

Final Team Project

Start Assignment

- Due Monday by 11:59pm
- Points 200
- Submitting a text entry box, a website url, a media recording, or a file upload
- Attempts 0
- Allowed Attempts 3

Introduction

The final project will assess the development of a production-ready data pipeline. You will be assigned to groups by your instructor. You and your teammate(s) will work closely to find a dataset, build an ETL or ELT pipeline to process it, and produce a practical output from it. Example outputs might be a triggered email if some condition is met or a dashboard to visualize the data.


Final projects and datasets from *prior courses cannot be used*. It is **required** that you and your team use **GitHub** as a code hosting platform to manage version control and collaboration during this project.

Project Timeline

- Module 1 (by the end of Week 1): The course instructor grouped students into teams of two to three members.
- Module 3 (by the end of Week 3): The team representative submitted the Initial Final Team Project Proposal Form to describe the concept for the project and the data that will ultimately support the concept.
- Module 5 (by the end of Week 5): The team submitted a status report as well as all materials developed to that point for instructor feedback. For the final submission, this feedback will have to be sufficiently addressed.
- Module 7 (by the end of Week 7): Each team should submit the following deliverables for the course project in the final week:
 1. **Design Document:** One **PDF** document outlining your system architecture. This should describe the source data, the specifics of your pipeline ETL or ELT, and the output. It will also include “next steps” that outline any shortcomings of your current system and what could be done to improve it. **Provide the link to your GitHub repository at the top of the Design Document.**
 2. **Demo Video:** One ~15 minutes video presentation by all team members. Submit one **mp4** file. This should be presented as a “demo day” presentation to a technical audience. You will present what your pipeline does, your system architecture, and how you interact with your system.

*It is critical to note that no extensions will be given for any of the final projects' due dates for any reason, and final projects submitted after the final due date will not be graded.

Project Datasets

You will be responsible for finding multiple datasets for this final team project in this course. Several free data resources' links are provided within the [Resources](https://usd-msads.github.io/resources/)  (<https://usd-msads.github.io/resources/>) page on the MS-ADS GitHub website.

Requirements

Divide the work equally between the team members for the following steps, and *everyone* needs to code and review the code. You **must** use SQL to transform and load your datasets. The database engine you run this SQL code in should be MySQL. This project requires that you and your team submit a link to your GitHub repository containing your code, create a design document, and conduct a demo day video presentation.

For your GitHub repository and code:

1. Your pipeline must be fully reproducible from the GitHub repository.
 - Ideally, your pipeline is deployed using infrastructure as code (IaC)
 - Document the deployment process in the README.
 - Manual processes can be required in the deployment process, but these must be fully documented in your README.
2. Your pipeline must load data to a persistent data store.
 - There are no restrictions on the type of data store.
 - Your pipeline must include at least one SQL transformation of your data.
3. You must address and incorporate any feedback provided by the instructor on your Module 5 status report.
4. The README must include the following:
 - An overview of the contents of the repository.
 - How to deploy your pipeline.

- How to monitor your pipeline.

Optional Strongly Suggested Requirements:

1. You must have a way to monitor your system.
2. You should perform code reviews.
 - Code reviews only become a requirement after Module 6 of the course.
3. Your repository should be configured for continuous integration.
 - Automated tests
 - Linters
4. Your pipeline should be triggered.
 - The trigger can be time-based (e.g., run every day) or event-based (e.g., run whenever new data is uploaded).
 - If your dataset is static, you should build a way to simulate data updates (e.g., load the data in chunks instead of all at once).

For your design document:

- You must provide the link to your GitHub repository at the top of the document.
- You must document your source datasets.
 - Where did you find the dataset, and why did you choose it?
- You must document the output of your pipeline.
 - Why is the output useful?
- You must include an architecture diagram.
 - A Diagram of the final schema that was used must be present
- You should document the gaps in your system.
 - Will the system scale as the dataset size grows?
 - Is the system secure?
 - Is the system extensible?


For your video presentation:

- Give a 10-15 minute presentation of your pipeline aimed at an audience of data engineers.
- Describe what your pipeline does at a high level.
- Present the architecture diagram and the core components of your system.
- Show how you interact with the system.
 - How is the source dataset ingested?
 - How do you view the output?
 - How do you monitor the pipeline while it is running?

Project Deliverables and Submission Format

- Prepare and submit your Design Document in **PDF** format.
 - **Include a link to your GitHub repository at the top of the design document.**
- Prepare a recorded video presentation of your project using a screencasting tool, such as Screencast-O-Matic or Zoom, to record your screen and provide a voice narration.
 - Ensure that the sound quality of your video is good and each member presents an equal portion of the presentation.
 - Export the video file to an **mp4** format.
 - You may use any recording software you wish. View the [Recording Video Presentation and Submission Guidelines for MS-ADS Students](https://sandiego.instructure.com/courses/17499/files/2562084?wrap=1) (https://sandiego.instructure.com/courses/17499/files/2562084?wrap=1) guide for additional recording instructions.
- Submit the ~15 minutes presentation mp4 video file and Design Document PDF file on the final team project submission page of Canvas. You will use the naming convention **Final-Project-Report-Team-Number.ext** (e.g., Final-Project-Report-Team-1.pdf and Demo-Presentation-Team-1.mp4). **Only one member** of your team will need to submit these deliverables.
- You will submit the peer evaluation form *individually* in [Assignment 7.1](https://sandiego.instructure.com/courses/17499/assignments/253068). (https://sandiego.instructure.com/courses/17499/assignments/253068)

