

Project Overview

Scenario

You are a data engineer at a data analytics consulting company. You have been assigned a project to **decongest 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**.

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**

Build ETL Data Pipelines with BashOperator using Apache Airflow

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 the staging area as follows:
/home/project/airflow/dags/finalassignment/staging.
 - Note: -p creates parent directory as well if it does not exists
 - mkdir for creating directory
 - chmod for setting permissions for the directory
 - curl command to extract data from web

```
sudo mkdir -p /home/project/airflow/dags/finalassignment/staging
sudo chmod -R 777 /home/project/airflow/dags/finalassignment
sudo curl https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Final%20Assignment/tolldata.tgz -o /home/project/airflow/dags/finalassignment/tolldata.tgz
```

```
theia@theiadocker-vaishnavis26:/home/project$ sudo mkdir -p /home/project/airflow/dags/finalassignment/staging
theia@theiadocker-vaishnavis26:/home/project$ sudo chmod -R 777 /home/project/airflow/dags/finalassignment
theia@theiadocker-vaishnavis26:/home/project$ ls -lrt
total 4
drwxrwsrwx 5 theia users 4096 Jun  3 07:23 airflow
theia@theiadocker-vaishnavis26:/home/project$ sudo curl https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DB0250EN-SkillsNetwork/labs/Final%20Assignment/tolldata.tgz -o /home/project/airflow/dags/finalassignment/tolldata.tgz
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left   Speed
100  516k  100  516k    0     0  2954k      0 --:--:-- --:--:-- --:--:-- 2968k
theia@theiadocker-vaishnavis26:/home/project$
```

```
theia@theiadocker-vaishnavis26:/home/project/airflow/dags/finalassignment$ ls -lrt
total 524
drwxrwsrwx 2 root users  4096 Jun  3 07:38 staging
-rw-r--r-- 1 root users 528994 Jun  3 07:39 tolldata.tgz
theia@theiadocker-vaishnavis26:/home/project/airflow/dags/finalassignment$
```

Creating Imports, DAG arguments, DAG definition

Create a new file named `ETL_toll_data.py` in `/home/project` directory and open it in the file editor.

```
touch ETL_toll_data.py
```

Import all the packages you need to build the DAG.

```
from airflow.models import DAG
from airflow.operators.bash import BashOperator
from datetime import timedelta
from airflow.utils.dates import days_ago
```

Define the DAG arguments as per the following details in the `ETL_toll_data.py` file:

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

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

Define the DAG in the ETL_toll_data.py file using 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),
)
```

```
from airflow.models import DAG
from airflow.operators.bash import BashOperator
from datetime import timedelta
from airflow.utils.dates import days_ago
```

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

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

you can put a trailing comma is optional but perfectly valid in Python, especially in multi-line argument lists. Cleaner diffs in version control (Git): If you add a new line after, only one line changes

Create the tasks using BashOperator

Create a task named `unzip_data` to unzip data. Use the data downloaded in the first part of this assignment in Set up the lab environment and uncompress it into the destination directory using `tar`.

Note: It is important to give extraction directory Airflow executes tasks in a **temporary isolated working directory**

```
unzip_data = BashOperator(
    task_id='unzip_data',
    bash_command='tar -xzf /home/project/airflow/dags/finalassignment/tolldata.tgz -C /home/project/airflow/
dags/finalassignment',
    dag=dag,
)
```

Create a task 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`.

```
extract_data_from_csv = BashOperator(
    task_id='extract_data_from_csv',
    bash_command='cut -d"," -f1-4 /home/project/airflow/dags/finalassignment/vehicle-data.csv > /home/
project/airflow/dags/finalassignment/csv_data.csv',
    dag=dag,
)
```

Create a task named `extract_data_from_tsv` to extract the fields Number of axles, Tollplaza id, and Tollplaza code from the `tollplaza-data.tsv` file and save it into a file named `tsv_data.csv`.

