EECS4415 Big Data Systems

Fall 2018

Team Project on Big Data Analytics

Objective

An important learning component of the class consists of a substantial course project. The project involves defining a problem, understanding the storage and processing needs (short term and long term) of a big data analytics solution and select an appropriate technical approach. You are expected to define the problem, find the appropriate data, and create an initial proof-of-concept implementation of the solution that demonstrates your understanding of stitching together a viable end-to-end big data analytics components. This will equip you with necessary skills and knowledge for applying a state-of-the-art data science approach in real-world problems, as they arise in industry or research settings.

Guidelines

- This is a team project, where each team consists of up to four (4) members.
- You need to identify an analytics problem based on an existing or new data set. There are no constraints on the data as long as all privacy or confidentiality constraints are met.
- You need to implement a process that computes the result in a repeatable fashion. Hence, it cannot just be a one-time computation. It should be sufficiently easy to kick-off a new end to end execution of the process. It should also be easy to analyze larger amounts of the data as they become available over time.
- The result of the processing should be accessible for review through some kind of serving layer and presented in a form that would make sense in an intended real world scenario.
- You can pick any of the technical solutions discussed in the course as long as you can justify why you picked that solution. The justification must be grounded in a real world use case.

Datasets & Additional Software Tools and Libraries

Check the course website, under resources. Many Open Data initiatives and tools are available online.

Defining a Problem

- You should formulate the specific problem and use case for the system/application.
- It does not need to be a big data problem, but it should involve complexity along some dimensions such as the size of the data (*Volume*), the quality and variety of the data (*Variety*), the speed at which data arrives and need to be analyzed (*Velocity*).
- It is permitted, but not required, to select a problem that requires advanced processing such as data mining or machine learning algorithms.

- You can make assumptions about the data, number of processes, users, etc. One such assumption could be, for example, that the data is cleaned to a certain degree. All such assumptions must be explicitly defined, and when appropriate reflected in the solution and the report. In other words, the focus of the problem is not on accuracy or importance of the findings, but rather on the big data analytics solution provided that can support the analysis.

Suggestions

Start by finding a data set. Based on the data set identify an interesting insight from the data using exploratory analysis. Implement a processing pipeline that can process and derive the insight repetitively. Determine what needs to be computed, how frequently should be computed and how frequently you expect the data to be updated. Based on this determine what kind of architecture you need. Make sure that the architecture you choose can scale. If there are limitations on scale, document them, so it becomes clear that the limitation is acceptable. Aim to implement a steel thread of the solution quickly (a steel thread is a subset of the functionality implemented end-to-end).

What and how to Submit?

When you have completed a part of the project deliverables, move or copy it in a directory (e.g., project), and use the following command to electronically submit your files within that directory:

```
% submit 4415 project proposal.pdf report.pdf presentation.pdf code.zip team.txt
```

The `team.txt` file includes information about the team members (*first name, last name, student ID, login, yorku email*). You should submit the appropriate files individually after you complete each part of the project– simply execute the *submit* command and give the filename that you wish to submit. Make sure you name your files **exactly** as stated (including lower/upper case letters). Failure to do so will result in a mark of 0 being assigned. Also, make sure that the **same student** should submit all parts of the project. You may check the status of your submission using the command:

```
% submit -l 4415 project
```

Project Deliverables, Due Dates & Evaluation

The table below presents the breakdown of the project deliverables, weights and due dates.

Project Deliverables	Weight	Due Date
Project proposal (~2 pages)	20%	Fri, Nov 16
Project report (~5 pages) and source code	50%	Fri, Nov 30
Project In-class presentation	30%	Tue, Dec 4

In the next sections, details about each of the project deliverables are provided.

