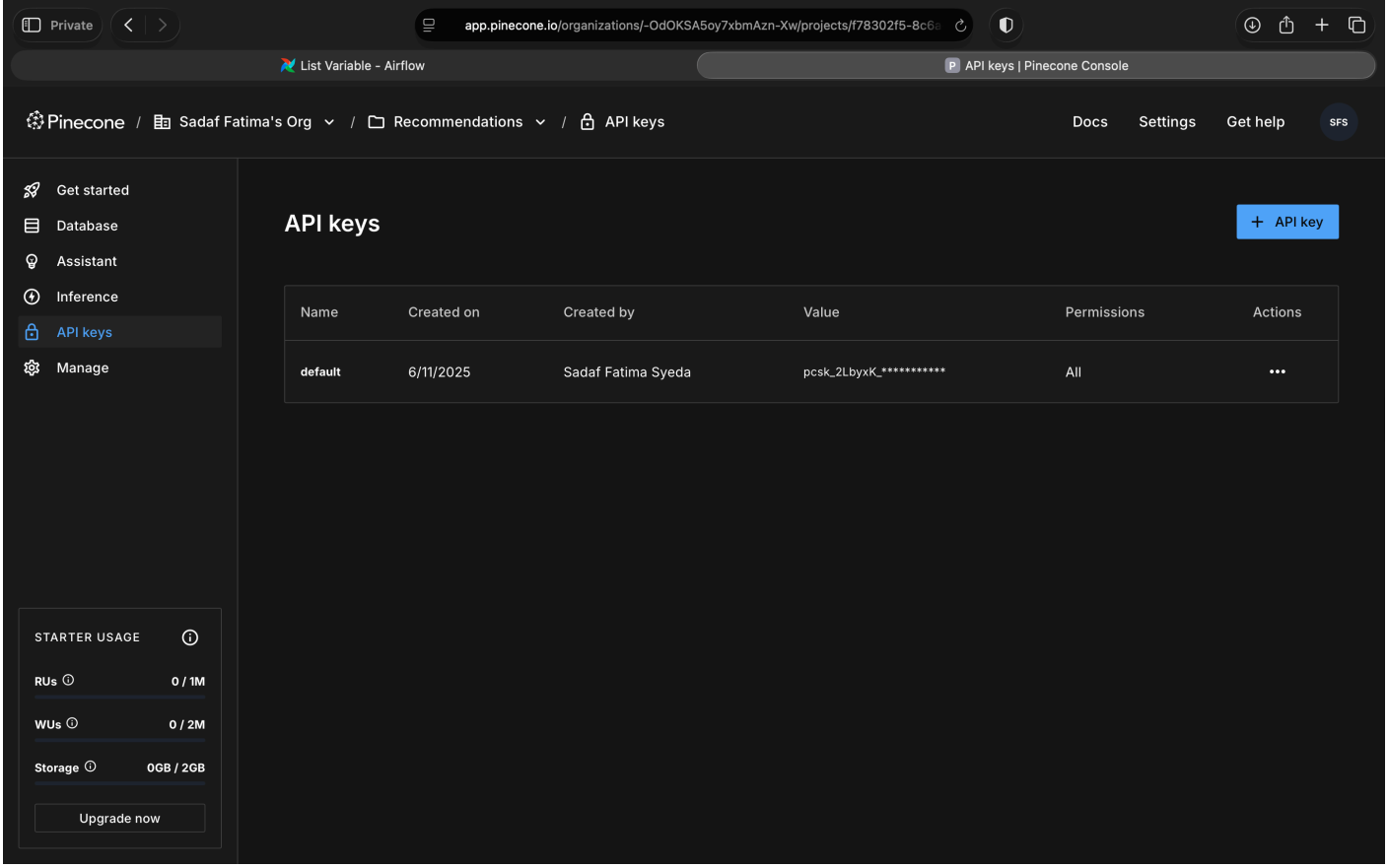
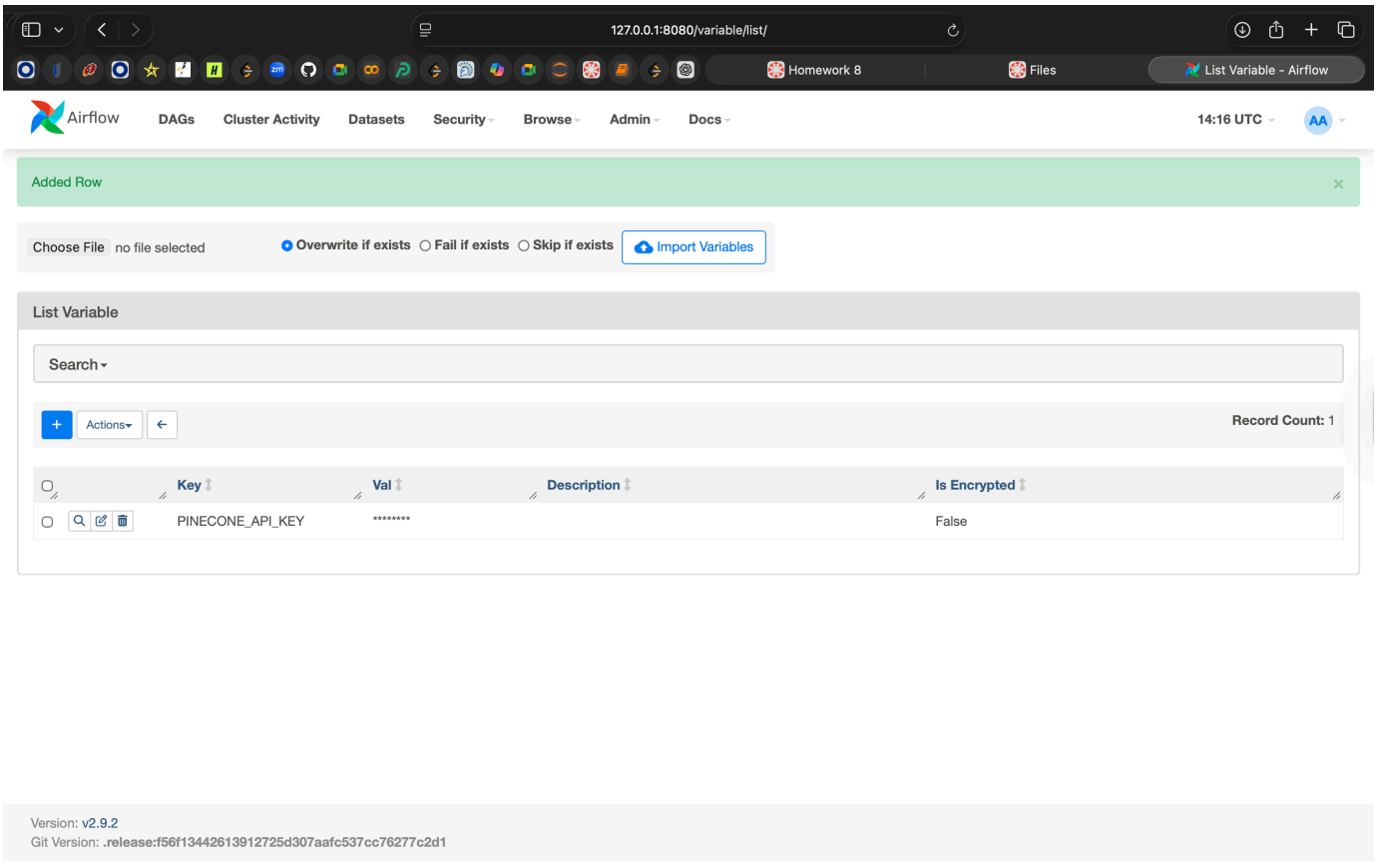
1. Modify your docker-compose.yaml file to include sentence-transformers and pinecone packages (1pt)
    - Restart your docker containers ("docker compose down" and "docker compose up")
    - docker-compose.yaml should be a part of your github submission.

