Project

Working in a group, you will apply the concepts you've learned about the data science pipeline in the context of a big data application.

Groups

It is recommended that you undertake the project in a group of three to five team members. Once you've formed a team, send me an email with the names and email addresses of your team members so that I can set up your teams.

Datasets

The first thing your team will want to do is identify an appropriate source of data for your projects. Keep in mind that you will want to demonstrate how to best explore this data source using parallel, distributed and cloud-based computing techniques, so take some time to find something that will be appropriate for this project. Here are some places you can start (although you are not limited to these as data sources!):

- NYC Taxi trip record data
- The GDELT project
- YouTube-8M Dataset
- Kaggle (searchable)

- Awesome Public Datasets (on GitHub)
- UCI Machine Learning Dataset

Application

Next, you'll want to figure out what you'd like to do with this data. There are many possibilities, but we recommend that you choose something:

- related to the data science pipeline (collection, storage, cleaning/wrangling, processing, data modelling, machine learning, visualization, etc)
- incorporating some kind of team practices relevant to the DevOps tools and techniques we will cover in class
- makes use of parallel and/or distributed computing techniques

Here are some suggested projects you can use as a starting point:

- Further explore the capabilities of libraries covered in class (Dask, Parquet, Spark (Streaming, SQL, ML), Bokeh)
- Apply the concepts covered in class to perform data analytics and visualization to gain insight from your chosen dataset.
- Setup and/or explore big data systems and libraries not covered in class (e.g. Hive, Kafka, Cassandra, GraphX, Tensorflow, Datashader etc.)
- Compare and contrast the performance (accuracy and efficiency) of machine learning models.
- Build an interactive data analytics and visualization dashboard.

Deliverables

The following are deliverables for this project. All will be assessed as team submissions, although individual grades may be adjusted throughout the project if it seems that progress and contributions to the work have not been equitable.

Pitch

On March 7, each team will have 5 minutes to pitch their project idea in the form of an oral presentation. The team should use no more than one slide per team member (not including a title slide and one for references).

This will allow everybody to receive rapid feedback about their idea from instructors as well as their classmates.

Proposal

Following this, your team will submit a brief project proposal (no more than two pages) via D2L, incorporating the feedback you have received.

- Keep it brief (max four pages)
- Convince me that your project is feasible for the approximately four weeks you will have to complete it
- What is your dataset(s)? Where can you find it, what do you think the strengths and weaknesses of this dataset are?
- Methodology:
    - Don't make it expensive
        - I do have Google Cloud and Cybera instances available
        - Be mindful of how you use your compute resources (If using cloud, make sure you know how this works!)
    - Make sure all the pieces are available
    - Make sure all the pieces are possible
    - If machine learning - what is the story you will tell? How will you determine if the machine learning was successful?

- Learning objectives. Each person in the group should list **one or two SMART objectives which will be something new they will learn or do** by completing the project. Ensure that we can all evaluate whether your team members have met the objectives by the time the project is complete.
    - A good learning objective should include something specific with an action verb which could be answered by a yes/no question, with a deadline. For example: "I will deploy my own Docker container onto a Google Cloud instance by March 20."
    - A poor learning objective may lack clarity or boundaries: "I will know how deep learning is used in pipeline simulations." Here, "know" is a verb which is not clear, and there is no deadline.
- No jargon. If we haven't covered this in DATA 604 or 608, unpack any terms (for example, HuggingFace, Cuda, Kafka, Databricks) with a brief explanation.
- Incorporate your feedback into the proposal.

Check-in

In between the proposal and presentation, your team will meet with me in-class during the time I use to have teams work on projects, to give me an update, run as a standup meeting. You should let me know as a team:

- What platform(s) you are using.
- What you are currently working on.
- Whether your proposal is still feasible, or what modifications you are making.

Individually, you will let me know:

- How you have worked towards your learning objective.
- What is one thing that is going well.

- What is one obstacle or potential obstacle.

All team members must be part of this meeting, as you will be given a grade for your check-in.

Presentation

Each team should have ready (by the last week of the course) a video approximately 12-15 minutes long to present their projects to their classmates. These videos will be uploaded to a D2L discussion board.

You will be expected to participate in providing feedback and constructive questions to 2-3 other teams using the discussion board - we recommend open-ended feedback that can elicit responses besides "yes," "no," or "thank you."

Skipping this step (or providing minimal feedback to other teams) may result in penalties which will be applied to your own project presentation grade.

In our synchronous presentations on April 4, and 6, half our teams (either volunteering or chosen at random on the day) will present each day, with the following format:

- A brief summary of your project for your classmates who may not have seen your video yet
- A discussion of your project using questions/feedback from the discussion board as the basis
- Slides are not required, though you may find them useful to structure the discussion of your project
- You SHOULD NOT recapitulate your entire presentation.

Project Report

As your major deliverable for the project, you'll be asked to produce a report for your project. The format can be variable depending on the nature of your project, but this report should showcase all the

work performed by your team this semester, including relevant research, the approach and methodology of the project, your results, and any discussion arising from your results.

I will refer questions about grading to the University of Calgary's grading system for graduate courses: https://www.ucalgary.ca/pubs/calendar/current/f-1-2.html.

Reflection (optional, not graded)

An important component of learning in teams and working on group projects is taking an opportunity to think about what you have done and you have learned. As you wrap up your final courses in the Diploma, take some time to think about your project for this course, and how you will continue to build upon it in the future.

This milestone is also an opportunity for you to provide feedback to me for your teammates, particularly if the distribution of work was somehow inequitable. Where I feel that an adjustment is well-justified based on what I know about your team and the reflections, your project grade may change.