IS3107 Data Engineering

Academic Year 2024/2025 Semester 2 Lecturer: Dr. Yiliang ZHAO

Group Project Guidelines and Grading Criteria

Updated as of 4th Jan 2025

Project Description

This hands-on project provides an opportunity to apply and integrate key concepts of data engineering covered in lectures and tutorials by developing an end-to-end data pipeline using a real-world dataset.

By completing this project, you will deepen your understanding of how to leverage:

1. Open-source libraries,
2. DBMS/Data warehouses, and
3. Workflow management tools,

to build efficient data pipelines that solve real-world problems and support downstream applications such as interactive dashboards, visualizations, or machine learning models. This project is designed to bridge theory and practice, enabling you to work with diverse datasets, apply innovative problem-solving, and gain experience building scalable pipelines for real-world applications.

Example project ideas

The following example project ideas are provided as references to spark your creativity and help you conceptualize the scope and direction of your own project. Feel free to draw inspiration from these examples or adapt them to suit your interests and goals. You are encouraged to explore new ideas and tackle challenges that resonate with your personal or professional aspirations.

1. **Tracking Cutting-Edge Research Trends with DBLP**
   Build a pipeline to analyze and visualize trends in computer science research using the DBLP dataset. Automate the regular ingestion of updated publication data and support downstream applications like identifying rising stars in various CS fields, mapping the co-authorship network, and analyzing publication trends over time. This tool could assist researchers in discovering new collaborations, conferences, and cutting-edge topics. Enhance it further by integrating NLP techniques to classify publications into trending research areas.

2. **Social Media Insights from Twitter Data**
   Develop a pipeline to analyze trending topics on Twitter by collecting tweets via the

Twitter API. Your pipeline could uncover patterns like frequently used hashtags, tweet sentiment, or the correlation between social media activity and real-world events. Build interactive dashboards that visualize emerging conversations and influencers. Extend this idea by predicting trends using machine learning or analyzing bot activity patterns. This project could benefit businesses, policymakers, and social media analysts aiming to understand public sentiment.

3. **Optimizing Airbnb Listings for Hosts and Guests**
   Create a pipeline to analyze Airbnb listings in a city to uncover insights on pricing strategies, popular amenities, and seasonal demand patterns. Use the data to recommend optimal pricing strategies to hosts and highlight the most attractive listings for travelers. Add a machine learning component to predict demand based on listing attributes or external factors like local events. This project could empower property owners to optimize revenue and assist travelers in finding tailored options for their stays.

4. **Smart City Traffic Management Using Public Data**
   Build a pipeline to process and analyze real-time traffic data from public APIs or IoT sensors. Identify bottlenecks, predict congestion, and recommend alternate routes for urban commuters. Use geospatial analysis to visualize traffic patterns and propose data-driven interventions for urban planners. Integrate this with weather data or event schedules for a comprehensive solution to traffic management challenges.

5. **E-Commerce Customer Behavior Analysis**
   Design a pipeline to analyze customer behavior in an e-commerce store by processing sales, clickstream, and customer review data. Identify purchasing trends, seasonal preferences, and factors driving customer satisfaction. Build dashboards that provide actionable insights for marketing teams or use machine learning models to predict customer churn and recommend personalized products.

Disclaimer

All group members are expected to contribute equally to the project. While group members will generally receive the same scores for all components of the project, points will be deducted for any member who is found to have not actively participated or contributed meaningfully. Any concerns about unequal contributions should be reported to the instructor or teaching assistants for review.

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	27 Apr 2025 @ 23:59	Presentation Slides Recorded Presentation
Group Project Final Report	27 Apr 2025 @ 23:59	PDF Report Completed Code

Group Project Final Report

Submission Details

- When to submit: 27 April 2025, Sunday, by 23:59
- Who to submit: A representative from each group
- What to submit: A PDF report (include the URL to your GitHub repository in the report)
- Where to submit: Canvas > Assignments > Project > Group Project Final Report

You are required to submit an **8-10 page PDF report** (excluding the title page and appendix) using reasonable fonts and spacing.

