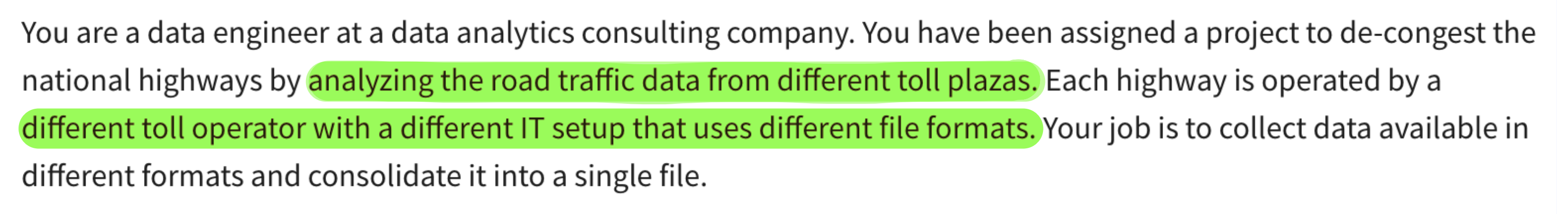
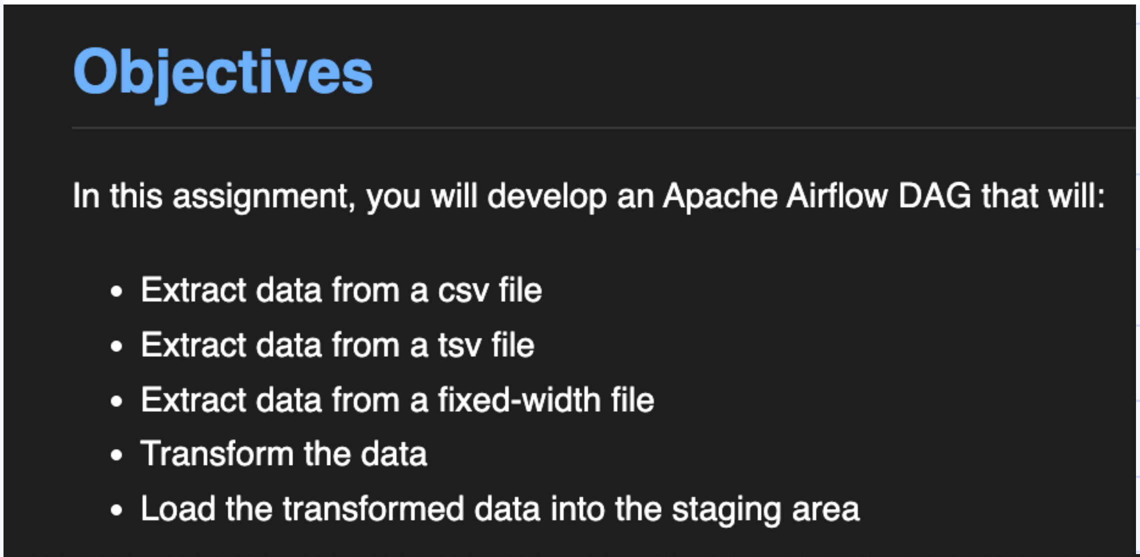


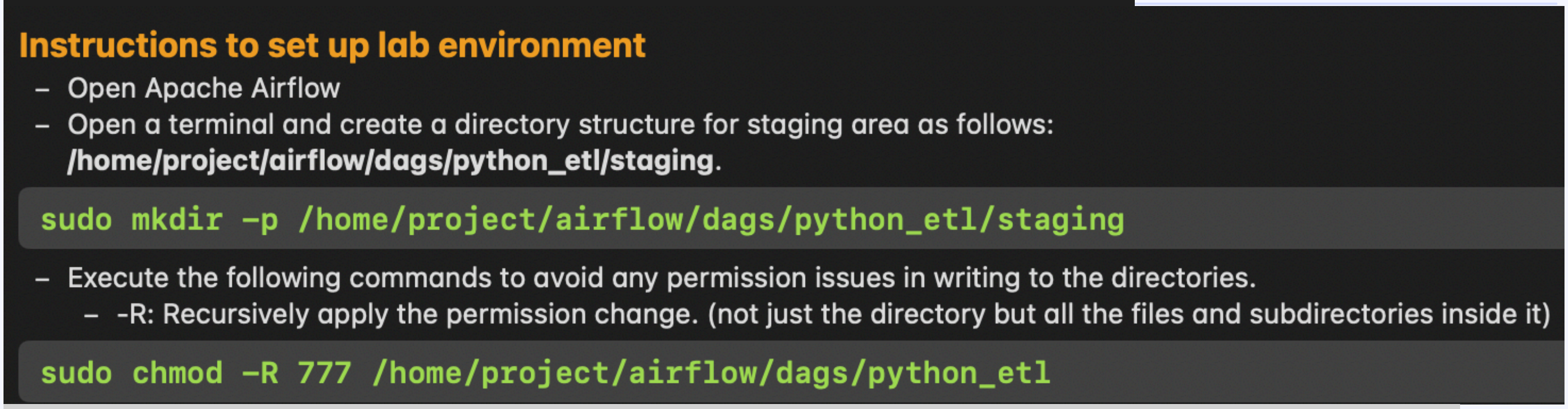
You are a data engineer at a data analytics consulting company. You have been assigned a project to de-congest the national highways by analyzing the road traffic data from different toll plazas. Each highway is operated by a different toll operator with a different IT setup that uses different file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



Objectives

In this assignment, you will develop 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



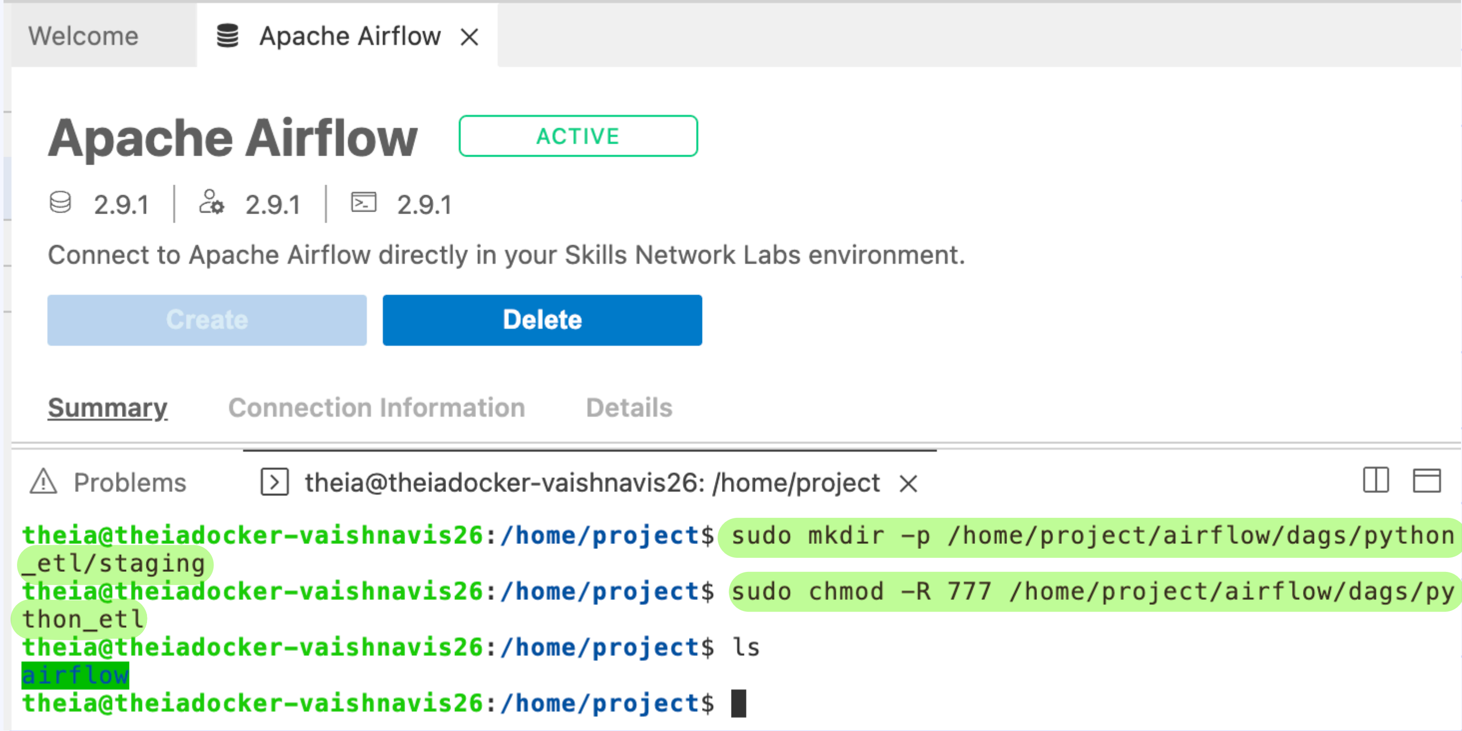
Instructions to set up lab environment

- Open Apache Airflow
- Open a terminal and create a directory structure for staging area as follows: **/home/project/airflow/dags/python_etl/staging**.

```
sudo mkdir -p /home/project/airflow/dags/python_etl/staging
```

- Execute the following commands to avoid any permission issues in writing to the directories.
  - -R: Recursively apply the permission change. (not just the directory but all the files and subdirectories inside it)

```
sudo chmod -R 777 /home/project/airflow/dags/python_etl
```



