SENG 550 – Scalable Data Analytics

Project Proposal

Projects should propose solutions that are scalable, i.e., "big data" solutions. You can use "small data" to develop, test, and demo your solutions, but your solutions should scale to larger datasets. This implies that you need to use a big data framework, e.g., Hadoop or Spark or Flink, to implement your solution.

Deliverables

Submit a short (1-page max) proposal that describes what your group wants to work on and submit it on Dropbox (on D2L) due November 3, 2023

Make sure to mention the following information such as:

1. Name and student ID of group members.
2. Title of project.
3. Project description.
4. Key project objectives.
5. Tools and technologies required.
6. Data Sources.
7. Expected Outcome of the project
8. Details of the project management plan such as – How will tasks be assigned, managed, and tracked? How will version control be set up?

Project Resources

Here are some resources that you can use for your project.

1) Databricks - www.databricks.com - provides a small-scale managed Spark cluster

2) https://cloud.google.com/dataproc - explore whether Google offers free credits for students

3) https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-spark.html - explore whether Amazon offers free credits for students

4) U of C Teaching and Learning Cluster - https://rcs.ucalgary.ca/index.php/TALC_Cluster - Free access to on-demand Spark cluster.

5) Cybera Rapid Access Cloud - provides VMs - https://www.cybera.ca/rapid-access-cloud/

Project Ideas

1. Propose your own project that involves some useful data analysis (e.g., related to your work or thesis topic)
2. Analysis of stack overflow data - https://www.ics.uci.edu/~duboisc/stackoverflow/ - build a classifier for each user that can predict whether that user will be able to answer a given question or not. More ideas can be found in the link.
3. Predict trip times of NYC taxis (https://www.kaggle.com/c/nyc-taxi-trip-duration/data)
4. Find publicly available Web server access logs - from these logs try to identify users with similar navigational behavior, e.g., in terms of URLs visited and/or in terms of whether they are robots or real users.
5. Tackle one of the COVID challenges (or a subset of a challenge) outlined at https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks?taskId=568. The proposed solution should be scalable, i.e., in Spark or similar frameworks, and should be original, i.e., not developed by others.
6. Propose a project based on publicly available NHL data (e.g., https://www.kaggle.com/martinellis/nhl-game-data?select=player_info.csv)

Project Datasets

The below datasets can be used to come up with some great project ideas.

1. Traffic Crashes – https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if
2. Bicycle Thefts Open Data – https://data.torontopolice.on.ca/datasets/TorontoPS::bicycle-thefts-open-data/about
3. Speed Camera Violations – https://data.cityofchicago.org/Transportation/Speed-Camera-Violations/hhkd-xvj4
4. Medical Insurance Fraud Detection – https://data.cms.gov/provider-summary-by-type-of-service/medicare-part-d-prescribers/medicare-part-d-prescribers-by-geography-and-drug
5. Auto Theft Open Data – https://data.torontopolice.on.ca/datasets/TorontoPS::auto-theft-open-data/about
6. Red Light Camera Violations – https://data.cityofchicago.org/Transportation/Red-Light-Camera-Violations/spqx-js37
7. Reviews on YELP – https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset
8. Log Analysis – https://www.sec.gov/about/data/edgar-log-file-data-sets