Data Pipelines with Airflow

Background Information

Our telecommunications company, MTN Rwanda, has a vast customer base, and we generate a large amount of data daily. We must efficiently process and store this data to make informed business decisions. Therefore, we plan to develop a data pipeline to extract, transform, and load data from three CSV files and store it in a Postgres database. We require a skilled data engineer who can use the Airflow tool to develop the pipeline to achieve this.

Problem Statement

The main challenge is that the data generated is in a raw format, and we need to process it efficiently to make it usable for analysis. This requires us to develop a data pipeline that can extract, transform and load the data from multiple CSV files into a single database, which can be used for further analysis.

Guidelines

The data pipeline should be developed using Airflow, an open-source tool for creating and managing data pipelines. The following steps should be followed to develop the data pipeline:

- The data engineer should start by creating a DAG (Directed Acyclic Graph) that defines the workflow of the data pipeline.
- The DAG should include tasks that extract data from the three CSV files.
- After extraction, the data should be transformed using Python libraries to match the required format.
- Finally, the transformed data should be loaded into a Postgres database.
- The data pipeline should be scheduled to run at a specific time daily using the Airflow scheduler.
- We can use the shared file (*mtnrwanda-dag.py*) as a starting point.

Sample CSV Files

The following are sample CSV files that will be used in the data pipeline:

- customer_data.csv
- order_data.csv
- payment_data.csv

All files for this project can be downloaded from here (link).

Deliverables

We will be expected to deliver a GitHub repository with the following:

- Airflow DAG file for the data pipeline.
- Documentation of the pipeline.
    - Highlight at least 3 best practices used during the implementation.
    - Recommendations for deployment and running the pipeline in a cloud-based provider.