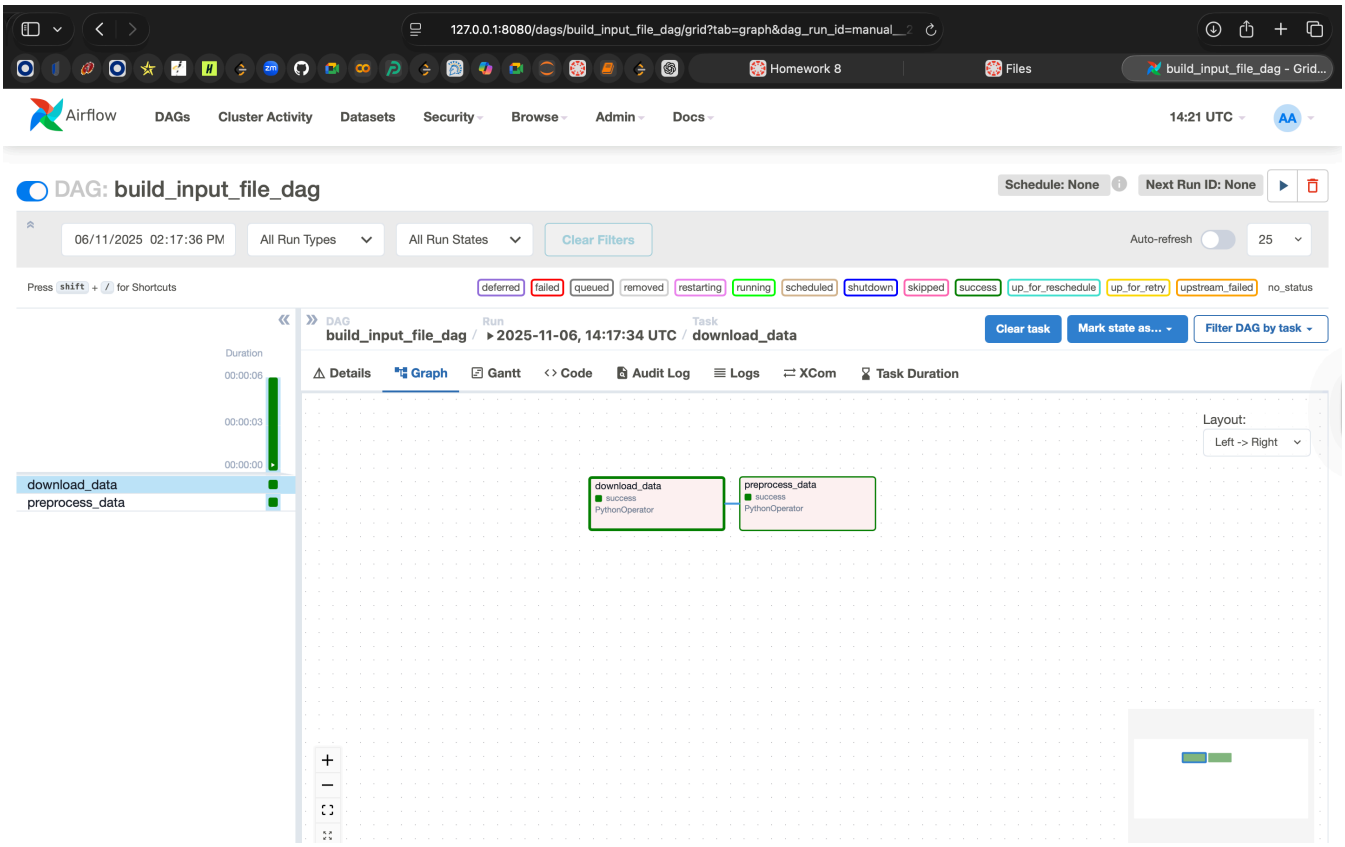
```
sadaf@Sadafs-MacBook-Air ~ % >....
