Project Phase #3

Due: November 25th @ 11:59PM

Content Covered

Data Processing and Machine Learning using Apache Spark MLib

Assignment Overview

In this phase of the project, you will engage in practical, hands-on learning by applying steps 6 and 7 of the data science pipeline to address issues within an application domain of your choice. Building on the work from the previous phases, you will now leverage Apache Spark to process large datasets and implement machine learning models using Spark's MLib. This phase emphasizes the scalability and efficiency of data processing and modeling using big data tools.

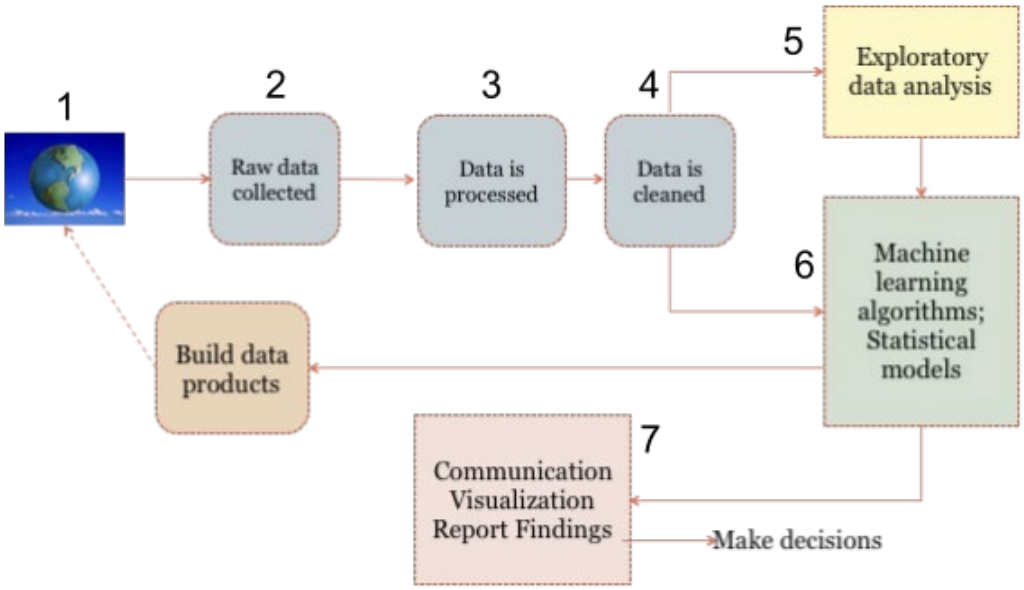

Figure 1: The Data Science Pipeline

Learning Outcomes for the Assignment:

1. **Distributed Data Preprocessing with PySpark**: Extending the data preprocessing steps from Project Phase 1 or by using new data preprocessing steps, perform distributed data preprocessing using PySpark. This includes cleaning, transforming, and preparing data for analysis. You are supposed to use RDD (Resilient Distributed Dataset) Operations, Windowing techniques etc to efficiently preprocess your dataset in a distributed fashion.

2. **Modeling with Spark MLib:** Implement machine learning algorithms (classification, regression, clustering, etc.) using Spark's MLib library. Compare these models that you used in Project Phase 2 (or you can use different model) , focusing on performance metrics like execution time, accuracy, precision, F1 Score. You can use libraries like matplotlib, seaborn to visualize your analysis.
3. **Performance Analysis:** Use Spark's DAG (Directed Acyclic Graph) visualizations to analyze the performance of your preprocessing steps and ML models. Evaluate the effectiveness of using Spark for large datasets.

Description:

Now that you have cleaned, processed, and performed exploratory analysis on your data and developed significant algorithms to extract intelligence from it, we will now shift our focus to handling large datasets using industry-standard techniques. Apache Spark is a powerful tool for effectively processing big data in a distributed environment. In this phase of the assignment, you will use PySpark to execute the data preprocessing steps and implement the machine learning algorithms you developed in earlier phases. The focus will be on distributed data processing using PySpark's powerful libraries, including Spark's MLib, to build and train machine learning models that can efficiently scale to large datasets.