Deliverable 1: Project Proposal (20%)

The project proposal should build on one or more large-scale open dataset(s) that are good candidates for analysis. The idea is to survey the data sets freely available and identify what are their strengths and weaknesses for the needs of the analysis. The proposal should then focus on what are some promising intuitions or questions that can be answered through rigorous analysis of the data. You should try to provide a concrete proposal for a big data analytics solution that can potentially be used for analysis of the same dataset over time. Emphasis should be given on how the solution scales. The header of the proposal should include the course title and an indication that this is a project proposal, the title of the project, your name(s) and contact information. The project proposal should have the following parts:

- **Domain Description and Motivation**: What is the data domain? What is the goal of your project? What is the motivation for rigorous data analytics? What are the questions you want to answer? Why the analysis is important? What are a few potential applications?
- **Architecture of Proposed Solution**: Provide a draft of your overall data analytics architecture. Describe the data collection/ingestion process, data storage, data processing, data serving, and (optionally) data visualization. Provide a draft figure that depicts the overall architecture and data flow in your system. Describe anticipated limitations or difficulties with your approach.
- **System Evaluation and Data Analysis**: How will you evaluate your system and architecture? What results you plan to obtain? What type of data analysis you will perform? How this type of analysis is adequate for the data, problem and questions posed? What other datasets can be used? What are the steps you need to take to scale your solution?

Formatting and Style

The suggested length of the project proposal is **2 pages** and it must be in PDF format. All reports should be formatted according to the new Standard ACM Conference Proceedings Template (pick the *sigconf* format). There are templates for both latex and MS Word users. More information can be found here: http://www.acm.org/publications/proceedings-template

Project Proposal Evaluation

The *project proposal* will be evaluated based on the following mark breakdown.

In-class Presenation Component	Weight	Due Date
Domain Description and Motivation	30%	Clear description of the domain and motivation for the big data analytics solution.
Architecture of Proposed Solution	30%	Clear description of the system architecture and data flow in the system.
System Evaluation and Data Analysis	30%	Clear set of tests and analytics to be performed.
Style and Language	10%	Overall organization, language, and style.

How to Submit?

```
% submit 4415 project proposal.pdf team.txt
```

Deliverable 2: Project Report and Source Code (50%)

The project report should represent all the completed work. The expectation is that most of the work has been completed and any major results are available. You should be able to provide a complete description of the project, even if a few minor parts have not yet been implemented or solved. The header of the report should include the course title and an indication that this is the project report, the title of the project, your name and contact information. The report should be structured as follows:

- **Abstract**: The abstract (limited to **150-200** words) should be a comprehensive but concise description of your project that aims to attract potential readers. It should briefly discuss the motivation, problem of interest, technical approach to solve it and main results of your work.
- **Introduction/Motivation**: What is the project about? What is the data domain? What is the goal of your project? What is the motivation for rigorous data analytics in this domain? What are the questions you want to answer? Why the analysis is important? What are potential applications?
- **Data and Data Analysis**: What are the data dimensions and processing dimensions of your solution? What are the pre-processing steps you need to perform? Are there any data cleaning, data transformation steps that are required? What are the data analytics methods you have employed? What type of data analysis you have performed? How this type of analysis is adequate for the data, problem and questions posed?
- **Architecture of Proposed Solution**: Describe your overall data analytics architecture. Describe the data collection/ingestion process, data storage process, data processing, data serving, and (optionally) data visualization. Provide a figure that clearly depicts the overall architecture and data flow in your system. Describe limitations or difficulties with your approach. Try to be as specific as possible.
- **Evaluation/Results**: How did you evaluate/test your work? What analysis did you perform? What datasets have been used? Provide summary statistics of your dataset. How your evaluation provides support (or not) of your solution. Show results of your analysis, discuss important findings, discuss implications of your analysis to applications.
- **Conclusions**: What are the conclusions of your work? Is your solution adequate for the problem? Are there any highlights of the analysis? What are some ideas for future work?
- **References**: The final report should include the full reference of the libraries, papers, code, tutorials that you have based your project on, your approach to solve the problems, and the tools and datasets that you have employed. Full citation is required. References should be specific and found inside the text, as appropriate.

You should try to fill in as many parts of the report as possible so it is representative of your term work.

Formatting and Style

The suggested length of the project report is **5 pages** and it must be in PDF format. All reports should be formatted according to the new Standard ACM Conference Proceedings Template (pick the *sigconf* format). There are templates for both latex and MS Word users. More information can be found here:
http://www.acm.org/publications/proceedings-template

Project Report Evaluation

The *project report* will be evaluated based on the following mark breakdown.

