MCDA5570: Big Data group project

Purpose: to demonstrate solid understanding of Big Data tools (Spark/Hadoop ecosystem) and ability to use them for data analysis

Groups: you should form groups of 3-5 students (exceptions are possible) by **February 23. Each group should send me an email with participants: full names, A# (nikita.neveditsin@smu.ca). If you can't find a group, please let me know as soon as possible.**

1. Find data to analyze (**due date: March 5**):

- It can be any dataset/text data/logs/other
- Data can be structured/unstructured/semi-structured
- It should be large enough (from ~10 megabytes to 1 gigabyte).
  - If you found some very large data (>1 gb) you can just get a random part of it
  - If you found a good data but it's considerably less than 10 megabytes you still can use it – just send me an email for approval (nikita.neveditsin@smu.ca)
- You can even use data from other courses if it's not protected by NDA

Each group should submit (Brightspace) a link to the data(set) and brief description of the data(set) by March 5. If you would like to analyze authorization logs, please let me know (I'll provide the logs to you). 1-2 groups can do it. First come first serve.

2. Using Big Data technologies (Hadoop/Spark ecosystem) analyze the data to get some insights from it

- You are expected to use the HDP instances that you used on Workshops
- You may use other Big Data technologies/services (for example, cloud ones) as long as they provide Hadoop/Spark capabilities (it won't affect your points)
- Try to use various Big Data tools (e.g., Pig for ETL, Hive and/or Spark SQL for queries, maybe Spark MLib to perform some classification/clustering/prediction, Zeppelin for visualizations)
- You may also use Tableau for visualization of results if you wish

4. Write a brief report with the following sections (5-20 pages: please be concise – you won't be marked for number of pages) (**due date: March 30**):

- Description of data
- Brief description of what is done with the data (ETL if applicable, data flows, etc.)
- Technologies used: provide code, scripts, queries.
- Results, insights (graphs/tables and their description)
- Any comments on the technologies that you learned in the course and/or used in your project (help future students to get better experience from workshops):
  - What was difficult to understand?
  - What was too easy to understand?
  - Maybe something was not covered but you'd like it to be covered?
  - Any other comments are welcome

Each group should submit (Brightspace) a report by March 30.

3. Prepare a 15-minutes presentation based on your report (**April 6**)

Marking criteria:

- Variety of Big Data tools used
- Complexity of the project
- Proper use of tools
- Quality of insights
- Quality of presentation
- All deadlines are met