  redis:
    image: redis:7-alpine

  airflow-webserver:
    <<: *airflow-common
    depends_on: [postgres, redis]
    ports: ["8080:8080"]
    command: bash -lc "pip install --no-cache-dir -r /requirements.txt && airflow webserver"

  airflow-scheduler:
    <<: *airflow-common
    depends_on: [airflow-webserver]
    command: bash -lc "pip install --no-cache-dir -r /requirements.txt && airflow scheduler"

  airflow-init:
    <<: *airflow-common
    entrypoint: /bin/bash
    command: -c "airflow db init && airflow users create --username admin --password admin --firstname a --lastname a --role Admin --email admin@example.com"
EOF

sadaf@Sadafs-MacBook-Air ~ % >....
  redis:
    image: redis:7-alpine

  airflow-webserver:
    <<: *airflow-common
    depends_on: [postgres, redis]
    ports: ["8080:8080"]
    command: bash -lc "pip install --no-cache-dir -r /requirements.txt && airflow webserver"

  airflow-scheduler:
    <<: *airflow-common
    depends_on: [airflow-webserver]
    command: bash -lc "pip install --no-cache-dir -r /requirements.txt && airflow scheduler"

  airflow-init:
    <<: *airflow-common
    entrypoint: /bin/bash
    command: -c "airflow db init && airflow users create --username admin --password admin --firstname a --lastname a --role Admin --email admin@example.com"
EOF

sadaf@Sadafs-MacBook-Air ~ % cd ~/pinecone-airflow
sadaf@Sadafs-MacBook-Air pinecone-airflow % docker compose down
WARN[0000] /Users/sadaf/pinecone-airflow/docker-compose.yaml: the attribute `version` is obsolete, it will be ignored, please remove it to avoid potential confusion
Cannot connect to the Docker daemon at unix:///Users/sadaf/.docker/run/docker.sock. Is the docker daemon running?
sadaf@Sadafs-MacBook-Air pinecone-airflow % docker compose up -d --build
WARN[0000] /Users/sadaf/pinecone-airflow/docker-compose.yaml: the attribute `version` is obsolete, it will be ignored, please remove it to avoid potential confusion
[+] Running 33/33