This assignment has [Turnitin](https://help.turnitin.com/integrity/student/canvas/assignments/submitting-an-assignment.htm)  (https://help.turnitin.com/integrity/student/canvas/assignments/submitting-an-assignment.htm) enabled for submissions which means that your instructor will obtain an Similarity Report that

identifies specific parts of your writing that may indicate a high level of matching to external content. You are strongly encouraged to review your work without penalty by activating the [Draft Coach extension in your Google Docs](https://help.turnitin.com/integrity/student/draft-coach/using-draft-coach.htm)  (<https://help.turnitin.com/integrity/student/draft-coach/using-draft-coach.htm>) prior to submitting your work for final grading.

NOTE: Team members may not get the same grade on the Final Team Project, depending on each team member's level of contribution.

To understand how your work will be assessed, view the scoring rubric below.

Click the **Start Assignment** button above to submit your assignment.

Final Team Project Scoring Rubric

Criteria	Ratings						Pts		
Pipeline - Yes/No Requirements 15%	30 pts Meets or Exceeds Expectations Project meets and exceeds all of the stated requirements and implements a successful data pipeline.	27 pts Approaches Expectations Project meets all of the stated requirements but may be lacking adequate implementation needed for a successful data pipeline.		24.6 pts Below Expectations One of the project requirements is missing.	21 pts Inadequate Attempt Two or three of the project requirements are missing.	0 pts Non-Performance Four or more of the project requirements are missing.	30 pts		
Pipeline - Reproducibility 15%	30 pts Meets or Exceeds Expectations Pipeline can be easily deployed. Documentation for deploying the pipeline is clear.	27 pts Approaches Expectations Pipeline is reproducible because documentation for deploying the pipeline is clear.		24.6 pts Below Expectations Pipeline is not easily reproducible.	21 pts Inadequate Attempt Pipeline is not reproducible without the help of the team members who built it.		0 pts Non-Performance Pipeline is not reproducible, and will not work without some modifications.	30 pts	
Addressed Instructor Feedback 10%	20 pts Meets or Exceeds Expectations Feedback provided is addressed and exceeds expectations.	18 pts Approaches Expectations Feedback provided is addressed and approaches expectations.		16.4 pts Below Expectations Feedback provided is addressed but below expectations.		14 pts Inadequate Attempt Feedback provided is inadequately addressed.	0 pts Non-Performance Feedback provided is not addressed.	20 pts	
Design Document - System Rationale 15%	30 pts Meets or Exceeds Expectations Design document is well written and describes the intent of the system as well as any gaps that may need to be addressed in future versions.		27 pts Approaches Expectations Design document describes the intent of the system as well as any gaps to be addressed in future versions.		24.6 pts Below Expectations Design document describes the intent of the system as well as any gaps to be addressed in future versions.		21 pts Inadequate Attempt Design document describes the intent of the system as well as any gaps to be addressed in future versions.	0 pts Non-Performance Design document is missing.	30 pts
Design Document - Pipeline Description 10%	20 pts Meets or Exceeds Expectations The descriptions of the source data and output are clear. The use case for the output is clear.	18 pts Approaches Expectations The descriptions of the source data and output are clear. The use case for the output is clear.		16.4 pts Below Expectations The descriptions of the source data and output are clear. The output is generic or could be sourced in a better format by an existing service.		14 pts Inadequate Attempt The descriptions of the source data or output are unclear.		0 pts Non-Performance The descriptions of the source data are not present.	20 pts

Criteria	Ratings						Pts
Design Document - Architecture Program 10%	20 pts Meets or Exceeds Expectations The associated diagrams are visually appealing and contain all relevant information.	18 pts Approaches Expectations The associated diagrams are present and contain all relevant information.	16.4 pts Below Expectations The associated diagrams present and contain most of the relevant information.	14 pts Inadequate Attempt The associated diagrams are lacking and do not contain all relevant information.	0 pts Non-Performance The associated diagrams are missing.		20 pts
GitHub - Process 10%	20 pts Meets or Exceeds Expectations The git log shows regular contributions from all members of the team. Tests or code quality tools are run automatically on the repository.	18 pts Approaches Expectations The git log shows contributions from all members of the team.	16.4 pts Below Expectations Code review is performed but reviews appear to be cursory. The git log shows contributions from most members of the team.	14 pts Inadequate Attempt Git is used ineffectively.	0 pts Non-Performance Git is not used.		20 pts
Video Presentation 15%	30 pts Meets or Exceeds Expectations Presentation is attractive and well organized. All members of the team participate in the presentation. Focus of the presentation is on the aspects that data engineers are interested in.	27 pts Approaches Expectations Presentation is fairly well organized. All members of the team participate in the presentation. Focus of the presentation is on the aspects that data engineers are interested in.	24.6 pts Below Expectations Presentation is organized. Team participation in the presentation is very unequal. Presentation is not focused on the aspects that data engineers are most interested in.	21 pts Inadequate Attempt Presentation is disorganized. Presentation is not at all aimed at the correct audience.	0 pts Non-Performance Presentation is not present.		30 pts
Total Points: 200							