Big Data Processing - class projects

As part of this course you are required to complete three projects in groups of up to five students. Each project will require that you apply different technologies that part of the Hadoop/Spark ecosystem. The first project is build around processing data and applying the machine learning library of Spark, the second project around analyzing streams of data, and the third around modeling graph data. It is important that your group selects a dataset all members are passionate about since you will be work with the same dataset throughout the entire semester.

Details about each project are provided below. You should prepare a report and an oral presentation. Reports should not exceed 5 pages. All your code should be commented and turned in as well.

Project 1 (data cleaning + MLlib)
Due date: October 1st
Goal: Prepare a dataset and predict behavior based on regression or classification.
Assignment: For you first project you get familiar with a dataset of your selection. You want to defining a target variable whose behavior should be predicted. Prediction may refer to estimating a continuous variable (i.e., carrying out a regression) or a classification tasks.

Your report should include:
1. A description of the initial dataset
2. A summary of the transformations performed on your data and a description of the final cleaned dataset
3. A comparison of at least the machine learning techniques that were evaluated to predict a given target variable (make sure to state clearly what you are trying to predict and how you are measuring performance).
4. A description of how the analysis/prediction can be extended in the context of a streaming and a graph application (see projects 2 and 3 below)

Project 2 (Spark Streaming + Kafka/Flume + MLlib)
Due date: October 29th
Goal: Predict behavior (based on regression or classification) based on batches of steams of data.
Assignment: You are now trying to perform the same analysis as for project but not all data is available once. Thus you will be analyzing batches of streams of data as it becomes available. Use the machine learning technique that worked best for your first project.

Your report should include:
1. A description of the streaming architecture (which technologies of the ecosystem are you using?)
2. An analysis of the impact of information. How much information do you need to make predictions with the same accuracy as for project 1? How big is your batch size (in time)?

Project 3 (Spark GraphX + Neo4j + MLlib)
Due date: November 26th
Goal: Predict behavior (based on regression or classification) based on measures of a graph (topological properties)
Assignment: Evaluate at least five measures (topological properties) of the resulting graph. Your objective is to evaluate the utility viewing data as graph data, that is data which explicitly takes into account relationships between entities.

Your report should include:
1. A comparison of the different topological measures what where evaluated. Which measure is most useful for prediction?
2. A description of Neo4j collection
3. A comparison of the prediction results with and without topological measures
4. Conclusions: A summary of all key findings throughout all projects

Grading
Reports are worth 4 points and your oral presentations are worth 1 point for a total of 5 points for each project. Projects weights are distributed as follows: project 1 (30), project 2 (30), and project 3 (40).

Reference code
The following scripts may help to get you started:
- RDD.py
- DataFrames.py
- MLlib_classification.py
- MLlib_streaming.py
- MLlib_streaming_with_filtering.py
- MLlib_streaming_max_temp.py
- MLlib_graphframe.py
- Cleaning_dataframes.py