IS3107 Data Engineering

Academic Year 2022/2023 Semester 2 Lecturer: Dr. Yiliang ZHAO

Group Project Guidelines and Grading Criteria

Updated as of 10th Jan 2023

Project Description

This hands-on project aims to reinforce and integrate the major concepts of data engineering that you have learned from the lectures and tutorials into a basic example of an end-to-end data pipeline, based on a real-world dataset.

After completing this project, you will gain a better understanding of how to leverage (1) open-source libraries, (2) DBMS/Data warehouses, and (3) workflow management tools, together, to build a data pipeline and solve a real-life problem to support downstream applications such as dashboard/visualization or machine learning applications.

Example project ideas

1. To better keep track of the state-of-the-art research progress, it’s common for researchers to get regular alerts about the latest publications of other leading scholars in relative fields. DBLP is a popular CS bibliography archive that provides open bibliographic information on major computer science journals and proceedings. Many researchers in the CS community use its reliable open-data services to access and maintain publication profiles, view coauthor information, and obtain links to electronic editions of publications, etc. DBLP is maintained by a specific team and is updated to include bulks of indexed conf/journal volumes on a daily basis. The core implementation of the project focuses on creating a data pipeline that regularly imports updated publication data from DBLP to the databases and data warehouses. The pipeline is able to support a few downstream applications, such as publication trend analysis, rising researchers in each field, etc.
2. This project aims to analyze trends in tweets on a particular topic or by a specific user. This will be accomplished by collecting a dataset of tweets using the Twitter API, preprocessing and cleaning the data, and using statistical and visualization techniques to uncover trends and insights. The analysis may include identifying the most common words or hashtags used, examining the sentiment of the tweets, determining the influence of certain users, and exploring any correlations between tweet activity and external events or factors. The results of this analysis will provide valuable information for understanding the conversation and sentiment surrounding a particular topic on Twitter, and may be useful for businesses, organizations, or individuals seeking to track and understand social media trends.
3. This project aims to analyze Airbnb listings in a specific city or region to understand trends in pricing, amenities, and overall demand. By collecting and analyzing data on various aspects of Airbnb listings, we hope to gain insights into the local vacation rental market

and make informed recommendations for property owners seeking to optimize their listings on the platform. Additionally, we aim to provide a comprehensive overview of the Airbnb landscape in the chosen location to assist travelers in finding the best options for their needs and budget. Through this analysis, we aim to improve the overall experience for both Airbnb hosts and guests in the chosen area.

Disclaimer

Individuals in the same group will generally receive the same scores for all components of the Group Project unless feedback is received that a particular member is only superficially participating and not doing actual work.

Important things to note:

Please name your files in the following format `report_<group number>.pdf` for Group Project Final Report and `presentation_<group number>.pptx` for Group Project Presentation.

Project Component	**Due Date and Time**	**Submission Items**
Group Project Presentation	21st Apr 2022 @ 23:59	Presentation Slides Recorded Presentation
Group Project Final Report	21st Apr 2022 @ 23:59	PDF Report Completed Code

Group Project Final Report

When to submit: 21st April 2023, Friday at 23:59
Who to submit: A representative of each group to submit
What to submit: PDF Report (Please include URL to your Github Repo in the report)
Where to submit: Canvas > Assignments > Project > Group Project Final Report

Guidelines

You are to submit an 8-10 page PDF report using reasonable fonts and spacing. A suggested outline of the Group Final Report includes:

1. A description of the use case that you want to address. Discuss why the problem is useful and important.

2. Some basic facts about the data, including the following:

- Source of the dataset(s)

- Describe the dataset(s) - E.g., number of observations, number of variables, type of variables. (string, integer, etc.) You may present the variables in a tabular form.
- An explanation of what you think are the interesting aspects that you can look at in the dataset(s)
- A discussion of what you hope to mine from the dataset(s)/some hypotheses that you might have in mind that you hope to verify - E.g. spot increasing trends etc.

3. A detailed discussion of the following:

- Pre-processing steps to sanitize/manipulate/combine your dataset(s).
- Design of databases and data warehouses in the form of an ER diagram. Explicitly specify how you define the primary/foreign keys of the tables.
- Snapshots of tables with data for both databases and data warehouses.
- Describe the rationales that you have considered for your design of the pipeline
- Snapshot of the graph visualization of your pipeline.
- Snapshot of the tree view of your pipeline after triggering your DAG.
- Run time of each step in your pipeline.
- What visualizations (line chart/bar chart/histogram, etc.) and machine learning models you have used (E.g., Unsupervised? Supervised?)
- The insights that you have gathered from the downstream applications.

4. Discussions

- What are the performances of the pipeline (speed, accuracy if ML models are used)?
- How can the model be integrated into the respective business process?
- What insights can be obtained from the dashboards?
- Etc.

5. Conclusions

Grading Criteria

The Group Project Final Report weighs 20% of your overall final grade. Your report will be evaluated based on:

(1) Clarity and completeness of the report
(2) Appropriateness of the models and methods applied for data processing and analysis
(3) The usefulness of your analysis in other similar real-world problems
(4) Concise summarization of your work
(5) Reasonability of the discussion of advantages and limitations of applied methods to the defined problem

Group Project Presentation

Who to present: Every member of the group should have a share in presenting
*What to present**: 12-15 minutes** worth of slides containing highlights of your project (data pipeline and analysis)*
When to submit: 21st Apr 2023, Friday at 23:59
What to submit: Presentation Slides and the Presentation Video
*Where to submit: **Canvas > Assignments > Project > Group Project Final Presentation***

Guidelines

1. You are to prepare **12-15 minutes** worth of slides for presentation and record your group presentation. The flow and content of the presentation can follow that of the Group Project Final Report.
2. You must include the title and group members at the beginning of the video, either as the first page of your slides or as an overlay text. Make sure that you leave the title for long enough of a duration to be read (up to 5 seconds).
3. The submitted video can be the recording video of your computer screen. Here, some screen capture software is recommended (For Windows users, **Camtasia** and **Camstido**. For Mac OS users, the **QuickTime** player). You can also use **zoom** to record the presentation session.
4. Record your presentation in the highest possible image and audio quality. Please ensure that your submitted video will play.

Grading Criteria

The Group Project Presentation weighs 20% of your overall grade. You are to submit your slides and code together with the Group Project Final Report on 21st April 2023 by 23:59. As long as your slides and codes are complete and error-free, you will be awarded 5% (specifically, the code shall be able to run smoothly and produce the required results). The remaining 15% of the grading of the presentation will be dependent on:

1. The clarity in highlighting the background and motivation of the use case.
2. Logical and reasonable explanations of selected tools and methods
3. Comprehensiveness and coherence of the presentation
4. Readiness and confidence of presenters
5. Self-explainability of slides content
6. Good time management during the presentation

All the best for your project!