source code for your local EDA (Jupyter Notebook or Python scripts).
2. The script used to import data into HDFS.
3. 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`
4. Any additional visualizations and supporting materials should be documented in the report.
5. Include only the files specified; do not add any extra files. You should provide details in the report.

Notes

1. Document all of your work and include explanations, observations, and justification for your choices wherever applicable.
2. All submissions must include proper references for any external resources used. Please note that citing a source does not permit directly copying and submitting it as your own work. Simply modifying minor changes to existing code may still be considered plagiarism. Your submission must reflect your original understanding and contricution. Kindly review the Academic Integrity statement for further guidance.

3. Only Phase I is released at this time. After Phase I submission, a separate document for Phase II will be provided.
4. Register your project within 1 week of the release of this document. Only one entry is required. Failure to do so will be penalized.
5. No late submissions are accepted, so start working on the project early on and submit as per the guidelines.
6. Phase II of the project will continue on the Phase I, so you would not be able to change the group or the project (dataset & problems).

Hadoop Installation

Local Installation

To prepare your development environment for this project you must first install and setup Java and Hadoop. Make sure to start this process as early as possible, issues WILL come up. To setup Hadoop, follow the instructions for pseudo-distributed operation here: `https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html`
If you do not have a compatible Java SDK installed, you will need to install that now as well. Recommended Java installations are listed at the above link as well.
Test out your installation on the provided examples to make sure it is working properly before moving on with the rest of the assignment.

Docker Setup

Access docker container at `https://buffalo.box.com/s/4eofd3kvfeirrejxxqtlpu4pzucheb4g`.
Access the documentation at
`https://towardsdev.com/setting-up-an-hdfs-docker-cluster-a-step-by-step-guide-d4846ff54b5d`
to learn how to run a `docker-compose.yml` and get started with containerization.
Access the documentation at `https://docs.docker.com/compose/` to learn more about docker compose.

Academic Integrity

Academic integrity is a fundamental part of the learning process. As a student, it is your responsibility to complete all work honestly and in accordance with the expectations set by your instructor. The goal is to ensure that you genuinely engage with and learn the course content, in alignment with UB's academic integrity principles.
This is an group assignment. Submitting code, report content, or any other assets (such as models, logs, or graphs) that are not entirely your own constitutes a violation of the academic integrity policy.
Thank you for upholding your personal integrity and contributing to UB's tradition of academic excellence.
For more information, please visit: `https://www.buffalo.edu/academic-integrity.html`