**Scenario**

You are a data engineer at a data analytics consulting company. You have been assigned to a project that aims to de-congest the national highways by analyzing the road traffic data from different toll plazas. Each highway is operated by a different toll operator with a different IT setup that uses different file formats. Your job is to collect data available in different formats and consolidate it into a single file.

**Objectives**

In this assignment you will author an Apache Airflow DAG that will:

Extract data from a csv file

Extract data from a tsv file

Extract data from a fixed width file

Transform the data

Load the transformed data into the staging area

**Note - Screenshots**

Throughout this lab you will be prompted to take screenshots and save them on your own device. These screenshots will need to be uploaded for peer review in the next section of the course. You can use various free screengrabbing tools or your operating system's shortcut keys (Alt + PrintScreen in Windows, for example) to capture the required screenshots.

Exercise 1 - Prepare the lab environment

Before you start the assignment:

**Start Apache Airflow.**

Download the dataset from the source to the destination mentioned below.

Note: While downloading the file in the terminal use the sudo command before the command used to download the file.

Source : https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Final%20Assignment/tolldata.tgz

Destination : /home/project/airflow/dags/finalassignment

Create a directory structure for staging area as follows

/home/project/airflow/dags/finalassignment/staging.

Firstly enter the command cd airflow/dags in the terminal to change the directory to the /home/project/airflow/dags .

cd airflow/dags

Step 2: Next enter the below given commands to create the directories finalassignment and staging

sudo mkdir finalassignment

cd finalassignment

sudo mkdir staging

cd staging

A screenshot of a computer

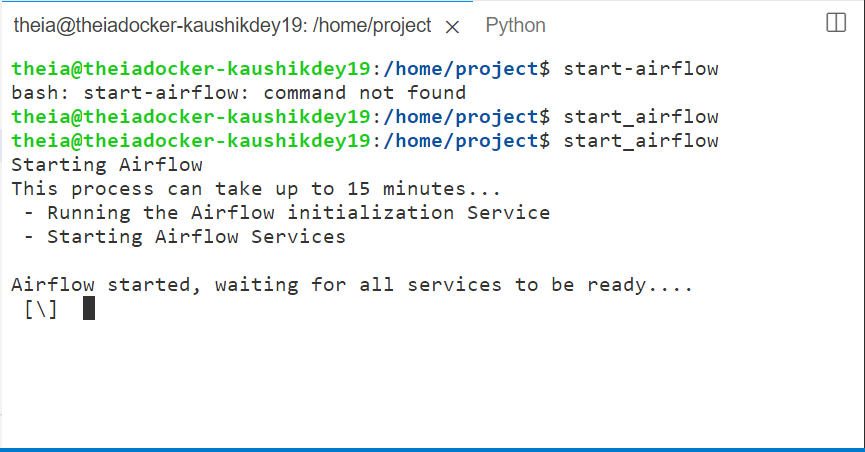
Description automatically generated

Source : sudo wget <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Final%20Assignment/tolldata.tgz>