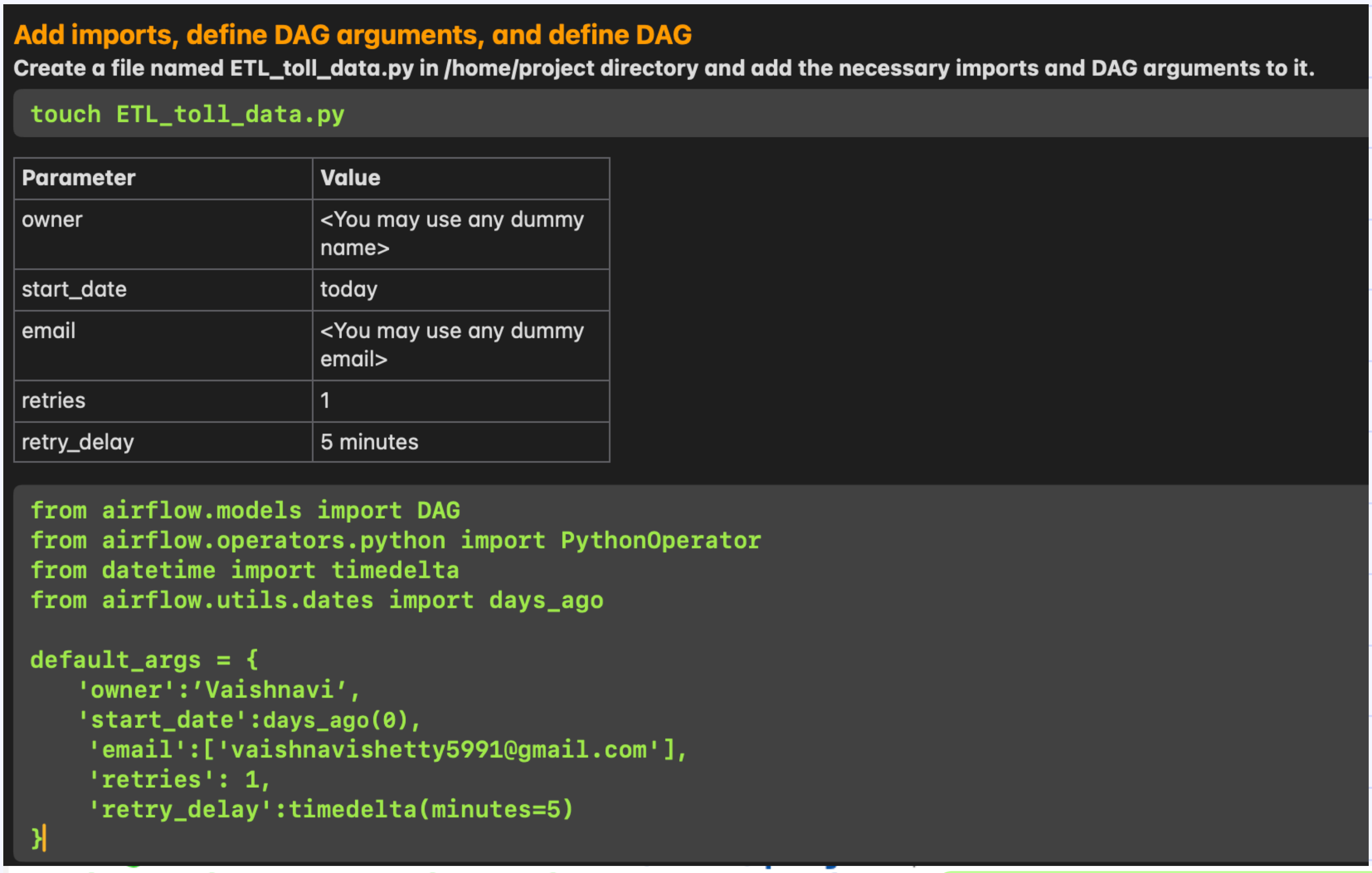
Apache Airflow

ACTIVE

2.9.1 | 2.9.1 | 2.9.1

Connect to Apache Airflow directly in your Skills Network Labs environment.

```
theia@theiadocker-vaishnavis26:/home/project$ sudo mkdir -p /home/project/airflow/dags/python
_etl/staging
theia@theiadocker-vaishnavis26:/home/project$ sudo chmod -R 777 /home/project/airflow/dags/py
thon_etl
theia@theiadocker-vaishnavis26:/home/project$ ls
airflow
theia@theiadocker-vaishnavis26:/home/project$ 
```



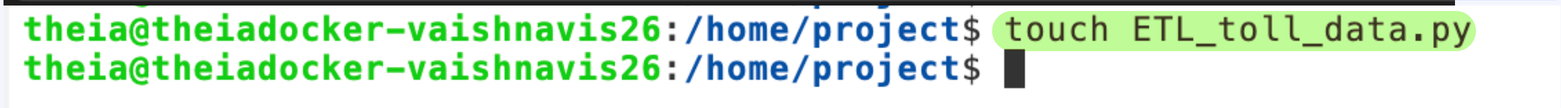
Add imports, define DAG arguments, and define DAG

Create a file named ETL_toll_data.py in /home/project directory and add the necessary imports and DAG arguments to it.

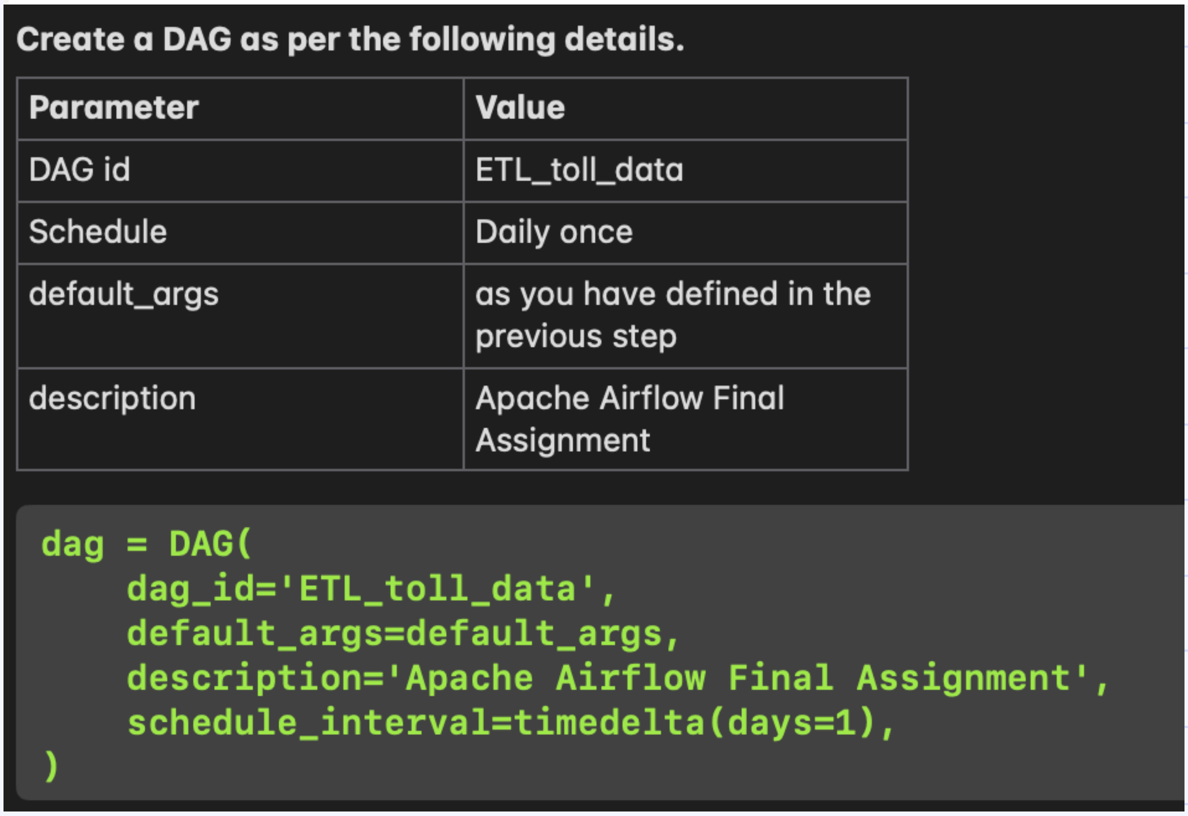
```
touch ETL_toll_data.py
```

Parameter	Value
owner	<You may use any dummy name>
start_date	today
email	<You may use any dummy email>
retries	1
retry_delay	5 minutes

```
from airflow.models import DAG
from airflow.operators.python import PythonOperator
from datetime import timedelta
from airflow.utils.dates import days_ago

default_args = {
    'owner':’Vaishnavi’,
    'start_date':days_ago(0),
     'email':['vaishnavishetty5991@gmail.com'],
     'retries': 1,
     'retry_delay':timedelta(minutes=5)
}
```



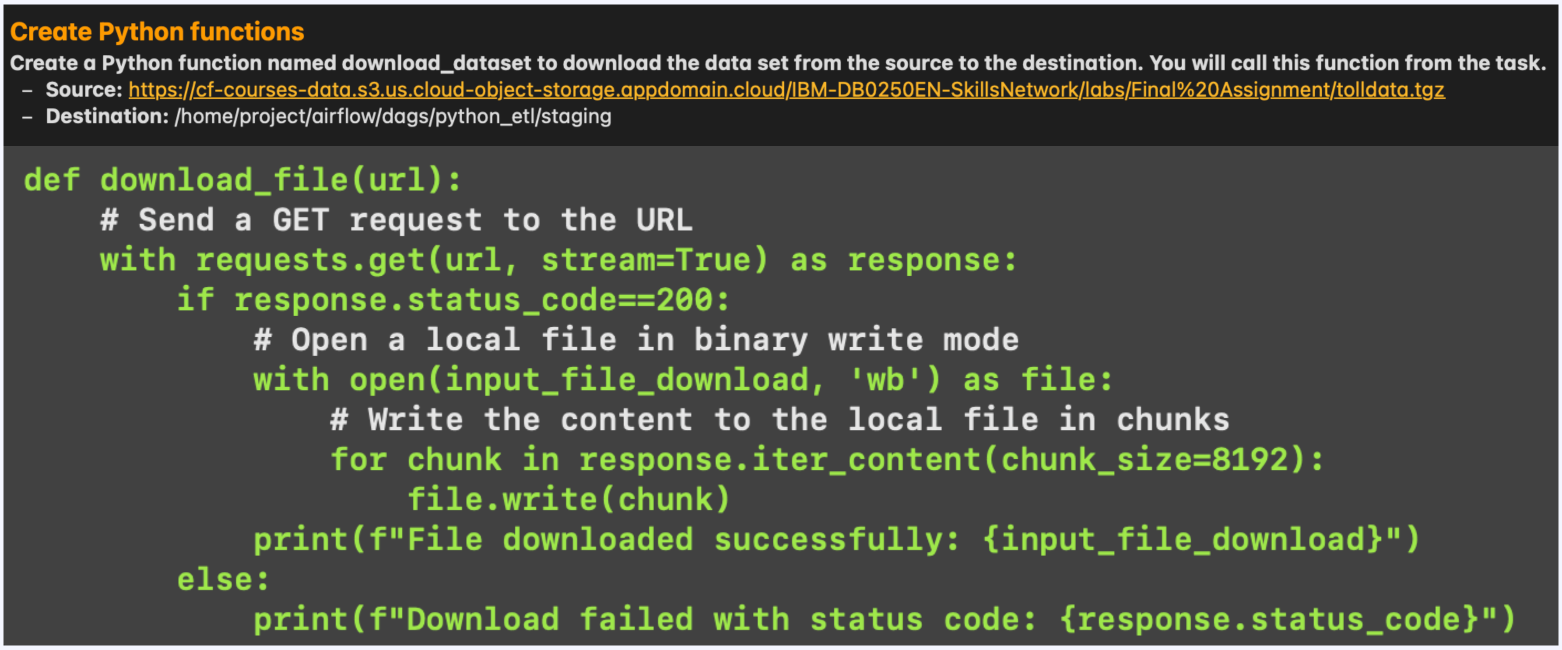
```
theia@theiadocker-vaishnavis26:/home/project$ touch ETL_toll_data.py
theia@theiadocker-vaishnavis26:/home/project$ 
```



Create a DAG as per the following details.

Parameter	Value
DAG id	ETL_toll_data
Schedule	Daily once
default_args	as you have defined in the previous step
description	Apache Airflow Final Assignment

```
dag = DAG(
    dag_id='ETL_toll_data',
    default_args=default_args,
    description='Apache Airflow Final Assignment',
    schedule_interval=timedelta(days=1),
)
```

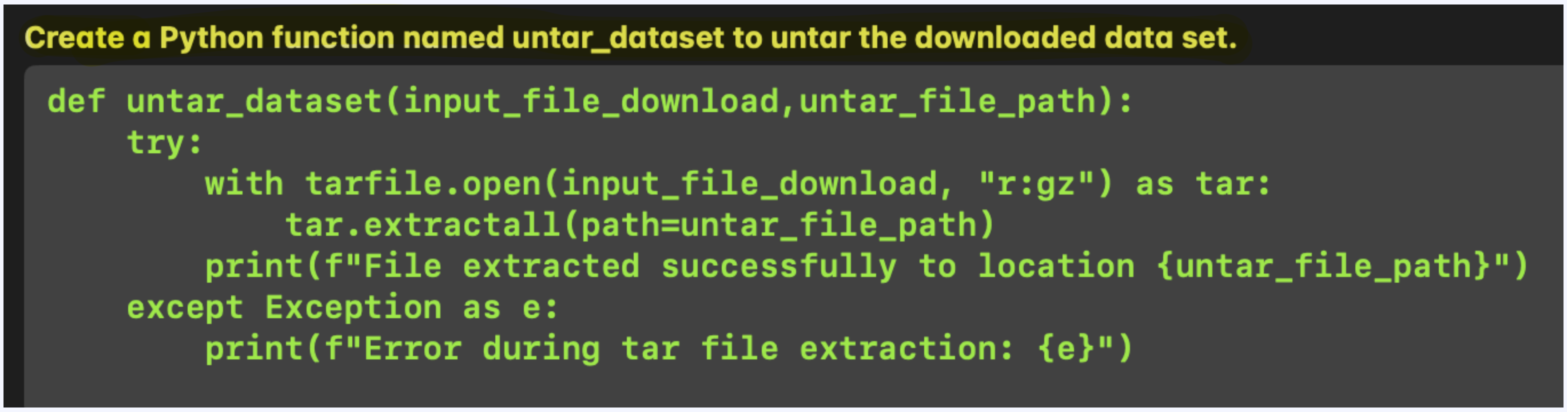


Create Python functions

Create a Python function named download_dataset to download the data set from the source to the destination. You will call this function from the task.

