**Steps to Run the Project**

**Prerequisites**

1. **Install Visual Studio Code.**
2. **Set Up Kaggle API Token:**
   * Navigate to your Kaggle account.
   * Create and download your API token (kaggle.json).
3. **Create Snowflake Account:**
   * Create SQL worksheets as shown in the snowflake\_processing folder and run them.
4. **Create AWS Account and S3 Bucket:**
   * Create an S3 bucket.
   * Create four folders within the S3 bucket to store normalized data.

**Setting Up the EC2 Instance**

1. **Create an Ubuntu EC2 Instance:**
   * Choose t2.small as the instance type.
   * Generate SSH access keys (.pem file).
2. **Configure Security Group Inbound Rules:**
   * Open port 22 for SSH access.
   * Open port 8080 for Airflow webserver access (Custom TCP anywhere).
3. **Connect to EC2 Instance:**
   * Use SSH to connect from your local machine to the EC2 instance:

ssh -i "path-to-your-pem-file" ec2-username@public-ip

**Installing Dependencies on EC2**

1. **Update Packages:**

sudo apt update

1. **Install Python and Virtual Environment:**

sudo apt install python3-pip

sudo apt install python3.10-venv

python3 -m venv airflow\_venv

1. **Activate the Python Environment:**

source airflow\_venv/bin/activate

1. **Install Required Python Packages:**

sudo pip install pandas

sudo pip install s3fs

pip install apache-airflow-providers-amazon

pip install python-dotenv

pip install kaggle

sudo pip install apache-airflow

**Setting Up Airflow**

1. **Transfer Kaggle API Token:**
   * From a new command prompt window, copy the Kaggle API token file to the EC2 instance:

scp -i "path-to-your-pem-file" "path-to-your-kaggle-json-api-tokenfile" ec2-username@your-ec2-public-ip:/home/ec2-username/.kaggle/kaggle.json

1. **Transfer DAG Script:**
   * Similarly, copy the DAG script (SNOWFLAKE\_ETL.py) to the Airflow DAGs folder on the EC2 instance:

scp -i "path-to-your-pem-file" "path-to-your-DAG-script" ec2-username@public-ip:/home/ec2-username/airflow/dags

1. **Start Airflow Components:**
   * Navigate back to your other command prompt window and start Airflow:

airflow standalone

**Accessing and Configuring Airflow**

1. **Access Airflow Webserver:**
   * Open a browser and navigate to <your-EC2-publicDNS>:8080.
2. **Set Up Connections:**
   * In the Airflow webserver, navigate to **Connections** and set up the following:
     + **HTTP API**
       - Connection Id: http-api\_kaggle
       - Connection Type: HTTP
       - Host: [Insurance Claims Dataset (kaggle.com)](https://www.kaggle.com/datasets/litvinenko630/insurance-claims)
     + **AWS Default**
       - Connection Id: aws-default
       - Connection Type: Amazon Web Services
       - AWS Access Key ID: <Your AWS access key>
       - AWS Secret Access Key: <Your AWS secret key>

**Editing and Triggering the DAG**

1. **Edit DAG Script:**
   * Open a new command prompt window and SSH into the EC2 instance:

ssh -i "path-to-your-pem-file" ec2-username@public-ip

* + Navigate to the DAGs folder and edit the SNOWFLAKE\_ETL.py script to understand and modify the load\_random\_sample() function as needed:

cd airflow/dags

nano SNOWFLAKE\_ETL.py

1. **Save and Refresh Airflow:**
   * Save the script changes and refresh the Airflow webserver.
   * Schedule and trigger the DAG according to your requirements.