Text

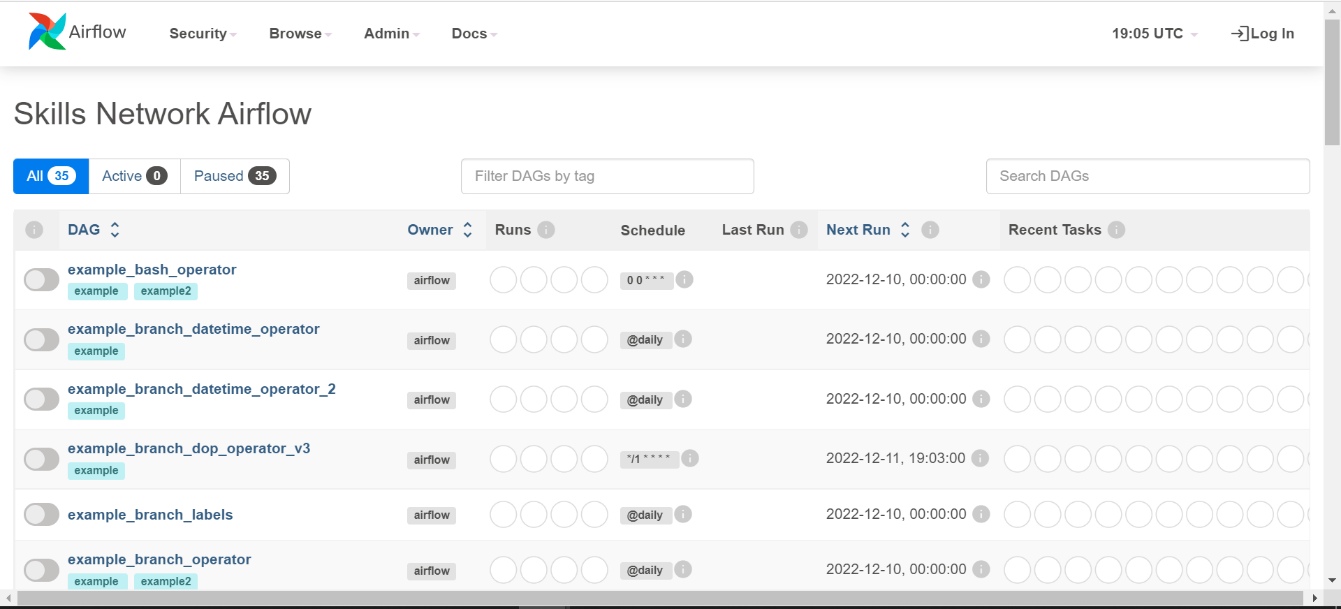
Description automatically generated

First, we have to up and running our Apache Airflow Server.



AirFlow Server Link: aa

<https://kaushikdey19-8080.theiadocker-2-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai/home>



# define tasks

# Task 1.3- Create a task named unzip\_data

# saved as unzip\_data.jpg

    # ---------------------------

unzip\_data = BashOperator(task\_id='unzip\_data',

                          bash\_command='tar -zxvf /home/project/airflow/dags/finalassignment/tolldata.tgz -C /home/project/airflow/dags/finalassignment'

                          , dag=dag)

# Task 1.4- Create a task to extract data from csv file

    # file saved as extract\_data\_from\_csv.jpg

extract\_data\_from\_csv = BashOperator(task\_id='extract\_data\_from\_csv',

        bash\_command='cut -d"," -f 1-4 /home/project/airflow/dags/finalassignment/vehicle-data.csv > /home/project/airflow/dags/finalassignment/staging/csv\_data.csv'

        , dag=dag)

# Task 1.5 - Create a task to extract data from tsv file.

extract\_data\_from\_tsv = BashOperator(task\_id='extract\_data\_from\_tsv',

        bash\_command='cut -f 5-7 /home/project/airflow/dags/finalassignment/tollplaza-data.tsv > /home/project/airflow/dags/finalassignment/staging/tsv\_data.csv --output-delimiter=","'

        , dag=dag)

# Task 1.6 - Create a task to extract data from fixed width file

extract\_data\_from\_fixed\_width = \

    BashOperator(task\_id='extract\_data\_from\_fixed\_width',

                 bash\_command='cut -c 59-61,63-68 /home/project/airflow/dags/finalassignment/payment-data.txt > /home/project/airflow/dags/finalassignment/staging/fixed\_width\_data.csv --output-delimiter=","'

                 , dag=dag)

# Task 1.7 - Create a task to consolidate data extracted from previous tasks

consolidate\_data = BashOperator(task\_id='consolidate\_data',

                                bash\_command='paste /home/project/airflow/dags/finalassignment/staging/csv\_data.csv /home/project/airflow/dags/finalassignment/staging/tsv\_data.csv /home/project/airflow/dags/finalassignment/staging/fixed\_width\_data.csv > /home/project/airflow/dags/finalassignment/staging/extracted\_data.csv'

                                , dag=dag)

# Task 1.8 - Transform and load the data

transform\_data = BashOperator(task\_id='transform\_data',

                              bash\_command='awk \'BEGIN{FS=","; OFS=","} {print $1,$2,$3,toupper($4),$5,$6,$7,$8,$9}\' /home/project/airflow/dags/finalassignment/staging/extracted\_data.csv > /home/project/airflow/dags/finalassignment/staging/transformed\_data.csv'

                              , dag=dag)

# task pipeline

# Task 1.9 - Define the task pipeline

unzip\_data >> extract\_data\_from\_csv >> extract\_data\_from\_tsv \

    >> extract\_data\_from\_fixed\_width >> consolidate\_data \

    >> transform\_data

theia@theiadocker-kaushikdey19:/home/project/airflow/dags$ airflow dags list

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Final%20Assignment/ETL_Peer_Review_Assignment.md.html>

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Apache%20Airflow/Build%20a%20DAG%20using%20Airflow/Build%20a%20DAG%20using%20Airflow.md.html>

<https://kaushikdey19-8080.theiadocker-2-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai/dags/ETL_toll_data/graph?root=>

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Apache%20Airflow/Monitoring%20a%20DAG/Hands-on_Lab-_Monitoring_a_DAG.md.html>