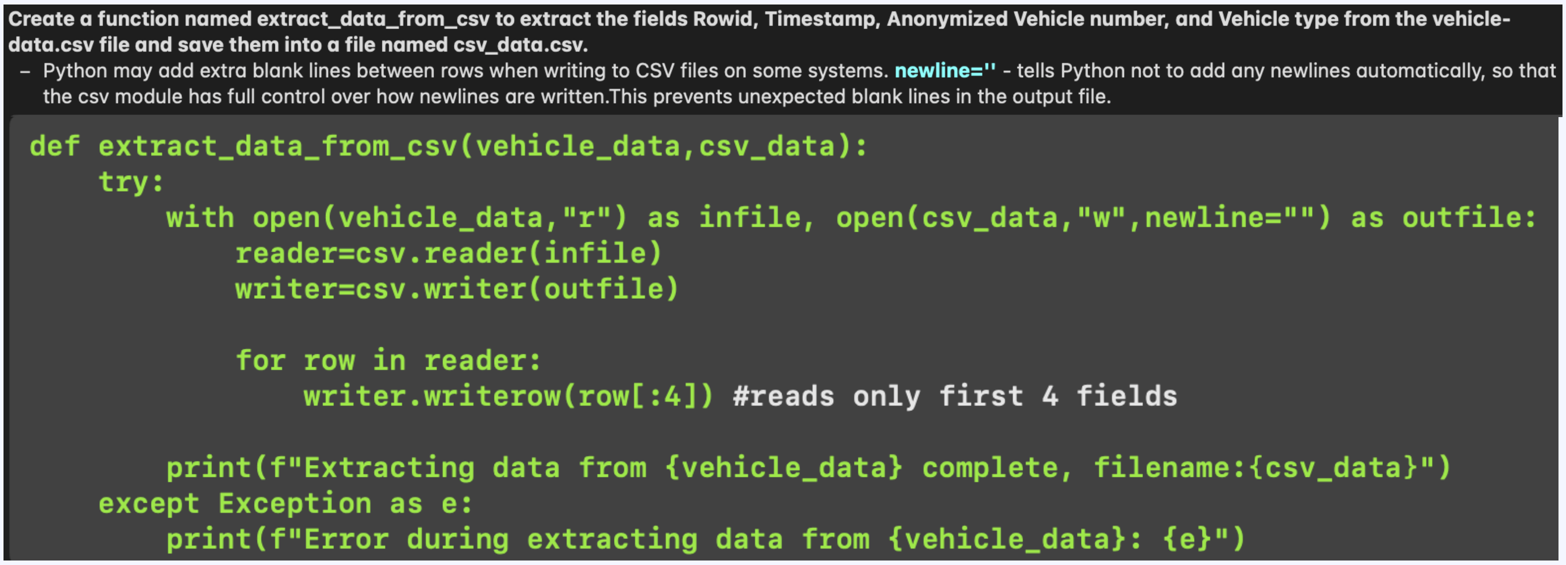
- **Source:** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Final%20Assignment/tolldata.tgz
- **Destination:** /home/project/airflow/dags/python_etl/staging

```
def download_file(url):
    # Send a GET request to the URL
    with requests.get(url, stream=True) as response:
        if response.status_code==200:
            # Open a local file in binary write mode
            with open(input_file_download, 'wb') as file:
                # Write the content to the local file in chunks
                for chunk in response.iter_content(chunk_size=8192):
                    file.write(chunk)
            print(f"File downloaded successfully: {input_file_download}")
        else:
            print(f"Download failed with status code: {response.status_code}")
```



Create a Python function named untar_dataset to untar the downloaded data set.

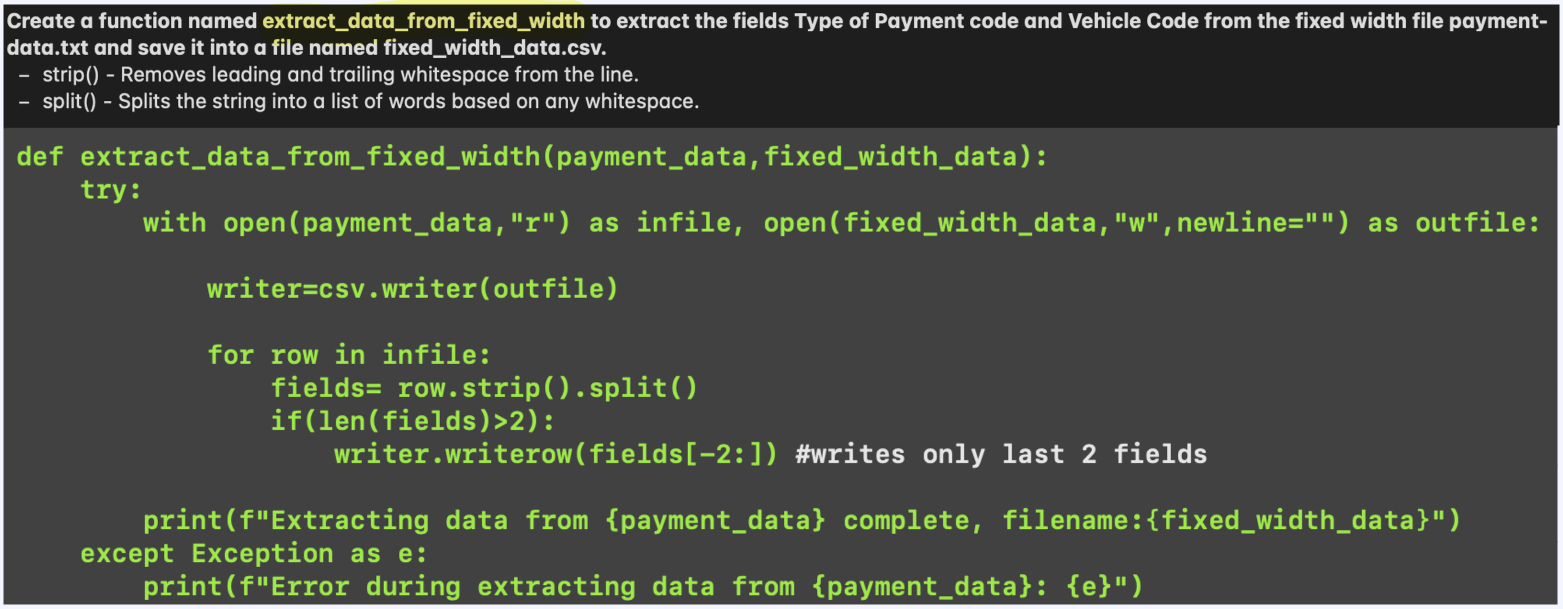
```
def untar_dataset(input_file_download,untar_file_path):
    try:
        with tarfile.open(input_file_download, "r:gz") as tar:
            tar.extractall(path=untar_file_path)
        print(f"File extracted successfully to location {untar_file_path}")
    except Exception as e:
        print(f"Error during tar file extraction: {e}")
```



Create a function named extract_data_from_csv to extract the fields Rowid, Timestamp, Anonymized Vehicle number, and Vehicle type from the vehicle-data.csv file and save them into a file named csv_data.csv.

- Python may add extra blank lines between rows when writing to CSV files on some systems. **newline=''** - tells Python not to add any newlines automatically, so that the csv module has full control over how newlines are written.This prevents unexpected blank lines in the output file.

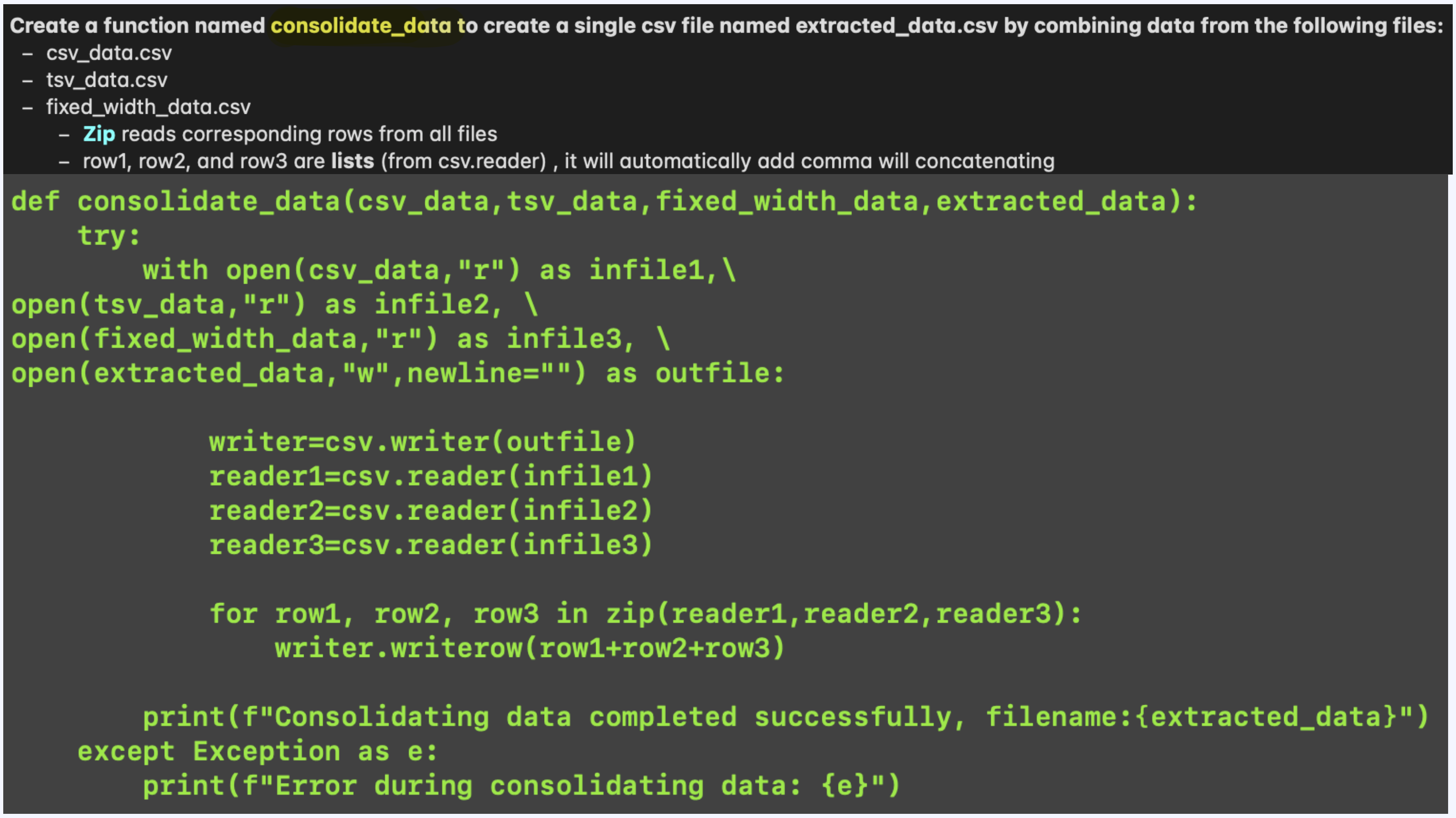
```
def extract_data_from_csv(vehicle_data,csv_data):
    try:
        with open(vehicle_data,"r") as infile, open(csv_data,"w",newline="") as outfile:
            reader=csv.reader(infile)
            writer=csv.writer(outfile)

            for row in reader:
                writer.writerow(row[:4]) #reads only first 4 fields

        print(f"Extracting data from {vehicle_data} complete, filename:{csv_data}")
    except Exception as e:
        print(f"Error during extracting data from {vehicle_data}: {e}")
```



Create a function named extract_data_from_fixed_width to extract the fields Type of Payment code and Vehicle Code from the fixed width file payment-data.txt and save it into a file named fixed_width_data.csv.

