Milestone 2

> Deadline: 25th of Jan 2025

In this milestone, your goal is to:

1. Build a data pipeline using Apache Airflow for the functions implemented in Milestone 1.
2. Create a visual dashboard using Apache Superset for analysis.
3. Prepare a PowerPoint presentation summarizing your work, outcomes, and key insights for an evaluation pitch.

Note: Ensure Docker is installed, as Airflow and Superset will be run using Docker containers.

Requirements:

Part 1: Functions & DAG

1.1. Functions

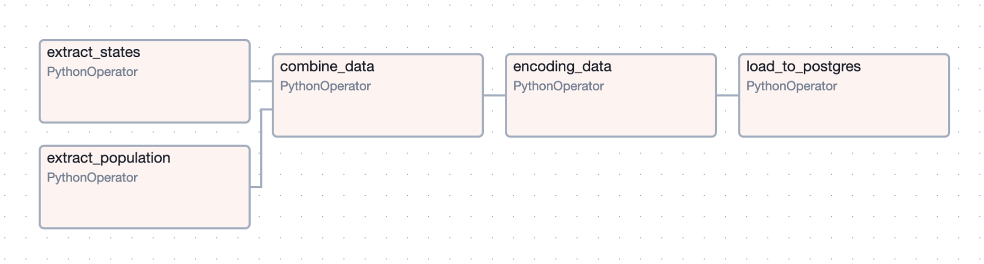
1. `extract_clean`
   - Extracts the dataset from the CSV source file.
   - Handles missing values.
   - Saves the cleaned dataset as `fintech_clean.parquet`.
2. `extract_states`
   - Extracts state names from states.csv (provided in the milestone Google Drive folder).
   - Saves the output as `fintech_states.parquet`.
3. `combine_sources`
   - Combines the datasets from `fintech_clean.parquet` and `fintech_states.parquet`.
   - Saves the merged output as `fintech_combined.parquet`.
4. `encoding`
   - Reads the combined dataset `fintech_combined.parquet`.
   - Applies necessary encodings (implemented in milestone 1).
   - Saves the encoded dataset as `fintech_encoded.parquet`.
5. `load_to_db`
   - Loads the encoded dataset`fintech_encoded.parquet` into a PostgreSQL database.

1.2. Airflow DAG

1. Create an Airflow DAG with the following structure:

   - Task 1: Execute `extract_clean`.
   - Task 2: Execute `extract_states`.
   - Task 3: Combine the outputs using `combine_sources`.
   - Task 4: Apply encoding using `encoding`.
   - Task 5: Load the final encoded data into the PostgreSQL database with `load_to_db`.

2. Ensure all tasks are properly linked using the following dependencies, it should look something like this (not same task names).



Part 2: Data Visualization Using Apache Superset

Use Apache Superset to create an interactive dashboard. Follow these steps:

1. Connect Superset to PostgreSQL
    - Add a connection to the database where the `fintech_encoded.parquet` dataset was loaded.
2. Create a Dashboard to display the following analysis:
    - Basic Info:
        - Total number of loans
        - Average loan amounts
        - Most State with the highest number of loans
        - Number of Unpaid loans
    - Time based analysis:
        - Avg. Loan amount over time
        - Total number of loans over time
    - Tabular View:
        - Display the avg. loan amount, total number of loans, and unpaid loans for each purpose.
    - Analysis Questions:
        - What is the distribution of loan amounts across different grades?
        - How does the loan amount relate to annual income across states?
        - What is the percentage distribution of loan grades (letter grades) in the dataset?
        - Top 5 states with the highest average loan amounts?
    - Extra Analysis (Bonus):
        - Any additional analysis you find interesting.

Part 3: Evaluation Pitch

In this part, you need to prepare a 15 minute pitch presentation to present your work, outcomes, and key insights.

In this presentation you need to showcase the Airflow DAG, Superset Dashboard, and any additional analysis you find interesting.

You can create presentation slides if you want to but it is not mandatory. You can also use any presentation tool you are comfortable with.

Note: Evaluation criterial will be sent out shortly

Submission

Deliverables

1. Airflow DAG scripts (dag file and functions files)
2. Superset dashboard screenshots (or a link to the dashboard if it is hosted)
3. Presentation slides (if any)

Make sure to include all of the files mentioned above in a folder named `Milestone2` and add it to the root drive folder you created in milestone 1.

Deadline

Submit the deliverables by the 25th of Jan 2025.

Best of Luck! 🚀