The suggested outline for your report includes the following:

1. **Use Case Description**
    - Describe the use case you aim to address.
    - Explain why the problem is both useful and important.
2. **Dataset Overview**
    - Provide the source(s) of the dataset(s).
    - Describe the dataset(s), including:
        - Number of observations
        - Number and types of variables (e.g., string, integer, etc.)
    - Optionally, present the variables in a tabular format.
    - Highlight interesting aspects of the dataset(s).
    - Discuss what you hope to mine from the dataset(s) and hypotheses you aim to verify (e.g., identifying trends).
3. **Detailed Analysis**

    - Pre-processing steps for sanitizing, manipulating, or combining dataset(s).
    - Design of databases and data warehouses (include an ER diagram and explicitly define primary/foreign keys).
    - Snapshots of data tables for databases and data warehouses.
    - Rationales for your pipeline design.
    - Include snapshots of:
        - The graph visualization of your pipeline
        - The tree view of your pipeline after triggering the DAG
    - Document the runtime of each pipeline step.
    - Visualizations (e.g., line charts, bar charts, histograms) and machine learning models (supervised or unsupervised) used.
    - Insights derived from downstream applications.
4. **Discussion**
    - Performance evaluation of your pipeline (e.g., speed, accuracy if ML models are used).
    - How the model can integrate into the respective business process.
    - Insights gathered from dashboards.
    - Additional discussions (e.g., challenges and future improvements).
5. **Conclusion**
    - Summarize your key findings and their implications.

Grading Criteria

Your report will be assessed based on the following:

1. Clarity and completeness of the report.
2. Appropriateness of models and methods applied for data processing and analysis.
3. Usefulness of your analysis in similar real-world scenarios.
4. Conciseness and thoroughness of your work summary.
5. Reasonability of the discussion on the advantages and limitations of your methods for the defined problem.

Group Project Presentation

Submission Details

- **Who presents**: Every member of the group must participate in the presentation.
- **What to present**: A 12-15 minute slide presentation highlighting key aspects of your project, including the data pipeline and analysis.
- **When to submit**: 26 April 2025, Sunday, by 23:59.
- **What to submit**: Presentation slides and a recorded video of the presentation.
- **Where to submit**: Canvas > Assignments > Project > Group Project Final Presentation.

Guidelines

1. Prepare slides for a 12-15 minute presentation and record your group presentation. The flow and content of the presentation can follow the structure of your **Group Project Final Report**.
2. At the beginning of the video, include the project title and group members. This can appear as the first slide or as overlay text on the video. Ensure this information is visible for at least **5 seconds** for easy readability.
3. For recording your presentation, you may use screen capture software. Recommended tools include:
    a. Loom (Web-based)
    b. ScreenRec (Cross-Platform)
    c. Zoom (Integrated Recording)
    d. OBS Studio (Cross-platform)
    e. Camtasia (Windows, Mac)
4. Record your presentation in the **highest possible image and audio quality**. Before submission, ensure the video plays correctly and is free from technical issues.

Grading Criteria

Your presentation will be graded based on the following:

1. **Clarity**: How well the background and motivation of the use case are highlighted.
2. **Logical Explanations**: Reasonability of the tools and methods selected for the project.
3. **Comprehensiveness**: The coherence and completeness of the presentation.
4. **Presentation Skills**: Readiness, confidence, and engagement of the presenters.
5. **Slide Quality**: Self-explanatory and visually clear slide content.
6. **Time Management**: Adherence to the 12-15 minute time limit.

To excel in this group project, focus on leveraging the knowledge and skills you have acquired throughout the course to design innovative and impactful solutions. Approach the project with curiosity, collaboration, and creativity, as it offers a unique opportunity to tackle real-world challenges while refining your technical and problem-solving abilities. Remember, this is not only a chance to showcase your proficiency in data engineering but also to develop teamwork and communication skills that will serve you in your future career. We look forward to seeing your innovative ideas and exceptional work. Good luck!