ADS507 Team 2 Final Project Plan

GitHub: <https://github.com/amcarr-ds/ads507_data_engineering>

**Current Aim**

Create an ETL or ELT pipeline to access Kaggle data and ingest it into COP28 internal storage.

**Goal State**

1. Automatically access Kaggle data via API.
2. Create a waypoint for data.
3. Perform transformation.
4. Automatically move data to storage.
5. Generate diagram of database structure.

**Questions**

1. Is it better to have one schema or multiple? Can foreign keys be implemented for separate schemas?
2. Does each of the separate tables have an ISO column, or do we need an additional lookup table?
3. How do we implement automatic updates? How are triggers implemented?
4. What is the automation expectation for the final assignment?
5. Assuming a batch update method is better than streaming based on the data, how do we best initiate a change data capture (CDC)?
6. What transformations are needed or necessary?
7. ETL vs ELT: What would work best for this situation?
8. What should the final storage container be? Online MySQL database, AWS S3, other?
9. What is required for deployment?
10. What is required for monitoring?
11. How do we implement security?