690A Final Project Guidelines

For your final project, you are required to design and implement a comprehensive data pipeline from scratch that demonstrates all three key elements of the ETL (Extract, Transform, Load) lifecycle. The goal is to apply the skills and tools covered in this course to a real-world data scenario.

Project Requirements:

- Select a data source that is relevant to the project. You can use APIs, cloud storage (e.g., S3), databases, or any other accessible data source.
- Retrieve data programmatically using tools such as Python, SQL, or Boto3.
- Clean, filter, and transform the extracted data using Python and libraries like Pandas or PySpark.
- Demonstrate proficiency in data manipulation techniques and ensure that the data is in the correct format for analysis or storage.
- Load the transformed data into a database, data warehouse (e.g., Snowflake or Redshift), or any analytics tool such as DBT, Streamlit, or Airflow.
- You should also consider performance optimization and scalability when loading your data.
- Automate the pipeline using a workflow orchestration tool such as Prefect or Airflow.
- Include automated scheduling, error handling, and notifications in your workflow.

Tools:

You must use at least one of the following tools from the syllabus:

- Flask for APIs and backend processes
- SQL/PostgreSQL for database interaction
- DBT for analytics engineering
- Docker for containerization and deployment
- AWS services (e.g., S3, EC2, Lambda, RDS, Athena)
- Airflow or Prefect for orchestration
- Pandas or PySpark for data manipulation

Deliverable:

- Submit the code for your data pipeline either by uploading it to GitHub (provide the link) or by zipping the code and uploading it directly.
- Include a brief README file that explains:
- The data source used
- The transformation steps
- The destination of the data
- How the pipeline is automated

Deadline:
Your final project is due by
This project will account for 50% of your total course grade, so ensure it showcases your mastery of data engineering techniques and tools.