Big Data – Fall 2021

Project Administration, Review and Evaluation

1. Project Mechanics and Team Forming

Projects are performed in teams of three (3) students. In your work, you must use either Hadoop or Spark and your solution must scale for large data. Note that you can do analysis and visualization of the results you produce using Hadoop/Spark results using your laptop/desktop.

Your code/scripts must be made available on GitHub and the outputs of your project must be *reproducible* -- you should include enough information so that others can re-run and reproduce the results you report.

2. Part 1: Profiling and cleaning a dataset – Due November 22nd, 2021

A dataset will be assigned to each group. The 3 project members will collaboratively explore the assigned dataset to identify data quality issues.

You should follow the best practices discussed in the Data Cleaning lectures, and apply the techniques you learned as well as other techniques available in the literature and in open-source tools. You will:

1) Profile the data: you will explore and learn about the data
2) Look for different types of data quality issues, including incorrect values (e.g., typos – brklyn), inconsistency between values (e.g., zipcodes or city names), missing data, outliers
3) Clean the data and create a new version of the dataset. Discuss your cleaning decisions, in particular when it was not clear what the right approach should be.

You can use existing tools, e.g.,

- Datamart Geo (https://pypi.org/project/datamart-geo)
- Datamart Profiler (https://pypi.org/project/datamart-profiler, https://docs.auctus.vida-nyu.org/python/datamart-profiler.html)
- OpenClean (https://github.com/VIDA-NYU/openclean)

and you can also implement your own methods.

Deliverable: You will submit a *reproducible and documented* Jupyter or Zepellin notebook that shows the different steps you followed.
Here's a an example of a reproducible and well-documented notebook:
https://github.com/VIDA-NYU/openclean/blob/master/examples/notebooks/Parking%20Violations%20-%20Profiling%20and%20Cleaning%20Example.ipynb

This will count as your Homework Assignment 3.

Datasets: Your group will work on one of the following datasets:
311 Service Requests from 2010 to present
https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

Historical DOB Permit Issuance Housing & Development
Construction permit issued between 1989 and 2013.
https://data.cityofnewyork.us/Housing-Development/Historical-DOB-Permit-Issuance/bty7-2jhb

DOB Job Application Filings
Job applications submitted through the Borough Offices
https://data.cityofnewyork.us/Housing-Development/DOB-Job-Application-Filings/ic3t-wcy2

Motor Vehicle Collisions - CrashesPublic Safety
details on crash events
https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Citywide Payroll Data (Fiscal Year)
how the City's budget is being spent on salary and overtime pay for all municipal employees.
https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e

NYPD Complaint Data HistoricPublic Safety
This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD)
https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

3. Part 2: Data cleaning at scale – Due December 12th, 2021

The goal of this task is to both refine the work you did by improving the strategies you used for your dataset, as well as to scale them to handle a large number of files.

You will Find at least 10 other datasets in NYC Open Data whose fields overlap with the dataset you worked on. You can find these by looking for datasets that contain similar columns (remember the lecture on similarity?)

1) Apply the techniques you used for Part 1 and measure their effectiveness. For this, you can use precision and recall (https://en.wikipedia.org/wiki/Precision_and_recall). Note that to measure recall, you will need to manually inspect the data. If the data is too large, select a sample that is large enough to give you confidence. To determine the required sample size, you can use a sample size calculator (e.g., https://www.surveysystem.com/sscalc.htm)
2) Improve/refine your techniques to cover the new data and compare its effectiveness with your original approach.
3) Create reference data for the data types you cleaned.
4) *Extra credit: contribute your reference data to https://github.com/VIDA-NYU/reference-data-repository*
5) Describe how you would run your approach on all NYC Open Data datasets. In particular, if given a list of all datasets, how would you run your cleaning strategy over all datasets?

You should submit the pdf of your report on Brightspace under “Final Project: Report”. The report should include

- The name of the group performing the evaluation and the names/nyuids of the group members
- A link to the github repo

Your github repo should contain your code, reference data, and your report.

All of the NYC open datasets are available on Peel HDFS at:
/user/CS-GY-6513/project_data
Do not make copies of the original datasets -- read them from our shared directory.

4. Part 3: Extra Credit - Reproducibility evaluation – Due December 19th, 2021 (20 points)

You will run the notebook submitted by another group (which we will assign to you) and produce a reproducibility report that describes what you did to re-run the experiment. For example, were all dependencies declared? Were you able to run the notebook as is? Did you have to install new packages? Were you able to get the same results?

You should also provide feedback on the techniques used and on potential improvements.

Your report will be embedded in the original notebook and will be shared with all students in the class.

You should submit the notebook file on Brightspace unde “Final Project: Extra credit”. The notebook should include

- The name of the group performing the evaluation and the names/nyuids of the group members
- The name of the group being evaluated and a link to their github repo
- Your reproducibility report and feedback

For more information and guidelines on reproducibility evaluation, see
https://reproducibility.sigmod.org/

Dec 12th: Project report is due. For the report, you should follow the format of a research paper (see suggested outline below), and I suggest (but do not require) that you use LaTeX (Overleaf) and the ACM format (https://www.overleaf.com/gallery/tagged/acm-official#.WOuOk2e1taQ).

You are not expected to complete a paper that is ready to be published, but I expect your report to be a starting point for a publication.

The suggested structure for the report and evaluation metrics are as follows:

- Introduction – 5 points

- Problem formulation – 10 points
- Related work – 10 points
- Methods, architecture and design – 30 points
- Results –25 points
- References (cited in the report and related work)

In addition, we will evaluate
- Technical depth and innovation – 10 points
- Code repository, correctness, and readability – 10 points

Dec 13th-20th: Project presentations (20 points)
You will give an 8-minute presentation about your project in which you will summarize your key findings/contributions. You should use Google Slides for your presentation and include the names of the group members, project name, and link to github repo.

Here's a suggested outline:

- State the question(s) you investigated
- Describe the method and data you used to answer the question(s)
- Present your findings and insights you gained
- Discuss challenges you encountered, and limitations of your approach

You should not repeat everything that is in your report. Instead, focus on what you deem is most interesting about your methods, findings, and the experience you gained.

Your instructor and classmates will be your audience. All teammates must be present on the day of their team's presentation.

Groups will be randomly assigned to one of the presentation dates.

Report review questions for graders
1. Briefly summarize the project.
2. What are the key strengths or positive aspects of the work?
3. What are the limitations of the work?
4. Were previous approaches considered?
5. Is the problem clearly stated? Does it make sense?
6. Is the project related to big data and the topics covered in class?
7. Are the findings and methods clearly described? If not, which paragraphs or statements are unclear.

8. Is there a new or useful component to the project?
9. Is there a Github code repository? Is the code readable and working? Are the results/outputs described in the report reproducible?