- strip() - Removes leading and trailing whitespace from the line.
- split() - Splits the string into a list of words based on any whitespace.

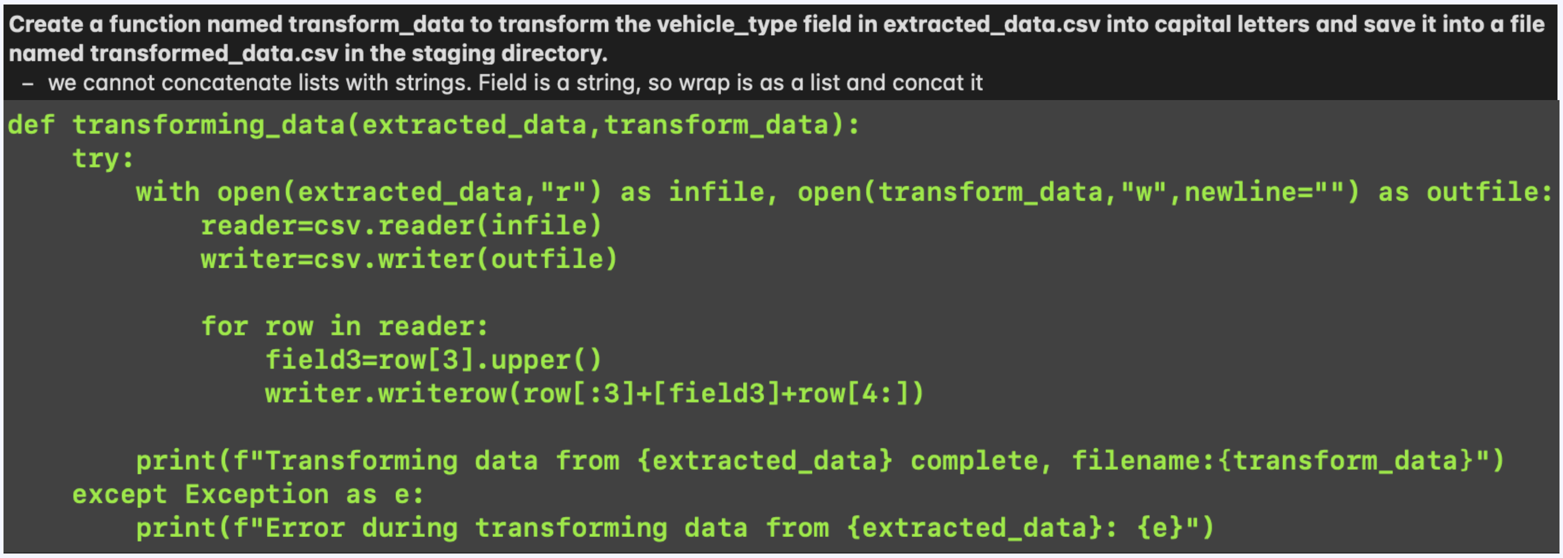
```
def extract_data_from_fixed_width(payment_data,fixed_width_data):
    try:
        with open(payment_data,"r") as infile, open(fixed_width_data,"w",newline="") as outfile:

            writer=csv.writer(outfile)

            for row in infile:
                fields= row.strip().split()
                if(len(fields)>2):
                    writer.writerow(fields[-2:]) #writes only last 2 fields

        print(f"Extracting data from {payment_data} complete, filename:{fixed_width_data}")
    except Exception as e:
        print(f"Error during extracting data from {payment_data}: {e}")
```



Create a function named consolidate_data to create a single csv file named extracted_data.csv by combining data from the following files:

- csv_data.csv
- tsv_data.csv
- fixed_width_data.csv
    - **Zip** reads corresponding rows from all files
    - row1, row2, and row3 are **lists** (from csv.reader) , it will automatically add comma will concatenating

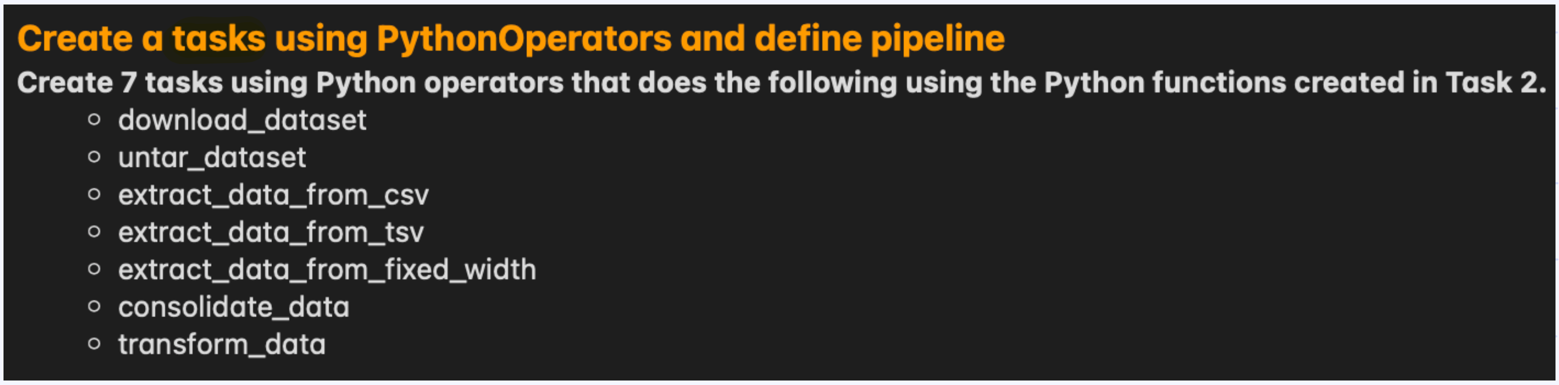
```
def consolidate_data(csv_data,tsv_data,fixed_width_data,extracted_data):
    try:
        with open(csv_data,"r") as infile1,\
open(tsv_data,"r") as infile2, \
open(fixed_width_data,"r") as infile3, \
open(extracted_data,"w",newline="") as outfile:

            writer=csv.writer(outfile)
            reader1=csv.reader(infile1)
            reader2=csv.reader(infile2)
            reader3=csv.reader(infile3)

            for row1, row2, row3 in zip(reader1,reader2,reader3):
                writer.writerow(row1+row2+row3)

        print(f"Consolidating data completed successfully, filename:{extracted_data}")
    except Exception as e:
        print(f"Error during consolidating data: {e}")
```



Create a function named transform_data to transform the vehicle_type field in extracted_data.csv into capital letters and save it into a file named transformed_data.csv in the staging directory.

- we cannot concatenate lists with strings. Field is a string, so wrap is as a list and concat it

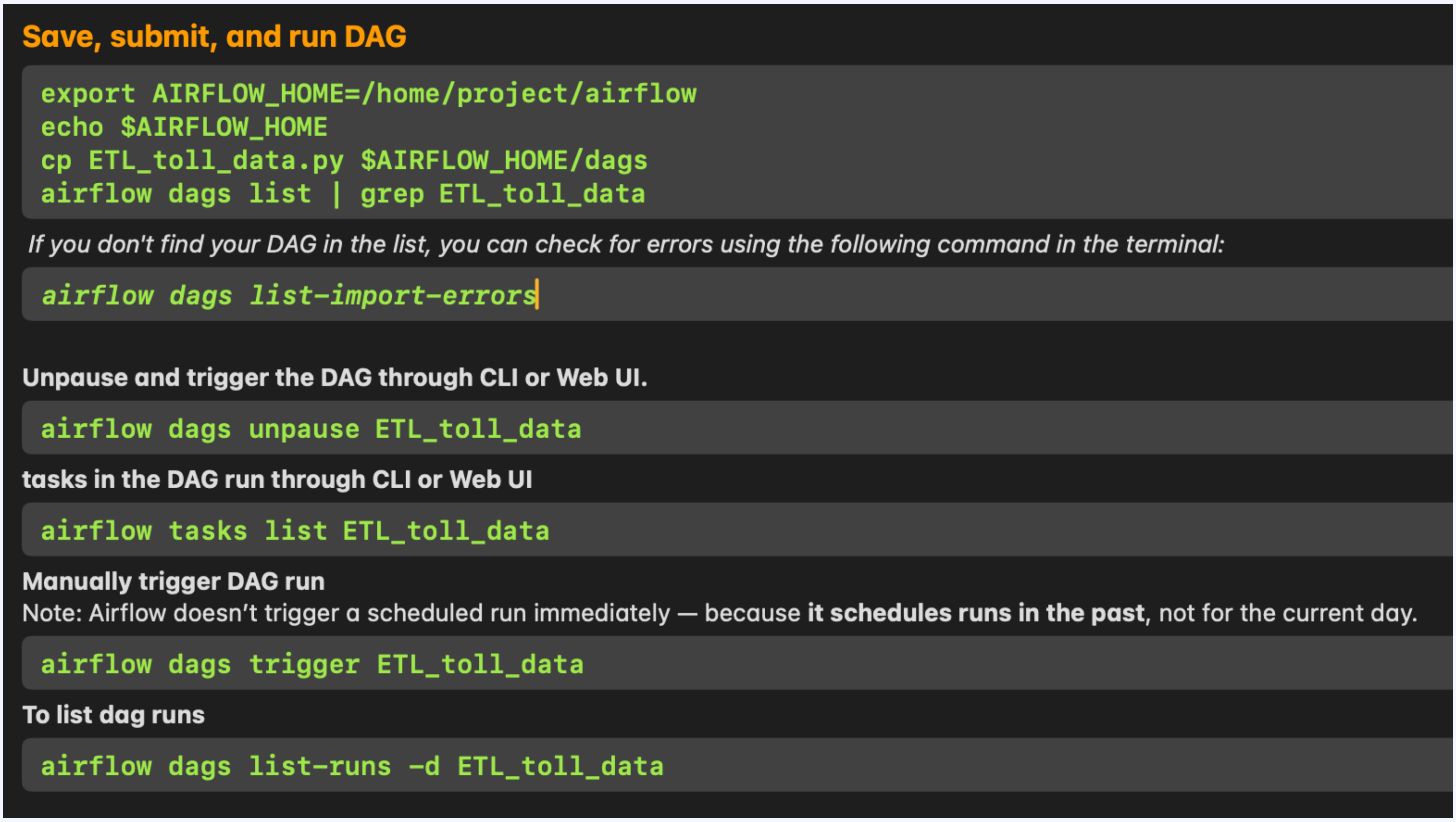
```
def transforming_data(extracted_data,transform_data):
    try:
        with open(extracted_data,"r") as infile, open(transform_data,"w",newline="") as outfile:
            reader=csv.reader(infile)
            writer=csv.writer(outfile)

            for row in reader:
                field3=row[3].upper()
                writer.writerow(row[:3]+[field3]+row[4:])

        print(f"Transforming data from {extracted_data} complete, filename:{transform_data}")
    except Exception as e:
        print(f"Error during transforming data from {extracted_data}: {e}")
```



Create a tasks using PythonOperators and define pipeline

Create 7 tasks using Python operators that does the following using the Python functions created in Task 2.

