

# MLOps

## Assignment# 2

**Submitted by:**

Huda

22i8790

**Submitted to:**

Pir Sami Ullah Shah



# NASA APOD ETL Pipeline with Airflow, DVC, and Postgres

This project implements an ETL pipeline that extracts data from NASA's Astronomy Picture of the Day (APOD) API, transforms it, loads it into PostgreSQL and CSV files, and versions it using DVC and Git. The pipeline consists of 5 sequential steps:

- Extract (E): Retrieves daily data from NASA APOD API
- Transform (T): Processes JSON data into structured format using Pandas
- Load (L): Persists data to both PostgreSQL database and CSV file
- Data Versioning (DVC): Versions the CSV file using DVC
- Code Versioning (Git): Commits DVC metadata to Git repository

This project successfully demonstrates a complete ETL pipeline with data versioning capabilities. The implementation showcases best practices in MLOps including workflow orchestration, data integrity, version control, and containerized deployment. The pipeline is production-ready and can be easily deployed to cloud platforms like Astronomer.

## Screenshots

The screenshot shows the Airflow web interface for the DAG `nasa_apod_etl_pipeline`. The top navigation bar includes links for Airflow, DAGs, Cluster Activity, Datasets, Security, Browse, Admin, and Docs. The current time is 19:34 UTC. The main header displays the DAG name and a brief description: "ETL pipeline for NASA APOD data with DVC versioning". Below the header, there are tabs for Grid, Graph, Calendar, Task Duration, Task Tries, Landing Times, Gantt, Details, Code, and Audit Log. The Grid tab is selected. The interface shows a timeline from 13/11/2025, 07:33:56 PM to 25/11/2025, 07:33:56 PM. A "Clear Filters" button is present. A legend at the bottom indicates various task states: deferred, failed, queued, removed, restarting, running, scheduled, skipped, success, up\_for\_reschedule, up\_for\_retry, upstream\_failed, and no\_status. On the left, a sidebar lists the tasks: extract\_nasa\_data, transform\_data, load\_data, version\_data\_with\_dvc, and commit\_to\_git. The main content area displays the DAG summary and details. The DAG summary table shows 5 total tasks and 5 PythonOperators. The DAG details table provides information about the DAG ID (nasa\_apod\_etl\_pipeline), description (ETL pipeline for NASA APOD data with DVC versioning), file location (/opt/airflow/dags/nasa\_apod\_etl\_dag.py), and other properties like import errors, task concurrency limits, and active status. At the bottom, a footer notes the version (v2.8.0) and Git version (release:db2b75c233e3e3c59ec9d0563b93dbe733ad0bf).

**DAG: nasa\_apod\_etl\_pipeline** ETL pipeline for NASA APOD data with DVC versioning

Schedule: 1 day, 0:00:00 | Next Run: 2025-11-13, 00:00:00

14/11/2025, 06:37:41 PM | 25 | All Run Types | All Run States | Clear Filters | Auto-refresh

Press shift + / for Shortcuts | deferred failed queued removed restarting running scheduled skipped success up\_for\_reschedule up\_for\_retry upstream\_failed no\_status

» DAG nasa\_apod\_etl\_pipeline

Details | Graph | Gantt | Code

Layout: Left -> Right

React Flow

Version: v2.8.0  
Git Version: release:db2b75c233e3e3c59ec9d0563b93dbe733ad0bf

**DAG: nasa\_apod\_etl\_pipeline** ETL pipeline for NASA APOD data with DVC versioning

Schedule: 1 day, 0:00:00 | Next Run: 2025-11-13, 00:00:00

14/11/2025, 06:38 PM | 25 | All Run Types | All Run States | Clear Filters | Auto-refresh

Press shift + / for Shortcuts | deferred failed queued removed restarting running scheduled skipped success up\_for\_reschedule up\_for\_retry upstream\_failed no\_status

» DAG nasa\_apod\_etl\_pipeline

Details | Graph | Gantt | Code

Parsed at: 2025-11-14, 18:37:33 UTC

```

11 from datetime import datetime, timedelta
12 from airflow import DAG
13 from airflow.operators.python import PythonOperator
14 from airflow.providers.postgres.operators.postgres import PostgresOperator
15 from airflow.providers.postgres.hooks.postgres import PostgresHook
16 import requests
17 import pandas as pd
18 import json
19 import os
20 import subprocess
21 import logging
22
23 # Default arguments for the DAG
24 default_args = {
25     'owner': 'mlops',
26     'depends_on_past': False,
27     'email_on_failure': False,
28     'email_on_retry': False,
29     'retries': 1,
30     'retry_delay': timedelta(minutes=5),
31 }
32
33 # DAG definition
34 dag = DAG(
    ...
)

```

Toggle Wrap

Version: v2.8.0  
Git Version: release:db2b75c233e3e3c59ec9d0563b93dbe733ad0bf

 Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs 19:01 UTC AA

DAG: **nasa\_apod\_etl\_pipeline** ETL pipeline for NASA APOD data with DVC versioning Schedule: 1 day, 0:00:00 Next Run: 2025-11-13, 00:00:00

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

14/11/2025, 07:01:26 PM 25 All Run Types All Run States Clear Filters Auto-refresh

Press shift + / for Shortcuts

Deferred Failed Queued Removed Restarting Running Scheduled Skipped Success Up\_for\_reschedule Up\_for\_retry Upstream\_failed No\_status

« » DAG nasa\_apod\_etl\_pipeline

Details Graph Gantt Code

**DAG Summary**

Total Tasks	5
PythonOperators	5

**DAG Details**

Dag id	nasa_apod_etl_pipeline
Description	ETL pipeline for NASA APOD data with DVC versioning
Filenloc	/opt/airflow/dags/nasa_apod_etl_dag.py
Has import errors	false
Has task concurrency limits	false
Is active	true
Is paused	true
Is subdag	false

Version: v2.8.0  
Git Version: release:db2b75c233e3e3c59ec9d0563b93dbe733ad0bf

Version: v2.8.0  
Git Version: release:db2b75c233e3e3c59ec9d0563b93ddbe733ad0bf

The screenshot shows the Airflow web interface for the 'nasa\_apod\_etl\_pipeline'. At the top, there's a navigation bar with links for DAGs, Cluster Activity, Datasets, Security, Browse, Admin, and Docs. The time is shown as 19:01 UTC. On the right, there are user settings and a language selector. Below the navigation is a search bar with the text 'DAG: nasa\_apod\_etl\_pipeline ETL pipeline for NASA APOD data with DVC versioning'. To the right of the search bar are buttons for 'Schedule: 1 day, 0:00:00' and 'Next Run: 2025-11-13, 00:00:00'. The main content area features a toolbar with 'Grid', 'Graph', 'Calendar', 'Task Duration', 'Task Tries', 'Landing Times', 'Gantt' (which is selected), 'Details', 'Code', and 'Audit Log'. Below the toolbar is a date range selector showing '14/11/2025, 07:01:26 PM' and a dropdown for '25'. There are also dropdowns for 'All Run Types' and 'All Run States'. A 'Clear Filters' button is located in the center of these dropdowns. To the right is an 'Auto-refresh' toggle. A keyboard shortcut keytip is displayed: 'Press shift + / for Shortcuts'. Below the toolbar, a series of colored status indicators are shown: 'deferred' (grey), 'failed' (red), 'queued' (light blue), 'removed' (pink), 'restarting' (yellow), 'running' (green), 'scheduled' (light green), 'skipped' (purple), 'success' (blue), 'up\_for\_reschedule' (light blue), 'up\_for\_retry' (light green), 'upstream\_failed' (orange), and 'no\_status' (yellow). The main workspace displays the DAG structure for 'nasa\_apod\_etl\_pipeline'. It includes a sidebar with the DAG's tasks: 'extract\_nasa\_data', 'transform\_data', 'load\_data', 'version\_data\_with\_dvc', and 'commit\_to\_git'. The main area shows a Gantt chart grid with columns representing dates from 19:01:45 UTC to 19:01:46 UTC. Each column has a corresponding row for each task. A message box at the top of the grid says 'Please select a dag run in order to see a gantt chart'. The entire interface is set against a light grey background.



DAGs

Cluster Activity

Datasets

Security

Browse

Admin

Docs

14:15 UTC

AA

**DAG: nasa\_apod\_etl\_pipeline** ETL pipeline for NASA APOD data with DVC versioning

Schedule: 1 day, 0:00:00 | Next Run: 2025-11-14, 00:00:00

Grid

Graph

Calendar

Task Duration

Task Tries

Landing Times

Gantt

Details

Code

Audit Log

▶

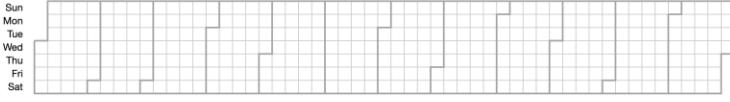
✖

success failed running planned no\_status

2024



2025



Version: v2.8.0

Git Version: .release:db2b75c233e3e3c59ec9d0563b93dbe733ad0bf

Airflow DAGs Cluster Activity Datasets Security - Browse - Admin - Docs - 14:16 UTC - AA -

DAG: nasa\_apod\_etl\_pipeline ETL pipeline for NASA APOD data with DVC versioning Schedule: 1 day, 0:00:00 | Next Run: 2025-11-14, 00:00:00

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

### DAG Details

Schedule Interval	1 day, 0:00:00
Catchup	False
Started	False
End Date	None
Max Active Runs	0 / 16
Concurrency	16
Default Args	{'depends_on_past': False, 'email_on_failure': False, 'email_on_retry': False, 'owner': 'mlops', 'retries': 1, 'retry_delay': datetime.timedelta(seconds=300)}
Tasks Count	5
Task IDs	['extract_nasa_data', 'transform_data', 'load_data', 'version_data_with_dvc', 'commit_to_git']
Relative file location	nasa_apod_etl_dag.py
Owner	mlops
Owner Links	None
DAG Run Timeout	None
Tags	dvc etl mlops nasa
DagModel debug information	
Attribute	Value
fileloc	/opt/airflow/dags/nasa_apod_etl_dag.py
has_import_errors	False
has_task_concurrency_limits	False
is_active	True
is_paused_at_creation	True
is_subdag	False
last_expired	None
last_parsed_time	2025-11-15 14:16:04.340568+00:00
last_picked	None
metadata	MetaData()
next_dagrun	2025-11-14 00:00:00+00:00
next_dagrun_create_after	2025-11-15 00:00:00+00:00
next_dagrun_data_interval	DataInterval(start=DateTime(2025, 11, 14, 0, 0, tzinfo=Timezone('UTC')), end=DateTime(2025, 11, 15, 0, 0, tzinfo=Timezone('UTC')))
next_dagrun_data_interval_end	2025-11-15 00:00:00+00:00
next_dagrun_data_interval_start	2025-11-14 00:00:00+00:00
parent_dag	None
pickle_id	None
processor_subdir	/opt/airflow/dags
registry	<sqlalchemy.orm.decl_api.registry object at 0xfffff8a63a250>
root_dag_id	None
safe_dag_id	nasa_apod_etl_pipeline
scheduler_lock	None
timetable_description	
timezone	Timezone('UTC')

Version: v2.8.0  
Git Version: .release:db2b75c233e3e3c59ec9d0563b93ddbe733ad0bf

The screenshot shows the Docker Desktop application window. On the left is a sidebar with navigation links: Ask Gordon (BETA), Containers, Images (selected), Volumes, Kubernetes, Builds, Models, MCP Toolkit (BETA), Docker Hub, Docker Scout, and Extensions. The main area is titled "Images" with a "Local" tab selected, showing "My Hub". It displays 10 images with the following details:

	Name	Tag	Image ID	Created	Size	Actions
untitledfolder-frontend	latest	b3defd88a3a1	23 days ago	76.83 MB	<a href="#">View</a> <a href="#">... Edit</a>	
untitledfolder-backend-service	latest	8610ae4f5232	23 days ago	453.02 MB	<a href="#">View</a> <a href="#">... Edit</a>	
untitledfolder-auth-service	latest	9a319c81f37e	23 days ago	449 MB	<a href="#">View</a> <a href="#">... Edit</a>	
mongo	7.0	c258b26dbb77	1 month ago	1.07 GB	<a href="#">View</a> <a href="#">... Edit</a>	
gcr.io/k8s-minikube/kicbase	v0.0.48	18b0fda1fa3d	2 months ago	1.73 GB	<a href="#">View</a> <a href="#">... Edit</a>	
gcr.io/k8s-minikube/kicbase	<none>	7171c97a5162	2 months ago	1.73 GB	<a href="#">View</a> <a href="#">... Edit</a>	
postgres	15	822f8795764a	10 days ago	648.55 MB	<a href="#">View</a> <a href="#">... Edit</a>	
mlopsa3-airflow-scheduler	latest	53299dc92a6b	25 minutes ag	1.95 GB	<a href="#">View</a> <a href="#">... Edit</a>	
mlopsa3-airflow-webserver	latest	8408c0133743	25 minutes ag	1.95 GB	<a href="#">View</a> <a href="#">... Edit</a>	
mlopsa3-airflow-init	latest	31d3f3b333c8	25 minutes ag	1.95 GB	<a href="#">View</a> <a href="#">... Edit</a>	

At the bottom, status information includes "Engine running", system resources (RAM 6.25 GB, CPU 0.40%, Disk: 18.92 GB used / limit 452.13 GB), and links for "Terminal" and "Update available".