```
extract_data_from_tsv = BashOperator(
    task_id='extract_data_from_tsv',
    bash_command='awk -F'\t' 'BEGIN {OFS=","} { print $5, $6, $7 }' /home/project/airflow/dags/
finalassignment/tollplaza-data.tsv > /home/project/airflow/dags/finalassignment/tsv_data.csv',
    dag=dag,
)
```

Create a task 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`

```
extract_data_from_fixed_width = BashOperator(
    task_id='extract_data_from_fixed_width',
    bash_command='awk '{ print $10 "\t" $11 }' /home/project/airflow/dags/finalassignment/payment-data.txt
> /home/project/airflow/dags/finalassignment/fixed_width_data.csv',
    dag=dag,
)
```

Create a task named `consolidate_data` to consolidate data extracted from previous tasks. This task should create a single csv file named `extracted_data.csv` by combining data from the following files:

```
consolidate_data = BashOperator(
    task_id='consolidate_data',
    bash_command='paste -d',' /home/project/airflow/dags/finalassignment/csv_data.csv /home/project/airflow/
dags/finalassignment/tsv_data.csv /home/project/airflow/dags/finalassignment/fixed_width_data.csv > /home/
project/airflow/dags/finalassignment/extracted_data.csv',
    dag=dag,
)
```

Create a task 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.

```
transform_data = BashOperator(
    task_id='transform_data',
    bash_command='awk -F',' 'BEGIN {OFS=","} { $4 = toupper($4); print }' /home/project/airflow/dags/
finalassignment/extracted_data.csv > /home/project/airflow/dags/finalassignment/transformed_data.csv',
    dag=dag,
)
```

Define the task pipeline as per the details given below:

Task	Functionality
First task	unzip_data
Second task	extract_data_from_csv
Third task	extract_data_from_tsv
Fourth task	extract_data_from_fixed_width
Fifth task	consolidate_data
Sixth task	transform_data

```
unzip_data >> extract_data_from_csv >> extract_data_from_tsv >> extract_data_from_fixed_width >> consolidate_data >> transform_data
```

Getting the DAG operational

Submit the DAG. Use CLI or Web UI to show that the DAG has been properly submitted

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


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
```

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

| Vaishnavi | True
theia@theiadocker-vaishnavis26:/home/project$ airflow dags
list| grep "ETL_toll_data"
ETL_toll_data |
/home/project/airflow/dags/ETL_toll_data.py

| Vaishnavi | True
theia@theiadocker-vaishnavis26:/home/project$
```

 Airflow

DAGsCluster ActivityDatasetsSecurityBrowseAdminDocs

04:09

Skills Network Airflow

All 60Active 0Paused 60Running 0Failed 0

Filter DAGs by tag

Search DAGs

DAG	Owner	Runs	Schedule	Last Run	Next Run
<div><div></div>ETL_toll_data</div>	Vaishnavi	<div><div></div><div></div><div></div><div></div></div>	1 day, 0:00:00	2025-06-04, 00:00:00	

Unpause and trigger the DAG through CLI or Web UI.

```
airflow dags unpause ETL_toll_data
```

```
theia@theiadocker-vaishnavis26:/home/project$
theia@theiadocker-vaishnavis26:/home/project$ airflow dags unpause ETL_toll_data
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:812 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.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:724 DeprecationWarning:
The auth_backend option in [api] has been renamed to auth_backends - the old setting has been us
ed, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:747 DeprecationWarning:
The auth_backend option in [api] has been renamed to auth_backends - the old setting has been us
ed, 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 conf
ig, 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_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 [database] - t
he old setting has been used, but please update your config.
/home/airflow/.local/lib/python3.9/site-packages/airflow/models/base.py:72 DeprecationWarning: Th
e 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.
[2025-06-04T04:18:44.934+0000] {dagbag.py:545} INFO - Filling up the DagBag from /home/project/ai
rflow/dags
dag_id          | is_paused
=====+=====
ETL_toll_data   | True

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



Airflow

DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

04:19 UTC

Log In

Skills Network Airflow

All 60

Active 1

Paused 59

Running 0

Failed 0

Filter DAGs by tag

Search DAGs

Auto-refresh



DAG

Owner

Runs

Schedule

Last Run

Next Run

Recent Tasks

ETL_toll_data

Vaishnavi

1 day, 0:00:00

2025-06-04, 00:00:00


tasks in the DAG run through CLI or Web UI


```
airflow tasks list ETL_toll_data
```

```

theia@theiadocker-vaishnavis26:/home/project$ airflow tasks list ETL_toll_data
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:812 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.
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:724 DeprecationWarning:
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 DeprecationWarning:
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:
auth_backends setting in [api] has had airflow.api.auth.backend.session added in the running, 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_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:
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/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.
consolidate_data
extract_data_from_csv
extract_data_from_fixed_width
extract_data_from_tsv
transform_data
unzip_data
theia@theiadocker-vaishnavis26:/home/project$