- download_dataset
- untar_dataset
- extract_data_from_csv
- extract_data_from_tsv
- extract_data_from_fixed_width
- consolidate_data
- transform_data



Save, submit, and run DAG

```
export AIRFLOW_HOME=/home/project/airflow
echo $AIRFLOW_HOME
cp ETL_toll_data.py $AIRFLOW_HOME/dags
airflow dags list | grep ETL_toll_data
```

If you don't find your DAG in the list, you can check for errors using the following command in the terminal:

```
airflow dags list-import-errors
```

Unpause and trigger the DAG through CLI or Web UI.

```
airflow dags unpause ETL_toll_data
```

tasks in the DAG run through CLI or Web UI

```
airflow tasks list ETL_toll_data
```

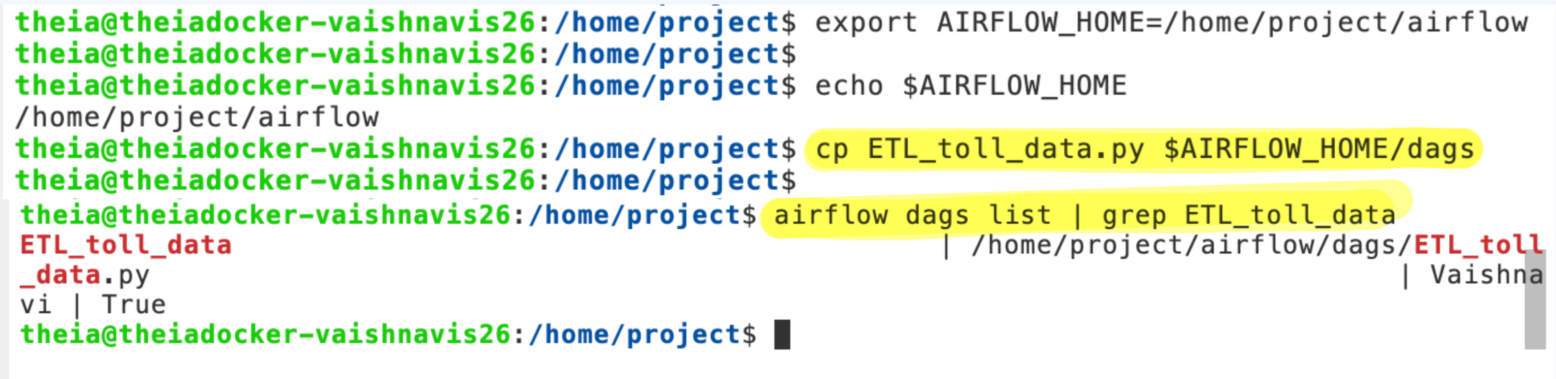
Manually trigger DAG run

Note: Airflow doesn't trigger a scheduled run immediately — because **it schedules runs in the past**, not for the current day.

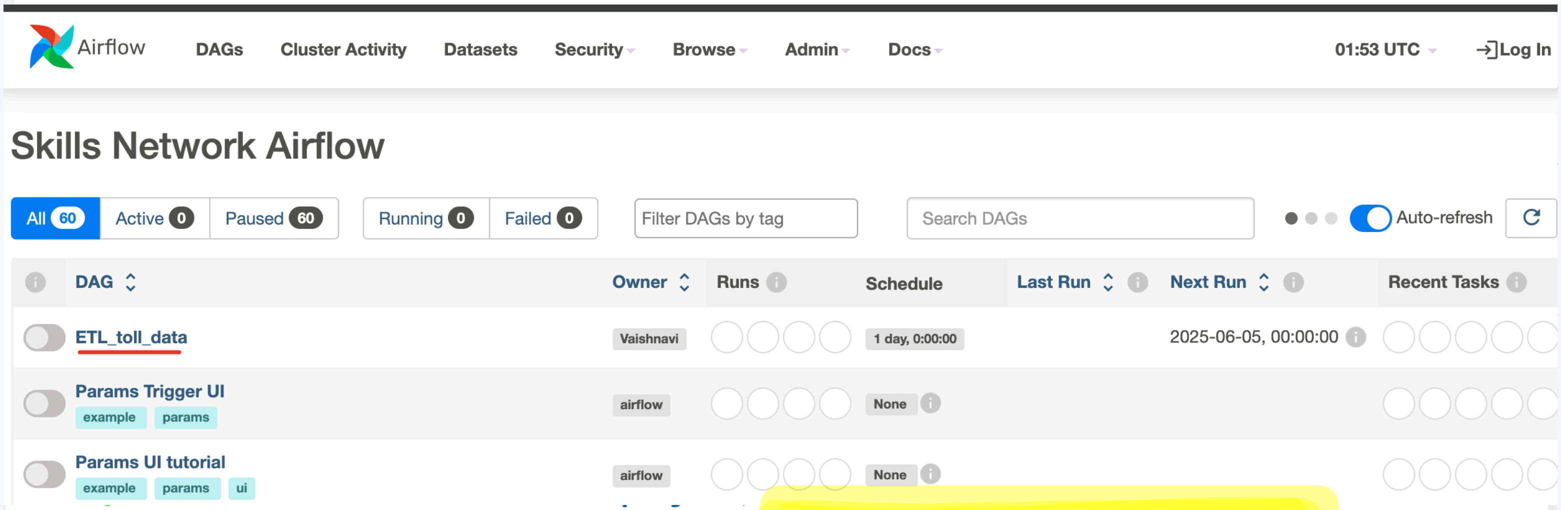
```
airflow dags trigger ETL_toll_data
```

To list dag runs

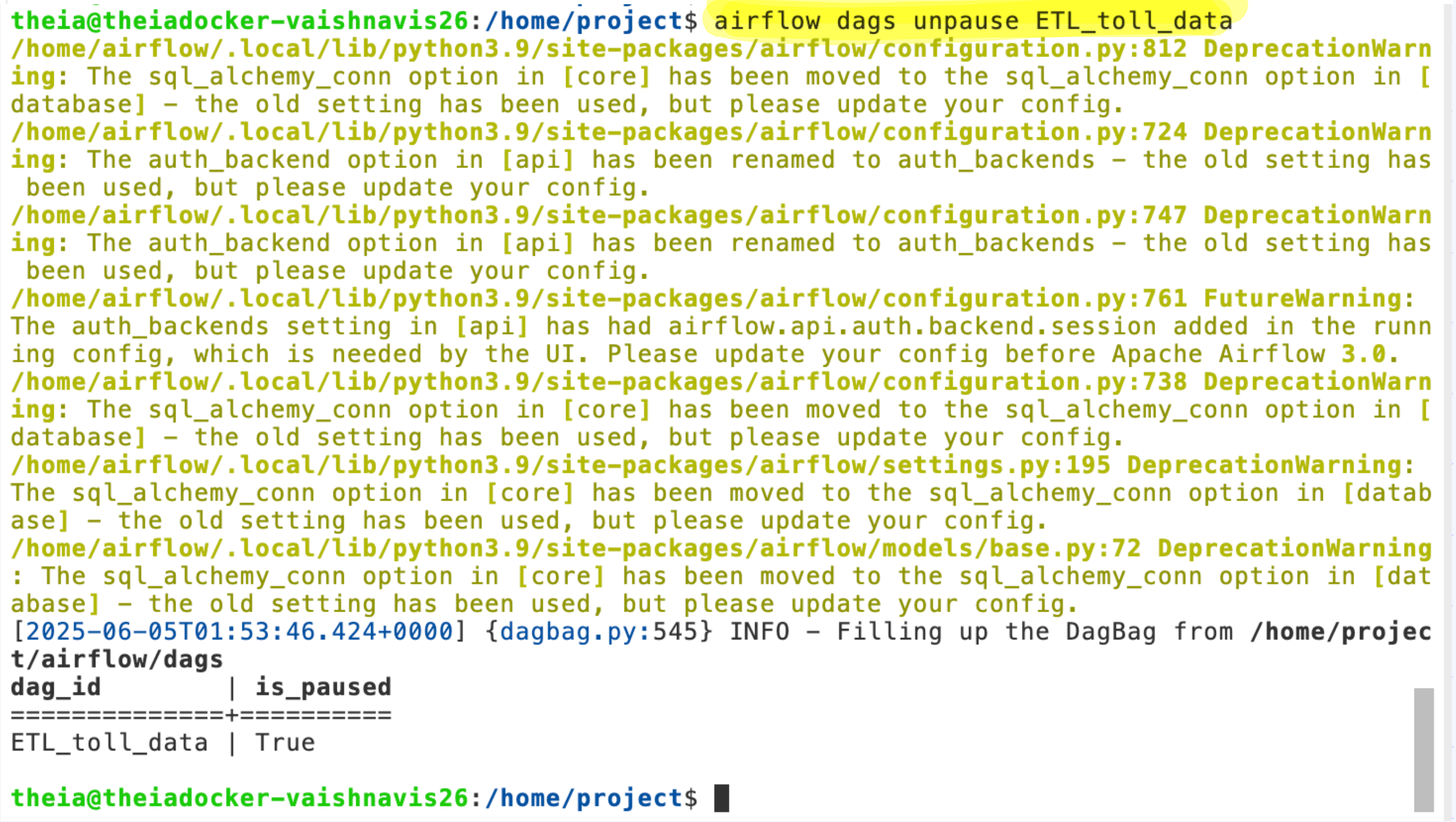
```
airflow dags list-runs -d ETL_toll_data
```



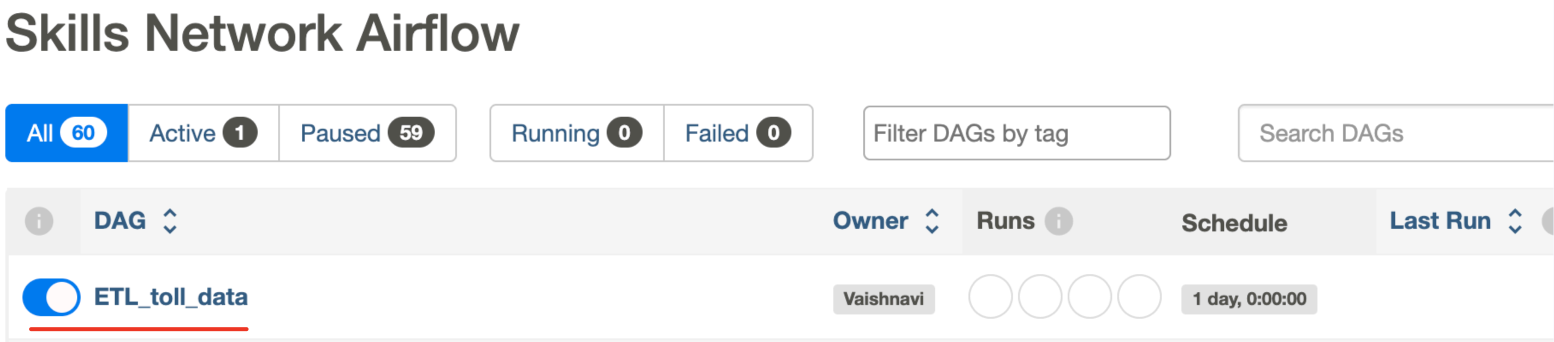
```
theia@theiadocker-vaishnavis26:/home/project$ export AIRFLOW_HOME=/home/project/airflow
theia@theiadocker-vaishnavis26:/home/project$
theia@theiadocker-vaishnavis26:/home/project$ echo $AIRFLOW_HOME
/home/project/airflow
theia@theiadocker-vaishnavis26:/home/project$ cp ETL_toll_data.py $AIRFLOW_HOME/dags
theia@theiadocker-vaishnavis26:/home/project$
theia@theiadocker-vaishnavis26:/home/project$ airflow dags list | grep ETL_toll_data
ETL_toll_data                                   | /home/project/airflow/dags/ETL_toll
_data.py                                                                   | Vaishna
vi | True
theia@theiadocker-vaishnavis26:/home/project$
```



