Parallel and Distributed Computing
Assignment2
Streaming ML

For this assignment, you will pick your own dataset and perform a steaming ML analysis of it.

Dataset

You will be picking your own dataset, ideally something of interest to you. You will need to run a SparkML pipeline on the data to classify/cluster/etc. the data in some way. Thus, your dataset will need to be suitable for that type of analysis. You will also have to stream your dataset, but we can simulate streaming with most datasets.

You can find your dataset anywhere you wish, but the following are a few sites that have some datasets you can explore for ideas:

https://www.kaggle.com/
https://archive.ics.uci.edu/ml/datasets.php
https://www.data.gov/

And the following site list even more listed by category:
https://lionbridge.ai/datasets/the-50-best-free-datasets-for-machine-learning/

Machine Learning

You will have to pick something in your dataset to analyze. This can be a simple binary classification or clustering or anything of your choice.

Obviously, the raw data in the dataset will need to be transformed to work with whatever ML algorithm you pick. Thus, you should create a pipeline to handle the fitting and transformations.

You will need a training set to train your ML algorithm. You will want to split off a portion of your dataset to use in training. Then the rest of the dataset can be used for testing.

You do not need to tune your ML algorithm parameters or compare and contrast different algorithms for this assignment.

Streaming

Lastly, you are to incorporate Spark Streaming into your assignment. That is, view the training set of data as "historical" data and simply use it to build your fitted model. Then the testing set of data should be streamed to your program. You can always simulate streaming by breaking up a larger DataFrame into several smaller files and stream by reading 1 file per trigger. Note you will still be using a ML pipeline model to transform your streaming data.

Writeup

You will also provide a writeup of your project. This should include a high-level description of your data. This includes where you found it and what the data represents. You should also discuss the ML problem you are trying to solve. This includes what ML algorithms you used and how accurate their results were. Some sample results along with the high-level analysis of your work would be good to include as well. You should also discuss how you handled the streaming section of the project. And lastly, please discuss any issues you ran into from cleaning data to applying ML algorithms to streaming.

Submission

You are to attach 3 documents to your submission. The first is your Python source code (PY). Make sure to have your name in a comment at the top of your source code. Second is your dataset. Third is your writeup document as described above.