General Assignment Requirements

1. **Work Environment:** Required language for the assignment is Python and the framework is PySpark. You can use Jupyter Notebook, Jupyter Lab or Google collab.
2. **Programming:** Prepare yourself to program by learning from the course textbooks and online resources.
3. **Academic Integrity: You will get an automatic F for the course if you violate the academic integrity policy. See the course syllabus for more detail.**
4. **Teams:** For this assignment, you may work in groups of one or two. Discussions regarding the assignment should only occur between you and your teammate, or you and course staff. Each team member must contribute to the assignment. There will be **one submission per team.**
5. **487 vs 587:** In certain instances, 587 students will be required to complete additional work, and in general their assignment will be held to higher standards. Instances of additional work will be clearly identified in the deliverables section.

Submission Requirements

1. **Deadlines:** Your submission is due by 11:59 PM on Nov.25th,2024. No late days will be allowed. Please start the assignment as soon as possible.
2. **Submission**: For the Phase 3 final submission, you are required to submit a zip file containing all the required deliverables. The zip file must be named: **member1_member2_member3_phase_3.zip.** It should contain a PDF for your project report named Phase3report.pdf and a src/ directory with your code files. in IEEE/ACM format, https://www.ieee.org/conferences/publishing/templates.html

Note: You also need to upload your code file.

Setup

- To prepare your development environment for this homework you must first install and set up PySpark. To install PySpark, follow the instructions here: https://spark.apache.org/docs/latest/api/python/getting_started/install.html

Deliverables **[100 marks total]**

1. **Distributed Data Cleaning/Processing [40 Marks]:** Leverage PySpark to perform distributed data cleaning and processing on your large dataset. Utilize big data techniques (E.g., using RDD operations, windowing techniques etc) to handle the scale and complexity of the data. You are required to document at least 4 distinct and unique distributed preprocessing and cleaning operations (6 for 587 students). Ensure your code is well-commented, with appropriate markup to explain the purpose and impact of each operation.

2. **Algorithms/Visualizations [50 marks]:** Develop 6 significant machine learning algorithms **(5 for 487 student and 6 for 587 students)** using PySpark's MLib. These algorithms should be relevant to your domain and specifically trained on large, distributed datasets. You can use the same models that you used on Phase-2. For 587 students, at least one of these algorithms must be selected from outside the class materials. Ensure that only PySpark's MLib is used for model training, taking advantage of its distributed processing capabilities.

3. **Explanation and Analysis [10 marks]:** Utilize Spark's DAG visualizations to explain the distributed stages of your data preprocessing and machine learning models. Provide a comparative analysis of the performance of your distributed models against those developed in Phase 2, focusing on metrics such as execution time, accuracy, precision and F1 Score. Discuss the effectiveness and advantages of using PySpark for distributed processing on large datasets, supported by relevant metrics and visualizations.

Additional Information and References

Some references and tutorials for a variety of algorithms and visualization techniques can be found here:

1. https://spark.apache.org/docs/latest/api/python/index.html
2. https://spark.apache.org/docs/latest/ml-guide.html

References

[1] C. O'Neill and R. Schutt. Doing Data Science., O'Reilly. 2013.
[2] J. VanderPlas, Python Data Science Handbook., O'Reilly. 2016.

Bonus Task [15 Marks] [Optional]

In Bonus task you will be building a data product from your Phase 2 models which would allow a user to interact with your models to gain insight into the data/problem statement you set out to solve. This could be as simple as allowing a user to input their own dataset for automatic analysis, or something more complicated tailored to your particular problem domain. For this Phase, it is important to put yourself in the shoes of your target users and imagine what type of product would be most useful to them.

Product Demo [15 marks]: submit a short video pitch/presentation (**no more than 5 minutes in length**), giving a brief demo of your product, how it is used, and what information people can learn from it.

a. [5 points] for showing a working user-interface (not just code you would expect a user to run)

b. [5 points] for showing how a user could input their own data (uploading their own datasets, filling out fields in your GUI, etc)

c. [5 points] for showing the feedback your product gives, explaining what it means, relevant manipulation/filtering of visualizations, and how a user could use it to help them solve a problem/answer a question.