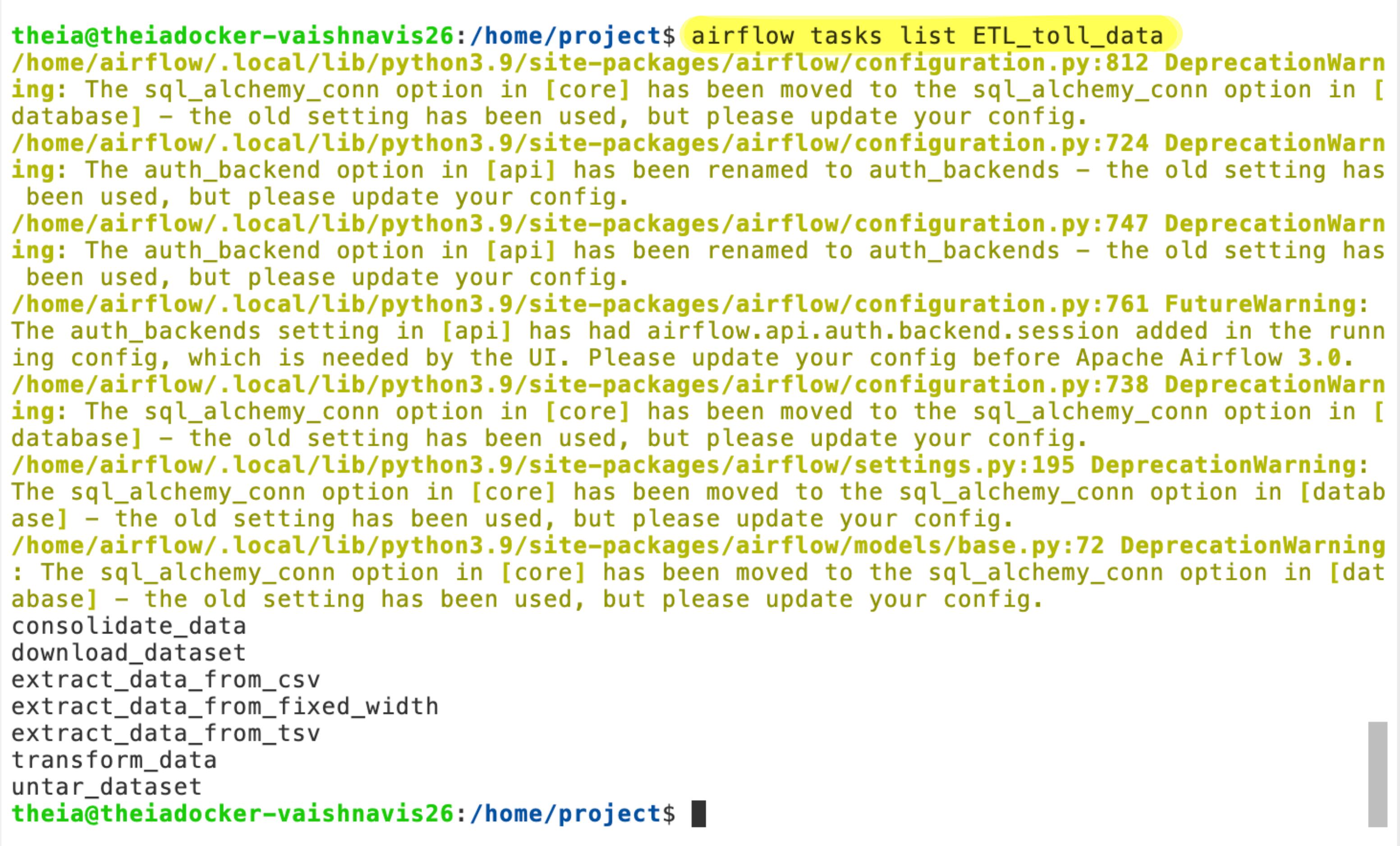
DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks
ETL_toll_data	Vaishnavi		1 day, 0:00:00		2025-06-05, 00:00:00	
Params Trigger UI example params	airflow		None			
Params UI tutorial example params ui	airflow		None			



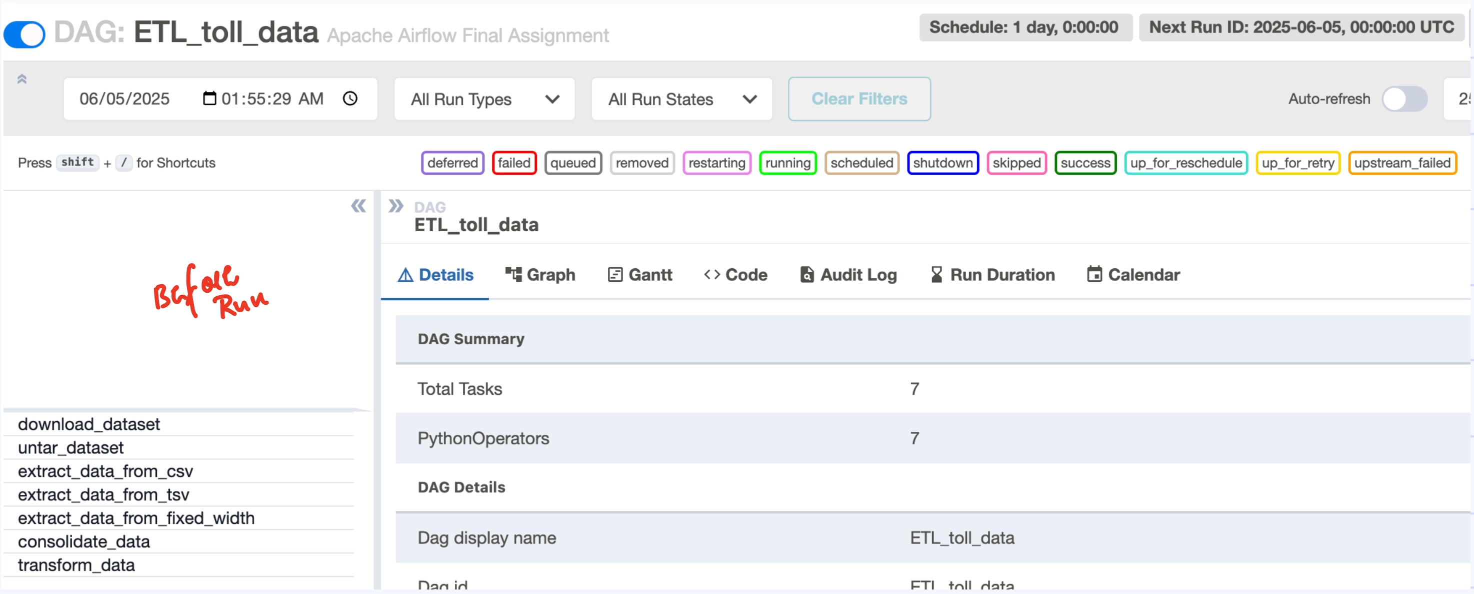
```
theia@theiadocker-vaishnavis26:/home/project$ airflow dags unpause ETL_toll_data
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:812 DeprecationWarn
ing: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [
database] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:724 DeprecationWarn
ing: The auth_backend option in [api] has been renamed to auth_backends - the old setting has
 been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:747 DeprecationWarn
ing: The auth_backend option in [api] has been renamed to auth_backends - the old setting has
 been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:761 FutureWarning:
The auth_backends setting in [api] has had airflow.api.auth.backend.session added in the runn
ing config, which is needed by the UI. Please update your config before Apache Airflow 3.0.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:738 DeprecationWarn
ing: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [
database] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/settings.py:195 DeprecationWarning:
The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [datab
ase] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/models/base.py:72 DeprecationWarning
: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [dat
abase] - the old setting has been used, but please update your config.
[2025-06-05T01:53:46.424+0000] {dagbag.py:545} INFO - Filling up the DagBag from /home/projec
t/airflow/dags
dag_id        | is_paused
==============+==========
ETL_toll_data | True

theia@theiadocker-vaishnavis26:/home/project$
```



DAG	Owner	Runs	Schedule	Last Run
ETL_toll_data	Vaishnavi	○○○○	1 day, 0:00:00	



```
theia@theiadocker-vaishnavis26:/home/project$ airflow tasks list ETL_toll_data
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:812 DeprecationWarn
ing: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [
database] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:724 DeprecationWarn
ing: The auth_backend option in [api] has been renamed to auth_backends - the old setting has
 been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:747 DeprecationWarn
ing: The auth_backend option in [api] has been renamed to auth_backends - the old setting has
 been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:761 FutureWarning:
The auth_backends setting in [api] has had airflow.api.auth.backend.session added in the runn
ing config, which is needed by the UI. Please update your config before Apache Airflow 3.0.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:738 DeprecationWarn
ing: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [
database] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/settings.py:195 DeprecationWarning:
The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [datab
ase] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/models/base.py:72 DeprecationWarning
: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [dat
abase] - the old setting has been used, but please update your config.
consolidate_data
download_dataset
extract_data_from_csv
extract_data_from_fixed_width
extract_data_from_tsv
transform_data
untar_dataset
theia@theiadocker-vaishnavis26:/home/project$ █
```