Final Report Component	Weight	Due Date
Introduction/Motivation	10%	Is the domain interesting? Is the motivation valid? Looking for clear motivation that encourages the reader to read on; clear objectives of the analysis and need for a data-driven approach.
Data and Data Analysis	20%	Is the data analysis approach and methods appropriate and well described? Looking for clear description of the data and processing dimensions; clear description of the analysis to be performed and its need for understanding the domain.
Architecture of Proposed Solution	30%	Is the proposed solution adequate for the data? Is the technical material correct? Are sufficient details provided? Looking for clear and well documented system architecture and data flow in the system. Clear discussion of any limitations.
Evaluation/Results/Conclusions	30%	Is the system evaluation adequate? Is the analysis “interesting”, or just a “toy” analysis? How original, important and well defined are the questions posed? Are there any novel/interesting applications of the analysis? Looking for clear and conclusive set of tests and analytics performed; comprehensive discussion of results.
Style and Language	10%	Is the report well organized? Is the write-up clear and the language adequate? Are results presented in the most appropriate manner? Are figures and tables used appropriately? Looking for good overall organization, language, and style.

How to Submit?

```
% submit 4415 project report.pdf code.zip team.txt
```

Deliverable 3: Project In-class Presentation (30%)

The in-class presentation should be seen as your opportunity to present your hard work in class, your ideas, your approach and solution, your results, and discuss further implications for future work. The expectation is that most of the work has been completed and any major results are available to share with an audience. At this stage you should be able to tell a story about your project, even if parts of it have not been completely implemented or solved. The title of the presentation should include the course title and an indication that this is the project's in-class presentation, the title of the project, your name and contact information.

The following are some guideline, tips and advice for preparing your presentation.

- You (or your group) will have 15 minutes to present your work in the classroom. Another 2 minutes will be allocated for question answering.
- You should prepare a set of ~15 slides, given that a slide should take around a minute to talk about on average.
- Presentations should be organized into thematic units. A typical outline includes:
    - Motivation of the project, its importance and potential applications.
    - Description of the data and processing dimensions.
    - Data analysis to be performed. Questions to be answered.
    - Architecture Overview. Components of the solution. Data flow in the system.
    - Highlights of the results.
    - Interesting variations and limitations of the system.
    - Concluding remarks.
- The talk should be self-sufficient, meaning that you should not make any assumption about prior knowledge of the audience or previous well-known results. All concepts should be introduced and appropriate notation or language should be used consistently throughout the presentation.
- Focus on the essential parts of the project and avoid too-many technical details. The goal is to give a summary of the project and convey the contribution of your work to other people. At the same time, you should make sure that important content is adequately covered.
- Prepare the slides carefully. Text should be easily readable and slides should not be overloaded with content. Avoid full text sentences and use of code snippets or math, unless necessary.
- **Practice** the talk several times, and time yourself to make sure you are within the time bounds.

Formatting and Style

For consistency, you are encouraged to use the official YorkU ppt template found here:
http://toolbox.info.yorku.ca/tools/templates/
(You'll be prompted for your passport YorkU account.)

More Advice & Tips

Some interesting advice on how to give a bad talk by David A. Patterson (UC Berkeley):
https://people.eecs.berkeley.edu/~pattrsn/talks/BadTalk.pdf

Some design tips for beautiful presentations:
https://visage.co/11-design-tips-beautiful-presentations/

In-class Presentation Evaluation

The *in-class presentation* will be evaluated based on the following mark breakdown.

In-class Presenation Component	Weight	Due Date
Introduction/Motivation	10%	Clear motivation that engages the audience; clear objectives of the analysis and need for a data-driven approach.
Data and Data Analysis	20%	Clear description of the data and processing dimensions; clear description of the analysis to be performed and its scope.
Architecture of Proposed Solution	30%	Clear description of the system architecture and data flow in the system. Clear discussion of any limitations.
Evaluation/Results/Conclusions	30%	Clear and conclusive set of tests and analytics performed; comprehensive discussion of results.
Style and Language	10%	Overall organization, language, and style.

How to Submit?

```
% submit 4415 project presentation.pdf team.txt
```