3252 TERM PROJECT

The Project

The 3252 Term Project is your opportunity to showcase what you've learned in the course. Your objective is to either research and report on a current evolving area of Big Data technology or to build a small but scalable application of your own. Some examples of projects/reports that would be appropriate:

- Find an appropriate dataset & analyze it using Spark's MLLib
- Develop a small concurrent application
- Develop a demo of reactive programming
- Research and explain differences between concurrency models such as Futures, Goroutines and Actors
- Research and explain the basic idea of Paxos or Raft
- Try out Jepsen on a distributed database
- Compare/contrast HBase & Cassandra or other NoSQL DBMS's
- Explain how to do the equivalent of a relational join in MongoDB
- Write an assessment of VoltDB or another NoSQL DBMS
- Experiment with Neo4j and time various operations
- Explain the difference between the different levels of serialization (e.g. Repeatable Read, Snapshot Isolation, etc.) of a relational database
- Compare and contrast RapidMiner and Pentaho (or other tools)

Data

The data you use for your analysis should be real (i.e. not randomly generated). You can use any source as long as it is open data or you have written approval to use it. The following are links to a variety of interesting datasets you might consider using:

https://www.analyticsvidhya.com/blog/2016/11/25-websites-to-find-datasets-for-data-science-projects/
https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/
https://www.ontario.ca/search/data-catalogue?sort=asc
http://open.canada.ca/en/open-data
http://blog.yhat.com/posts/7-funny-datasets.html
http://koaning.io/fun-datasets.html

Topic Approval

Prior instructor approval of your choice of project is not required for any of the suggestions above. If you would like to undertake a different topic please submit your idea to the instructor via email for feedback and approval by Week 9 and preferably earlier.

Requirements & Due Dates

You must submit two documents, a report and a presentation. The report is due the second-last class of the term and must be submitted to Blackboard. The presentation is due the day before you present, which may be the second-last or last class depending on scheduling.

Report

The precise structure of your report will depend on the nature of the topic but the report should *be of a quality that is suitable for a management briefing*. It should be about 8-10 pages in length (single-spaced). Any code and samples of data should be included as an appendix, not in the main body of the report. Data samples should be limited to less than a page. It should be structured as follows:

- Introduction
- Objectives
- Findings
- Appendices

Presentation

The presentation should be five PowerPoint slides long (excluding title slide). You will only have five minutes to present, which will go by quickly, so please rehearse and time your presentation in advance.

The presentation is due the day before you present so we won't lose time in class (for example, waiting for USB drivers to install).

Marking Scheme

Marks will be allocated as follows for a total out of 40:

- Suitability as a report and presentation to management – 10 marks
    - Spelling
    - Explanation of use of technical terms
    - Formatting
    - Easy to follow
- Correctness and thoroughness of analysis – 15 marks
    - Use of appropriate techniques if data analysis is part of the project
    - Breadth and depth of analysis
- Plausibility of conclusions – 10 marks
- Difficulty/Novelty – 5 marks
    - How challenging a project was it?
    - Is this an interesting and different analysis?