```


Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
04:26 UTC
Log In


DAG: ETL_toll_data
Apache Airflow Final Assignment

Schedule: 1 day, 0:00:00
Next Run ID: 2025-06-04, 00:00:00 UTC

06/04/2025
04:23:42 AM
All Run Types
All Run States
Clear Filters
Auto-refresh
25

deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status

unzip_data

extract_data_from_csv

extract_data_from_tsv

extract_data_from_fixed_width

consolidate_data

transform_data

unzip_data

extract_data_from_csv

extract_data_from_tsv

extract_data_from_fixed_width

consolidate_data

transform_data

ETL_toll_data

Details

Graph

Gantt

Code

Audit Log

Run Duration

Calendar

Layout:

Left -> Right

unzip_data

extract_data_from_csv

extract_data_from_tsv

extract_data_from_fixed_width

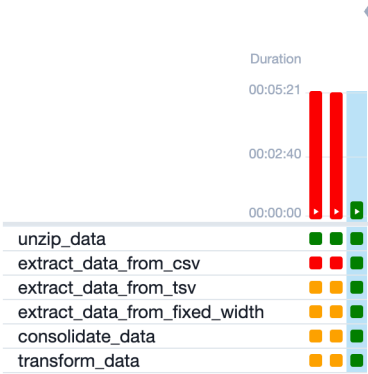
consolidate_data

transform_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



DAG

Run

ETL_toll_data / ▶ 2025-06-04, 04:58:16 UTC

Clear

Mark state

Details

Graph

Gantt

<> Code

Audit Log

Dag Run Details

Status	■ success
Run ID	manual__2025-06-04T04:58:16+00:00 🔗
Run type	▶ manual
Run duration	00:00:28
Last scheduling decision	2025-06-04, 04:58:45 UTC
Queued at	2025-06-04, 04:58:16 UTC
Started	2025-06-04, 04:58:17 UTC
Ended	2025-06-04, 04:58:45 UTC
Data interval start	2025-06-03, 04:58:16 UTC
Data interval end	2025-06-04, 04:58:16 UTC



△ Details  Graph  Gantt <> Code  Audit Log

WHEN ▾	TASK ID ↕	EVENT ↕	OWNER ↕	EXTRA ↕
2025-06-04, 04:58:44 UTC	transform_data	success	Vaishnavi	
2025-06-04, 04:58:42 UTC	transform_data	running	Vaishnavi	
2025-06-04, 04:58:38 UTC	consolidate_data	success	Vaishnavi	
2025-06-04, 04:58:37 UTC	consolidate_data	running	Vaishnavi	
2025-06-04, 04:58:33 UTC	extract_data_from_fixed_width	success	Vaishnavi	
2025-06-04, 04:58:33 UTC	extract_data_from_fixed_width	running	Vaishnavi	
2025-06-04, 04:58:29 UTC	extract_data_from_tsv	success	Vaishnavi	
2025-06-04, 04:58:29 UTC	extract_data_from_tsv	running	Vaishnavi	
2025-06-04, 04:58:26 UTC	extract_data_from_csv	success	Vaishnavi	
2025-06-04, 04:58:25 UTC	extract_data_from_csv	running	Vaishnavi	
2025-06-04, 04:58:22 UTC	unzip_data	success	Vaishnavi	
2025-06-04, 04:58:21 UTC	unzip_data	running	Vaishnavi	

 theia@theiadocker-vaishnavis26: /home/project

 theia@theiadocker-vaishnavis26: /home/project ×  

```
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_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 DeprecationWarning:  
The auth_backend option in [api] has been renamed to auth_backends - the old setting has been us  
ed, but please update your config.  
/home/airflow/.local/lib/python3.9/site-packages/airflow/configuration.py:747 DeprecationWarning:  
The auth_backend option in [api] has been renamed to auth_backends - the old setting has been us  
ed, 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 conf  
ig, 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_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 [database] - t  
he old setting has been used, but please update your config.  
/home/airflow/.local/lib/python3.9/site-packages/airflow/models/base.py:72 DeprecationWarning: Th  
e 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__2025-06-04T04:58:16+00:00	success	2025-06-04T04:58:16+00:00	2025-06-04T04:58:17.061256+00:00	2025-06-04T04:58:45.079294+00:00
ETL_toll_data	manual__2025-06-04T04:41:16+00:00	failed	2025-06-04T04:41:16+00:00	2025-06-04T04:41:17.767345+00:00	2025-06-04T04:46:37.185339+00:00
ETL_toll_data	manual__2025-06-04T04:2	failed	2025-06-04T04:2	2025-06-04T04:2	2025-06-04T04:35:

