MLOps

Assignment# 2

Submitted by:
Huda
22i8790

Submitted to:
Pir Sami Ullah Shah

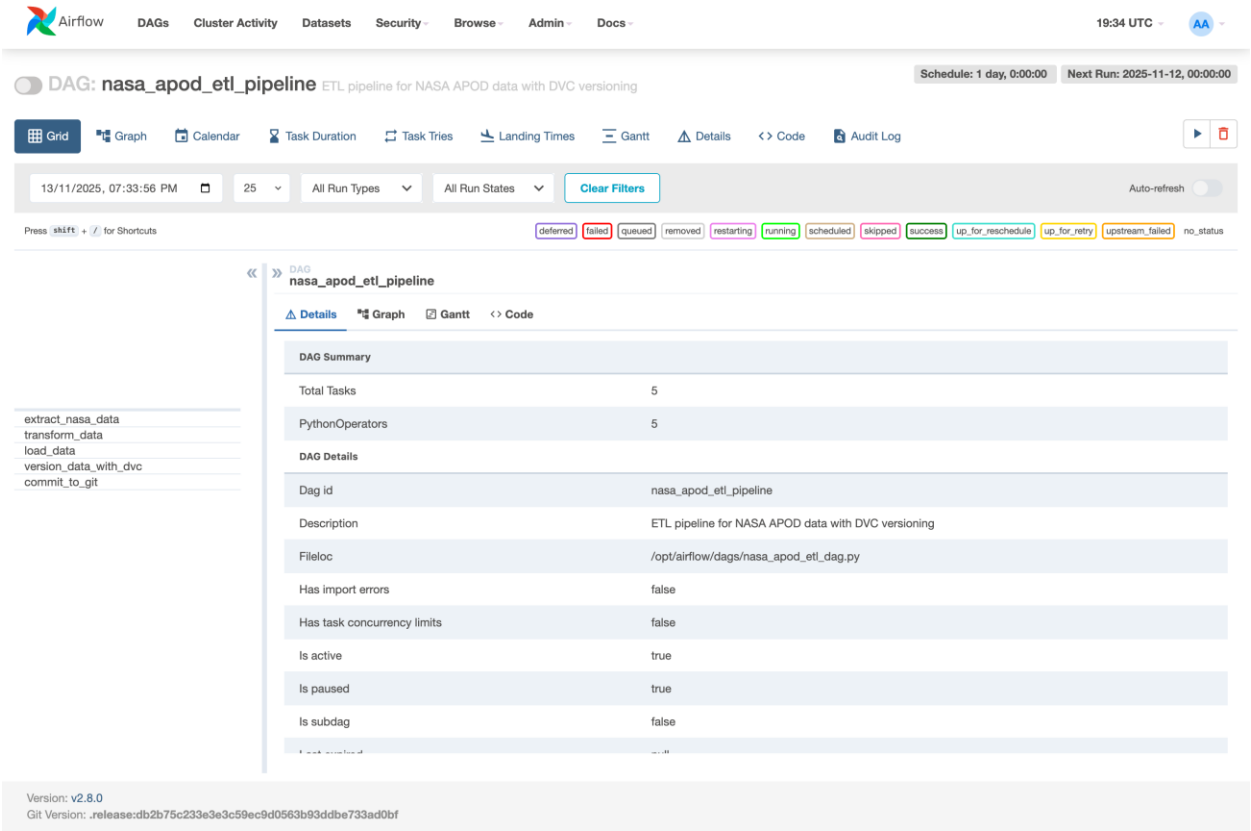
NASA APOD ETL Pipeline with Airflow, DVC, and Postgres

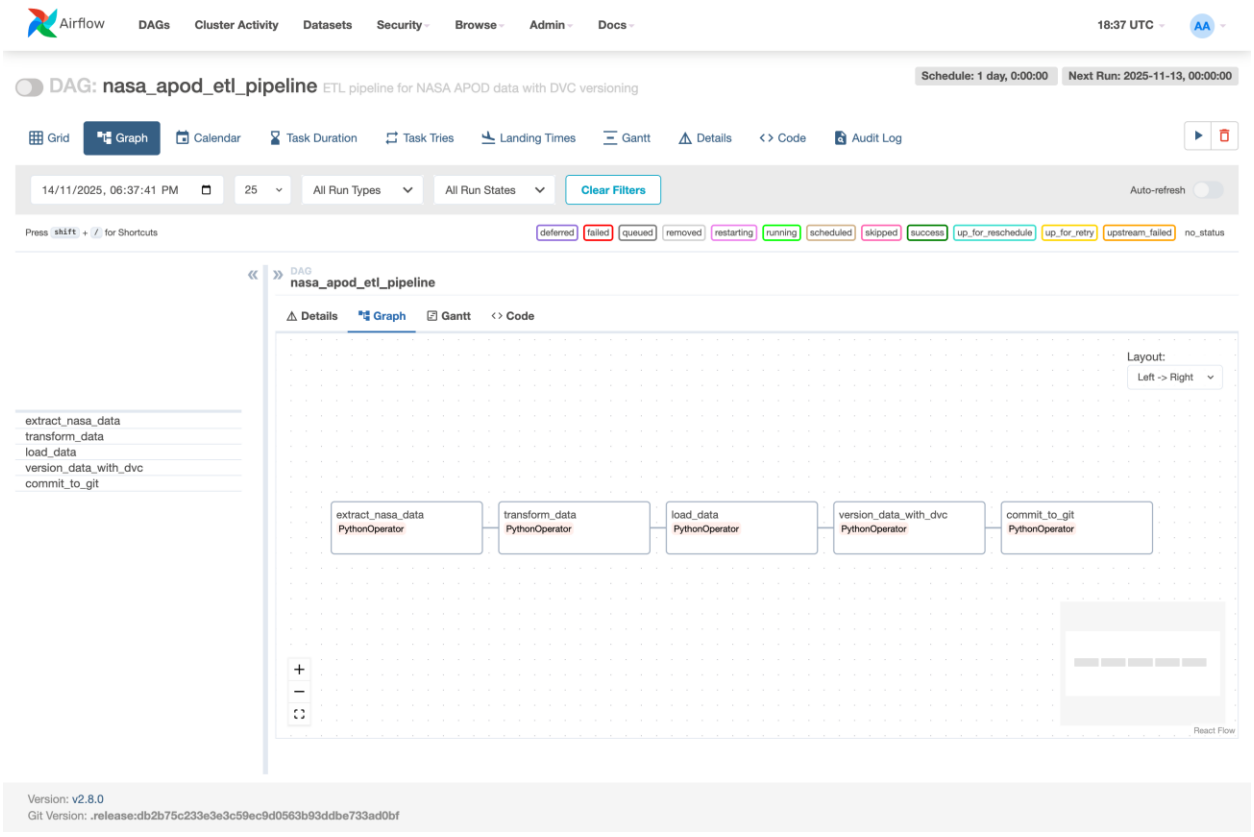
This project implements an ETL pipeline that extracts data from NASA's Astronomy Picture of the Day (APOD) API, transforms it, loads it into PostgreSQL and CSV files, and versions it using DVC and Git. The pipeline consists of 5 sequential steps:

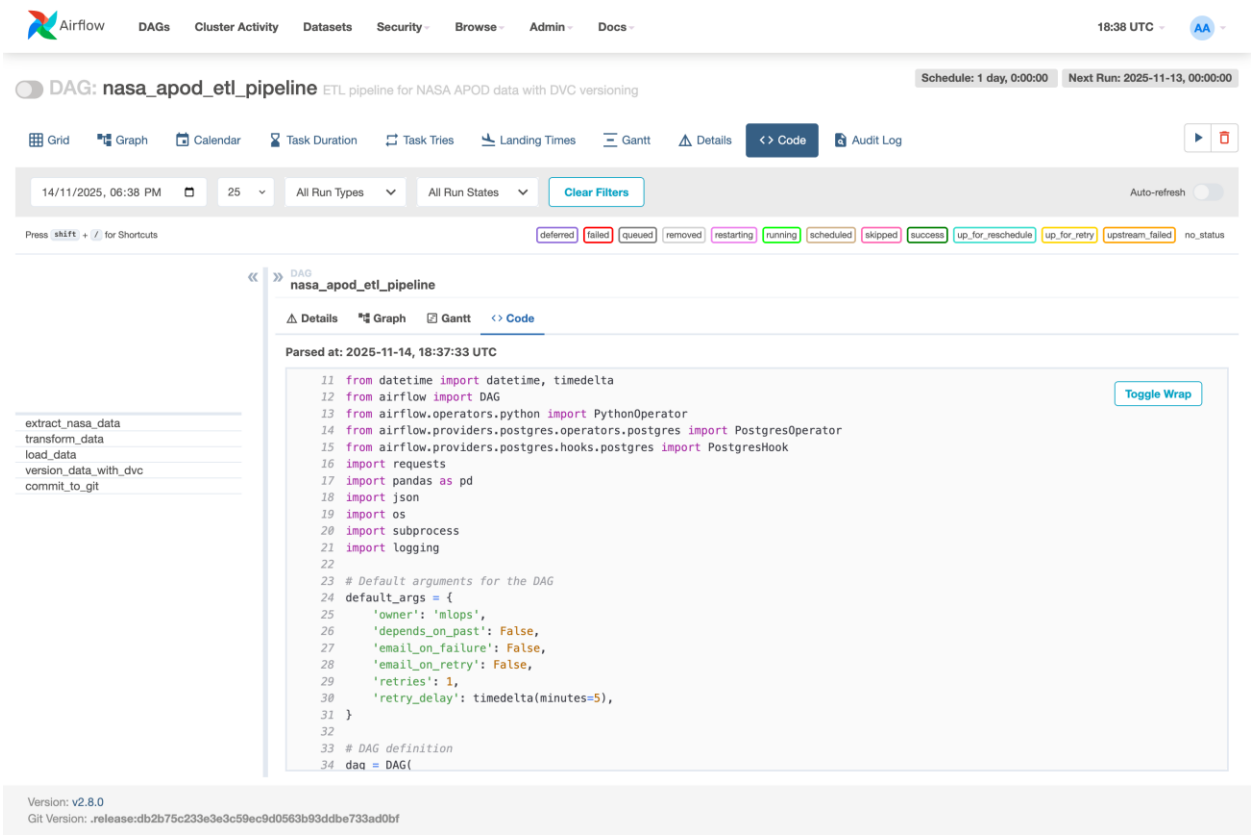
- Extract (E): Retrieves daily data from NASA APOD API
- Transform (T): Processes JSON data into structured format using Pandas
- Load (L): Persists data to both PostgreSQL database and CSV file
- Data Versioning (DVC): Versions the CSV file using DVC
- Code Versioning (Git): Commits DVC metadata to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

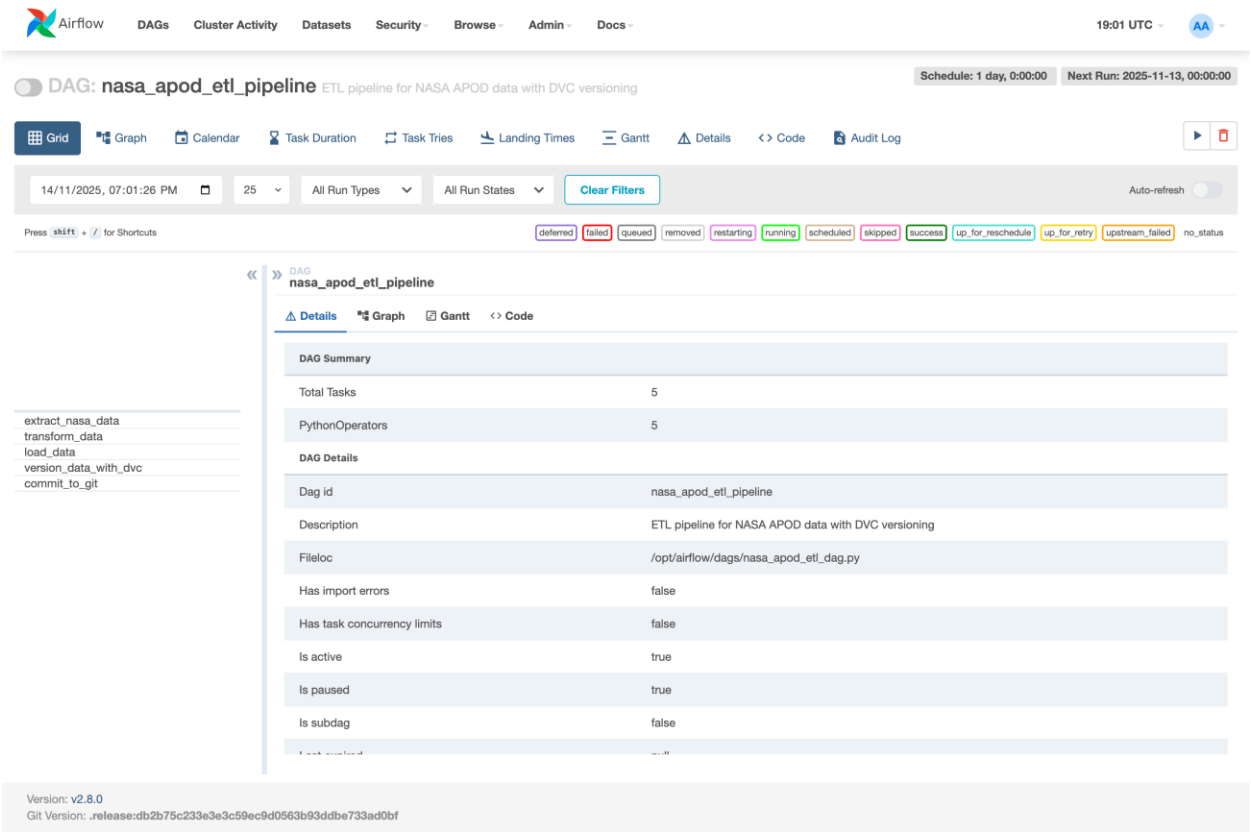
This project successfully demonstrates a complete ETL pipeline with data versioning capabilities. The implementation showcases best practices in MLOps including workflow orchestration, data integrity, version control, and containerized deployment. The pipeline is production-ready and can be easily deployed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

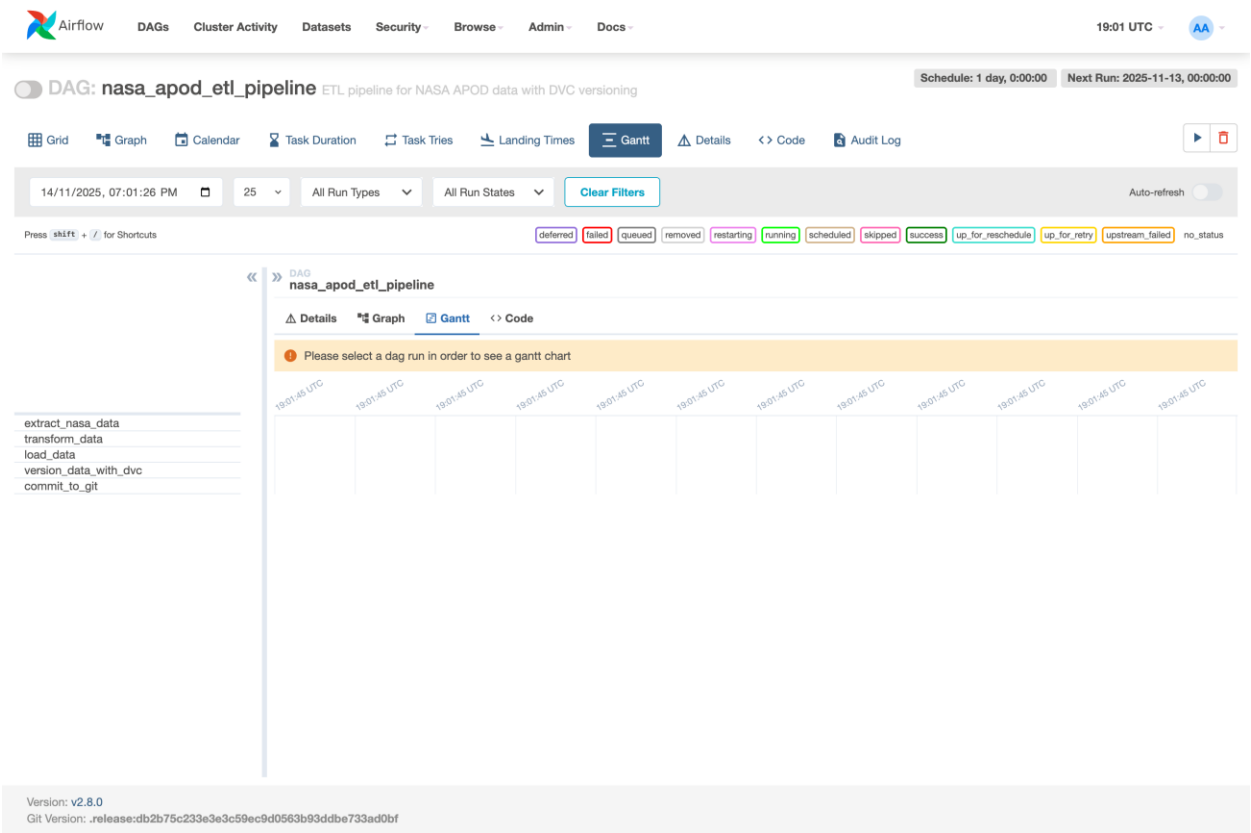
Screenshots



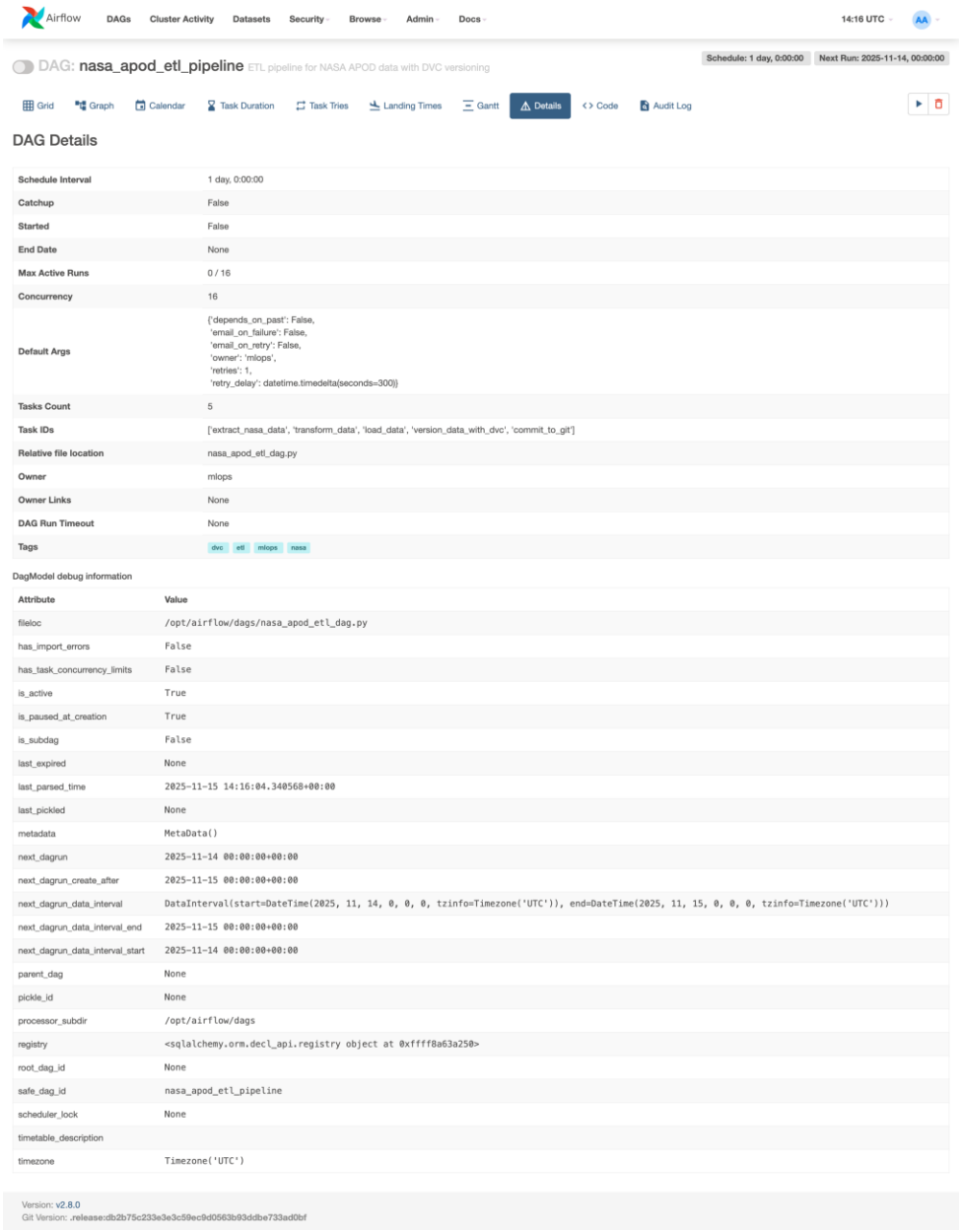

Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
18:37 UTC
AA
DAG: nasa_apod_etl_pipeline ETL pipeline for NASA APOD data with DVC versioning
Schedule: 1 day, 0:00:00
Next Run: 2025-11-13, 00:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
14/11/2025, 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
extract_nasa_data
transform_data
load_data
version_data_with_dvc
commit_to_git
extract_nasa_data
PythonOperator
transform_data
PythonOperator
load_data
PythonOperator
version_data_with_dvc
PythonOperator
commit_to_git
PythonOperator
React Flow
Version: v2.8.0
Git Version: .release:db2b75c233e3e3c59ec9d0563b93ddbe733ad0bf