DAG: ETL_toll_data Apache Airflow Final Assignment

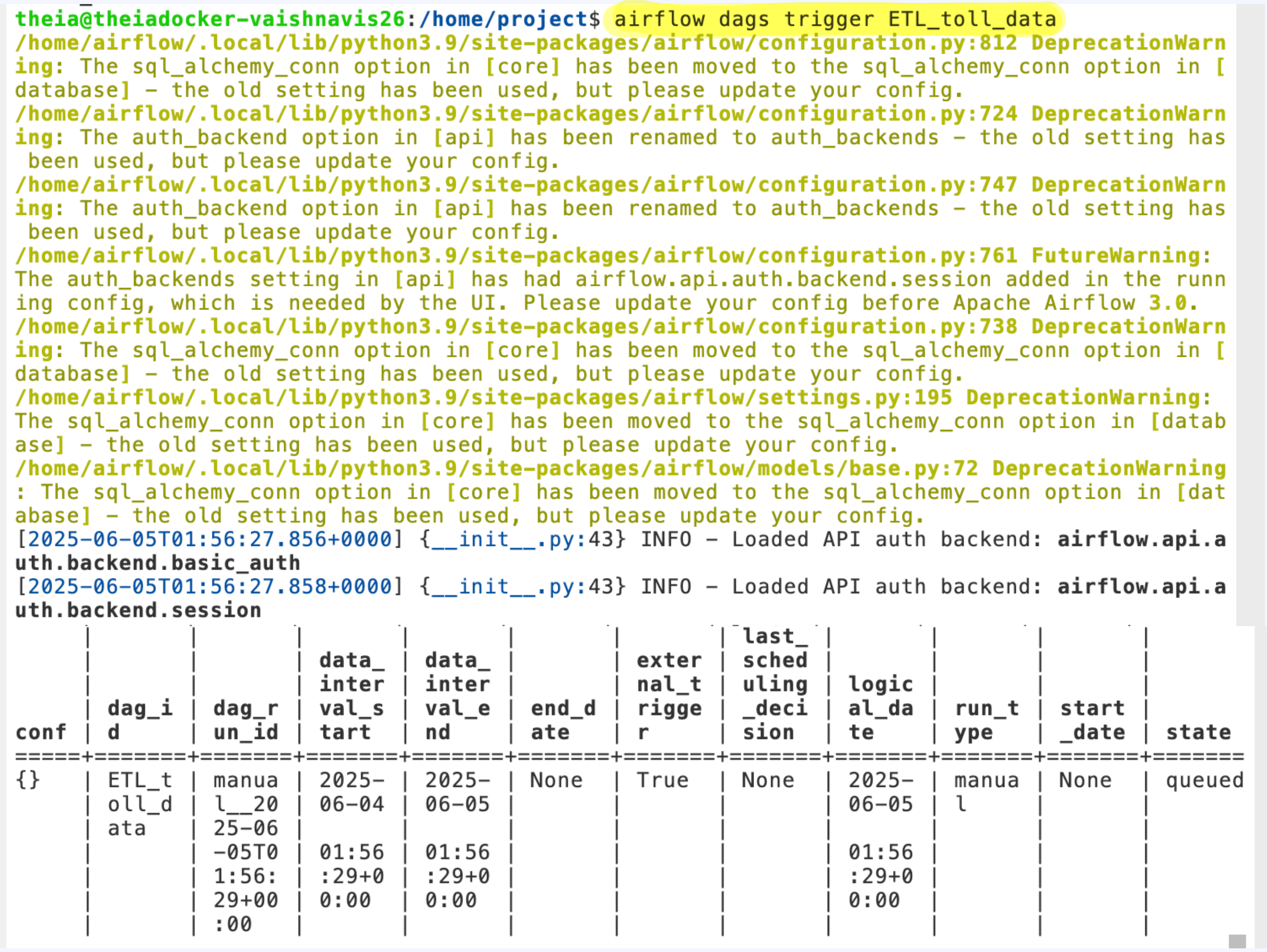
Schedule: 1 day, 0:00:00

Next Run ID: 2025-06-05, 00:00:00 UTC

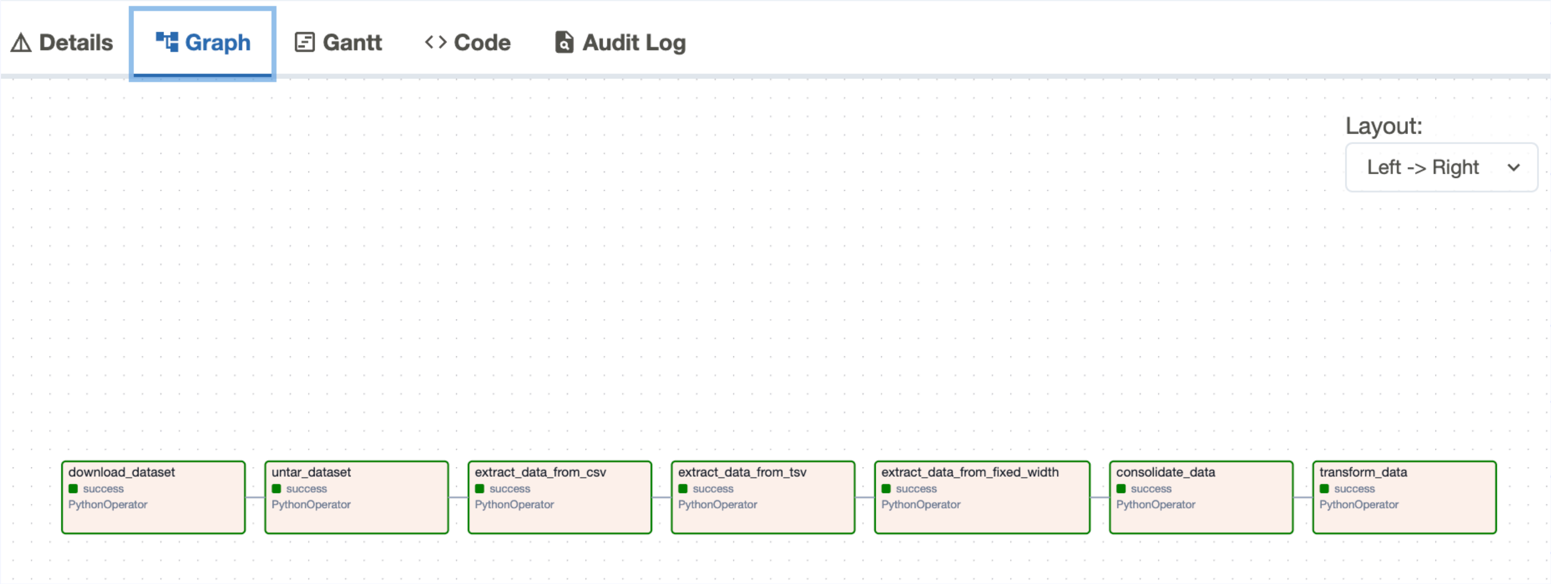
download_dataset
untar_dataset
extract_data_from_csv
extract_data_from_tsv
extract_data_from_fixed_width
consolidate_data
transform_data

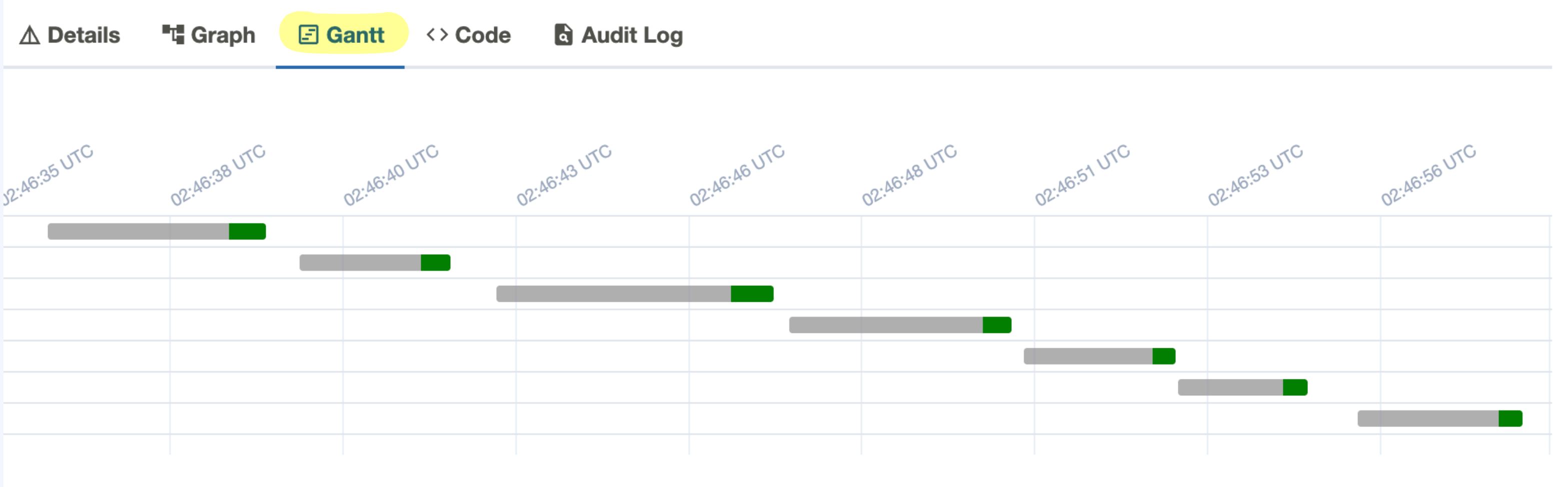
DAG
ETL_toll_data

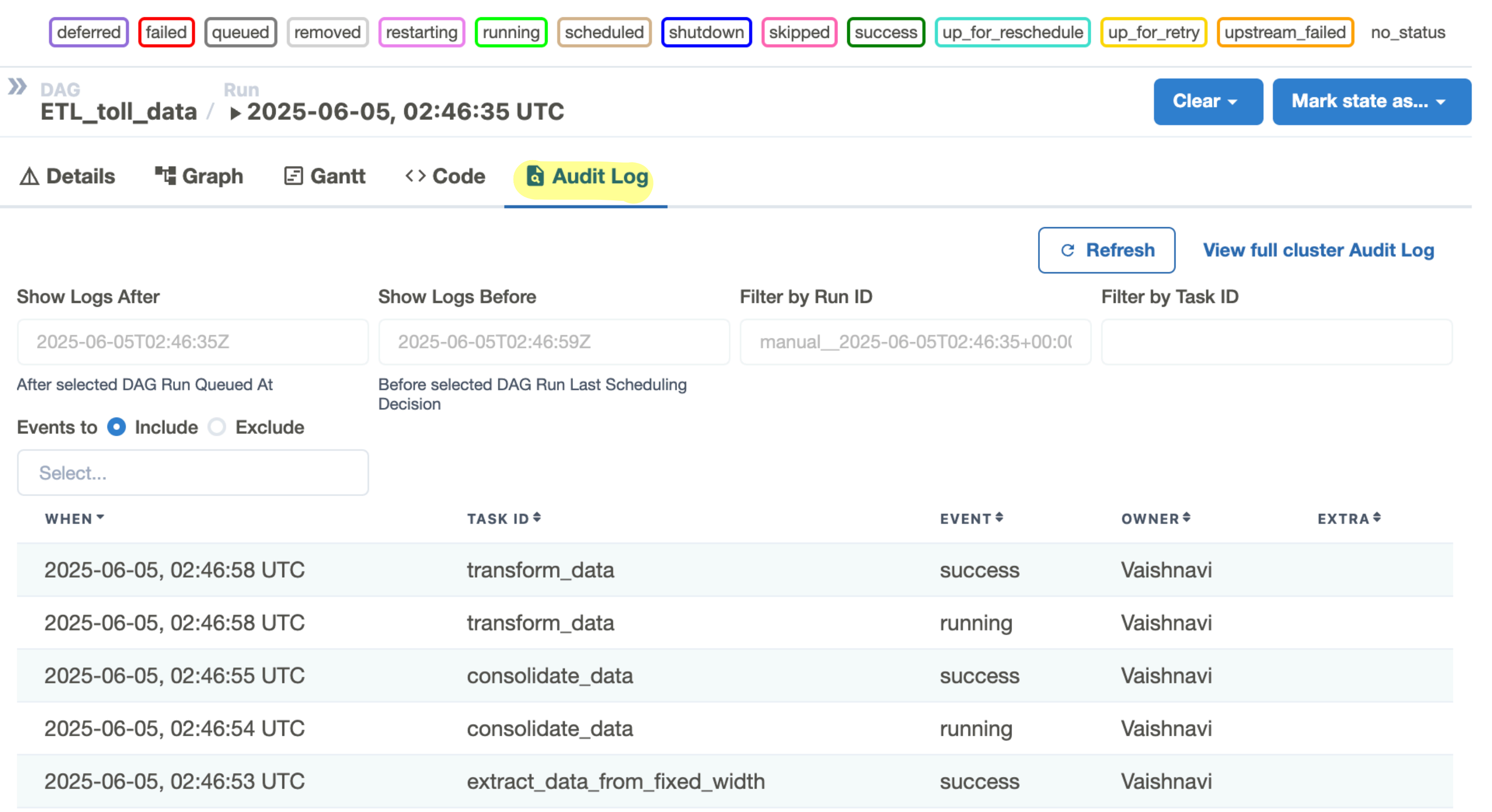
DAG Summary	
Total Tasks	7
PythonOperators	7
DAG Details	
Dag display name	ETL_toll_data
Dag id	ETL_toll_data



```
theia@theiadocker-vaishnavis26:/home/project$ airflow dags trigger ETL_toll_data
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:812 DeprecationWarn
ing: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [
database] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:724 DeprecationWarn
ing: The auth_backend option in [api] has been renamed to auth_backends - the old setting has
 been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:747 DeprecationWarn
ing: The auth_backend option in [api] has been renamed to auth_backends - the old setting has
 been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:761 FutureWarning:
The auth_backends setting in [api] has had airflow.api.auth.backend.session added in the runn
ing config, which is needed by the UI. Please update your config before Apache Airflow 3.0.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:738 DeprecationWarn
ing: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [
database] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/settings.py:195 DeprecationWarning:
The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [datab
ase] - the old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/models/base.py:72 DeprecationWarning
: The sql_alchemy_conn option in [core] has been moved to the sql_alchemy_conn option in [dat
abase] - the old setting has been used, but please update your config.
[2025-06-05T01:56:27.856+0000] {__init__.py:43} INFO - Loaded API auth backend: airflow.api.a
uth.backend.basic_auth
[2025-06-05T01:56:27.858+0000] {__init__.py:43} INFO - Loaded API auth backend: airflow.api.a
uth.backend.session
      |       |       |       |       |       |       | last_ |       |       |       |
      |       |       | data_ | data_ |       | exter | sched |       |       |       |
      |       |       | inter | inter |       | nal_t | uling | logic |       |       |
      | dag_i | dag_r | val_s | val_e | end_d | rigge | _deci | al_da | run_t | start |
conf  | d     | un_id | tart  | nd    | ate   | r     | sion  | te    | ype   | _date | state
======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======
{}    | ETL_t | manua | 2025- | 2025- | None  | True  | None  | 2025- | manua | None  | queued
      | oll_d | l__20 | 06-04 | 06-05 |       |       |       | 06-05 | l     |       |
      | ata   | 25-06 |       |       |       |       |       |       |       |       |
      |       | -05T0 | 01:56 | 01:56 |       |       |       | 01:56 |       |       |
      |       | 1:56: | :29+0 | :29+0 |       |       |       | :29+0 |       |       |
      |       | 29+00 | 0:00  | 0:00  |       |       |       | 0:00  |       |       |
      |       | :00   |       |       |       |       |       |       |       |       |
```