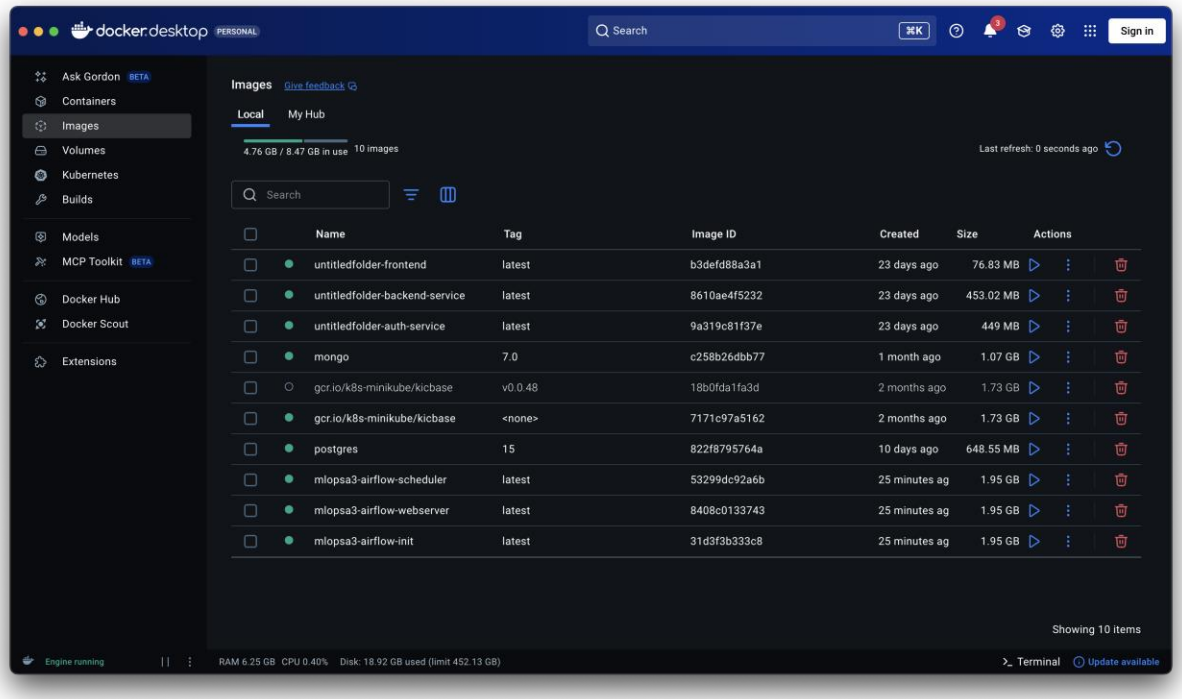

Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
18:38 UTC
AA
DAG: nasa_apod_etl_pipeline ETL pipeline for NASA APOD data with DVC versioning
Schedule: 1 day, 0:00:00
Next Run: 2025-11-13, 00:00:00
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
14/11/2025, 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
Toggle Wrap
extract_nasa_data
transform_data
load_data
version_data_with_dvc
commit_to_git
from datetime import datetime, timedelta
from airflow import DAG
from airflow.operators.python import PythonOperator
from airflow.providers.postgres.operators.postgres import PostgresOperator
from airflow.providers.postgres.hooks.postgres import PostgresHook
import requests
import pandas as pd
import json
import os
import subprocess
import logging
# Default arguments for the DAG
default_args = {
'owner': 'mlops',
'depends_on_past': False,
'email_on_failure': False,
'email_on_retry': False,
'retries': 1,
'retry_delay': timedelta(minutes=5),
}
# DAG definition
dag = DAG(
Version: v2.8.0
Git Version: .release:db2b75c233e3e3c59ec9d0563b93ddbe733ad0bf


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
extract_nasa_data
transform_data
load_data
version_data_with_dvc
commit_to_git
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
Fileloc
/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:db2b75c233e3e3c59ec9d0563b93ddbe733ad0bf


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
extract_nasa_data
transform_data
load_data
version_data_with_dvc
commit_to_git
DAG
nasa_apod_etl_pipeline
Details
Graph
Gantt
Code
Please select a dag run in order to see a gantt chart
19:01:45 UTC
Version: v2.8.0
Git Version: .release:db2b75c233e3e3c59ec9d0563b93ddbe733ad0bf


Airflow
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
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Version: v2.8.0
Git Version: .release:db2b75c233e3e3c59ec9d0563b93ddbe733ad0bf


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_pickled
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, 0, tzinfo=Timezone('UTC')), end=DateTime(2025, 11, 15, 0, 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 0xffff8a63a250>
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


docker desktop
PERSONAL
Search
⌘K
Sign in
Ask Gordon BETA
Containers
Images
Volumes
Kubernetes
Builds
Models
MCP Toolkit BETA
Docker Hub
Docker Scout
Extensions
Images
Give feedback
Local
My Hub
4.76 GB / 8.47 GB in use
10 images
Last refresh: 0 seconds ago
Search
Name
Tag
Image ID
Created
Size
Actions
untitledfolder-frontend latest b3defd88a3a1 23 days ago 76.83 MB
untitledfolder-backend-service latest 8610ae4f5232 23 days ago 453.02 MB
untitledfolder-auth-service latest 9a319c81f37e 23 days ago 449 MB
mongo 7.0 c258b26dbb77 1 month ago 1.07 GB
gcr.io/k8s-minikube/kicbase v0.0.48 18b0fda1fa3d 2 months ago 1.73 GB
gcr.io/k8s-minikube/kicbase <none> 7171c97a5162 2 months ago 1.73 GB
postgres 15 822f8795764a 10 days ago 648.55 MB
mlopsa3-airflow-scheduler latest 53299dc92a6b 25 minutes ag 1.95 GB
mlopsa3-airflow-webserver latest 8408c0133743 25 minutes ag 1.95 GB
mlopsa3-airflow-init latest 31d3f3b333c8 25 minutes ag 1.95 GB
Showing 10 items
Engine running
RAM 6.25 GB CPU 0.40% Disk: 18.92 GB used (limit 452.13 GB)
Terminal
Update available