DAG ETL_toll_data / Run ▸ 2025-06-05, 02:46:35 UTC

Show Logs After: 2025-06-05T02:46:35Z
After selected DAG Run Queued At

Show Logs Before: 2025-06-05T02:46:59Z
Before selected DAG Run Last Scheduling Decision

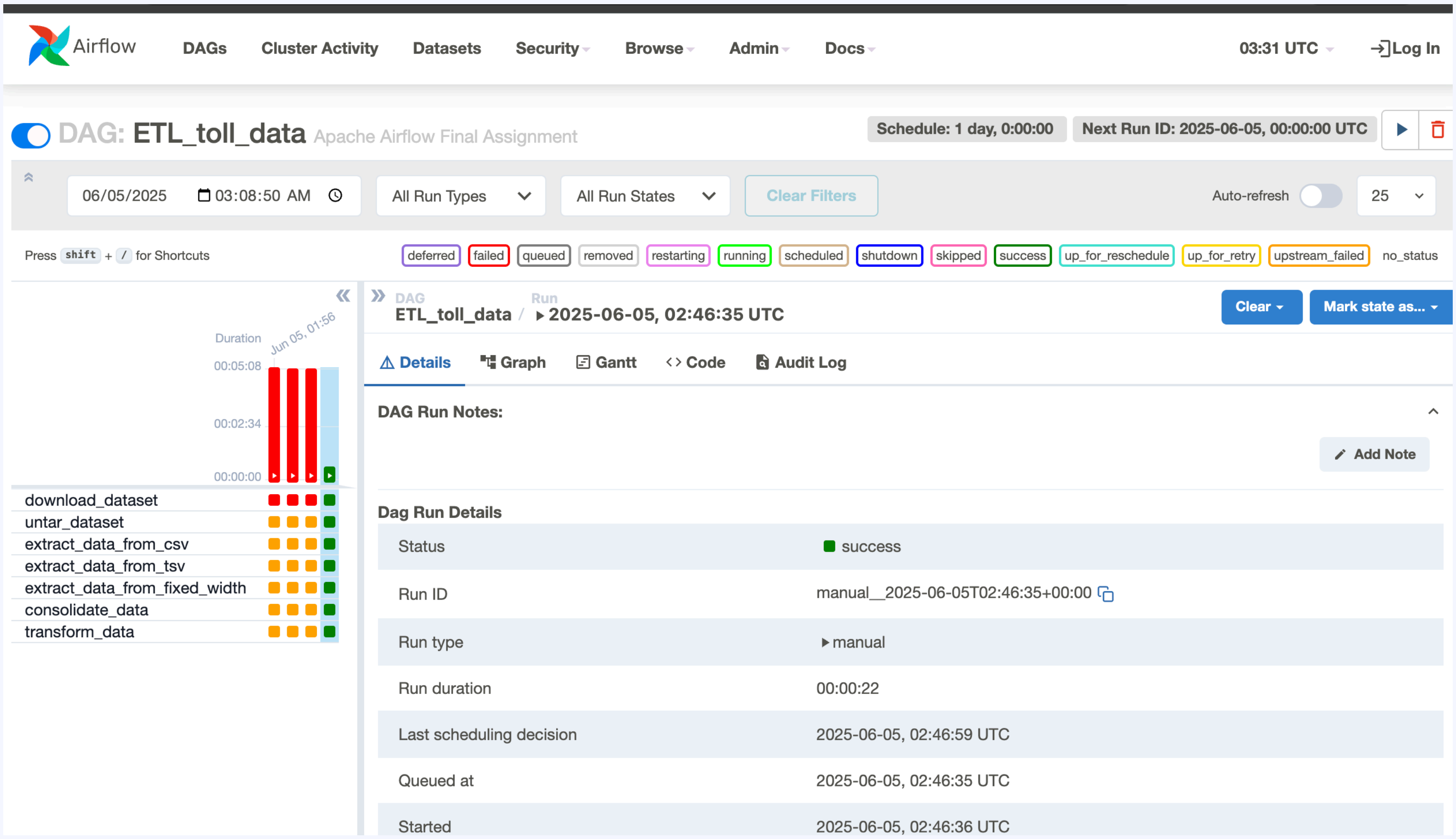
Filter by Run ID: manual__2025-06-05T02:46:35+00:0(

Filter by Task ID:

Events to ● Include ○ Exclude

Select...

WHEN	TASK ID	EVENT	OWNER	EXTRA
2025-06-05, 02:46:58 UTC	transform_data	success	Vaishnavi	
2025-06-05, 02:46:58 UTC	transform_data	running	Vaishnavi	
2025-06-05, 02:46:55 UTC	consolidate_data	success	Vaishnavi	
2025-06-05, 02:46:54 UTC	consolidate_data	running	Vaishnavi	
2025-06-05, 02:46:53 UTC	extract_data_from_fixed_width	success	Vaishnavi	



DAG: ETL_toll_data Apache Airflow Final Assignment

Schedule: 1 day, 0:00:00

Next Run ID: 2025-06-05, 00:00:00 UTC

06/05/2025 03:08:50 AM

All Run Types

All Run States

25

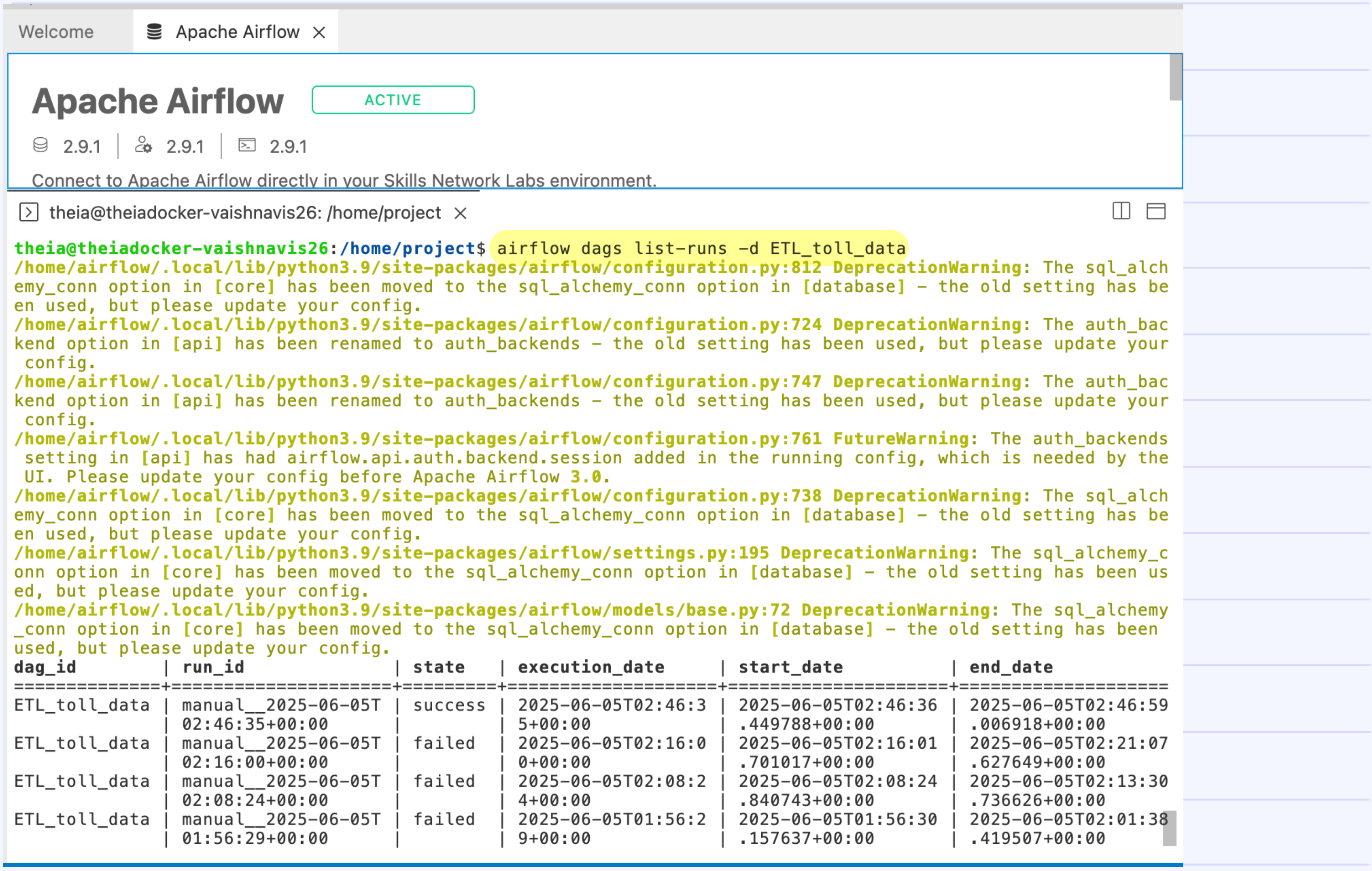
download_dataset
untar_dataset
extract_data_from_csv
extract_data_from_tsv
extract_data_from_fixed_width
consolidate_data
transform_data

DAG ETL_toll_data / Run 2025-06-05, 02:46:35 UTC

DAG Run Notes:

Dag Run Details

Status	🟩 success
Run ID	manual__2025-06-05T02:46:35+00:00
Run type	manual
Run duration	00:00:22
Last scheduling decision	2025-06-05, 02:46:59 UTC
Queued at	2025-06-05, 02:46:35 UTC
Started	2025-06-05, 02:46:36 UTC



Apache Airflow

2.9.1 | 2.9.1 | 2.9.1

Connect to Apache Airflow directly in your Skills Network Labs environment.

theia@theiadocker-vaishnavis26: /home/project

```
theia@theiadocker-vaishnavis26:/home/project$ airflow dags list-runs -d ETL_toll_data
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:812 DeprecationWarning: The sql_alch
emy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has be
en used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:724 DeprecationWarning: The auth_bac
kend option in [api] has been renamed to auth_backends - the old setting has been used, but please update your
 config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:747 DeprecationWarning: The auth_bac
kend option in [api] has been renamed to auth_backends - the old setting has been used, but please update your
 config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:761 FutureWarning: The auth_backends
 setting in [api] has had airflow.api.auth.backend.session added in the running config, which is needed by the
 UI. Please update your config before Apache Airflow 3.0.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:738 DeprecationWarning: The sql_alch
emy_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has be
en used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/settings.py:195 DeprecationWarning: The sql_alchemy_c
onn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been us
ed, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/models/base.py:72 DeprecationWarning: The sql_alchemy
_conn option in [core] has been moved to the sql_alchemy_conn option in [database] - the old setting has been
used, but please update your config.
dag_id        | run_id              | state   | execution_date      | start_date          | end_date
==============+=====================+=========+=====================+=====================+===================
ETL_toll_data | manual__2025-06-05T | success | 2025-06-05T02:46:3  | 2025-06-05T02:46:36 | 2025-06-05T02:46:59
              | 02:46:35+00:00      |         | 5+00:00             | .449788+00:00       | .006918+00:00
ETL_toll_data | manual__2025-06-05T | failed  | 2025-06-05T02:16:0  | 2025-06-05T02:16:01 | 2025-06-05T02:21:07
              | 02:16:00+00:00      |         | 0+00:00             | .701017+00:00       | .627649+00:00
ETL_toll_data | manual__2025-06-05T | failed  | 2025-06-05T02:08:2  | 2025-06-05T02:08:24 | 2025-06-05T02:13:30
              | 02:08:24+00:00      |         | 4+00:00             | .840743+00:00       | .736626+00:00
ETL_toll_data | manual__2025-06-05T | failed  | 2025-06-05T01:56:2  | 2025-06-05T01:56:30 | 2025-06-05T02:01:38
              | 01:56:29+00:00      |         | 9+00:00             | .157637+00:00       | .419507+00:00
```