[ ✔ airflow-webserver Pulled                                        101.0s ]
[ ✔ airflow-scheduler Pulled                                        101.0s ]
 ✔ airflow-init Pulled                                              101.0s
 ✔ redis Pulled                                                      55.9s
[+] Running 6/6
 ✔ Network pinecone-airflow_default                 Created           0.0s
 ✔ Container pinecone-airflow-redis-1               Started           0.5s
 ✔ Container pinecone-airflow-postgres-1            Started           0.5s
 ✔ Container pinecone-airflow-airflow-init-1        Started           0.5s
 ✔ Container pinecone-airflow-airflow-webserver-1   Started           0.5s
 ✔ Container pinecone-airflow-airflow-scheduler-1   Started           0.8s
sadaf@Sadafs-MacBook-Air pinecone-airflow % cd ~/pinecone-airflow
```

2. Configure Pinecone (account), get the API token and create Airflow Variable (1pt)


Private
app.pinecone.io/organizations/-OdOKSA5oy7xbmAzn-Xw/projects/f78302f5-8c6a
List Variable - Airflow
API keys | Pinecone Console
Pinecone / Sadaf Fatima's Org / Recommendations / API keys
Docs
Settings
Get help
SFS
Get started
Database
Assistant
Inference
API keys
Manage
API keys
+ API key
Name
Created on
Created by
Value
Permissions
Actions
default
6/11/2025
Sadaf Fatima Syeda
pcsk_2LbyxK_**********
All
STARTER USAGE
RUs 0 / 1M
WUs 0 / 2M
Storage 0GB / 2GB
Upgrade now


127.0.0.1:8080/variable/list/
Homework 8
Files
List Variable - Airflow
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:16 UTC
AA
Added Row
Choose File no file selected
Overwrite if exists
Fail if exists
Skip if exists
Import Variables
List Variable
Search
Actions
Record Count: 1
Key
Val
Description
Is Encrypted
PINECONE_API_KEY
****
False
Version: v2.9.2
Git Version: .release:f56f13442613912725d307aafc537cc76277c2d1

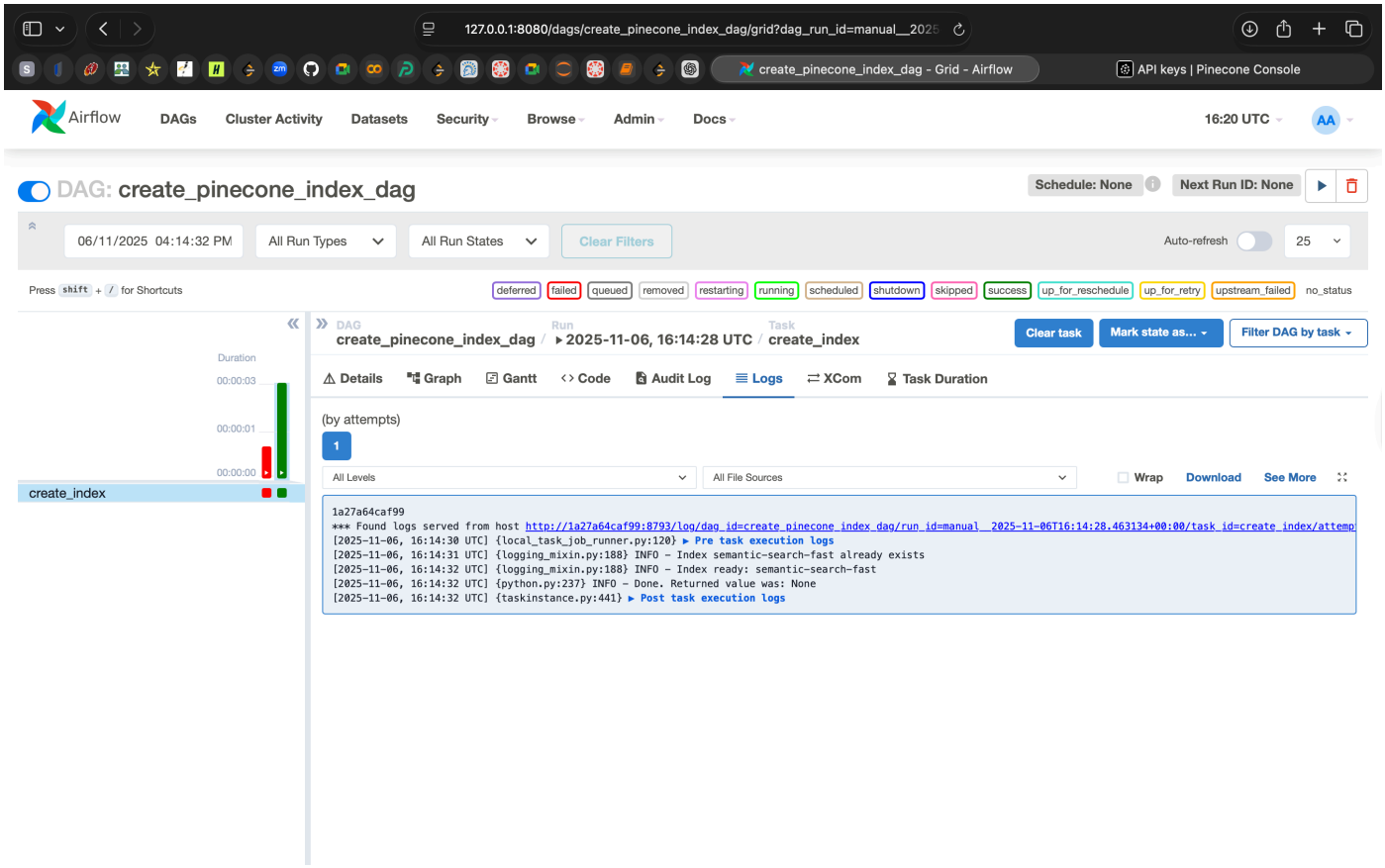
3. Download, Process and Generate an input file to Pinecone (2pt)



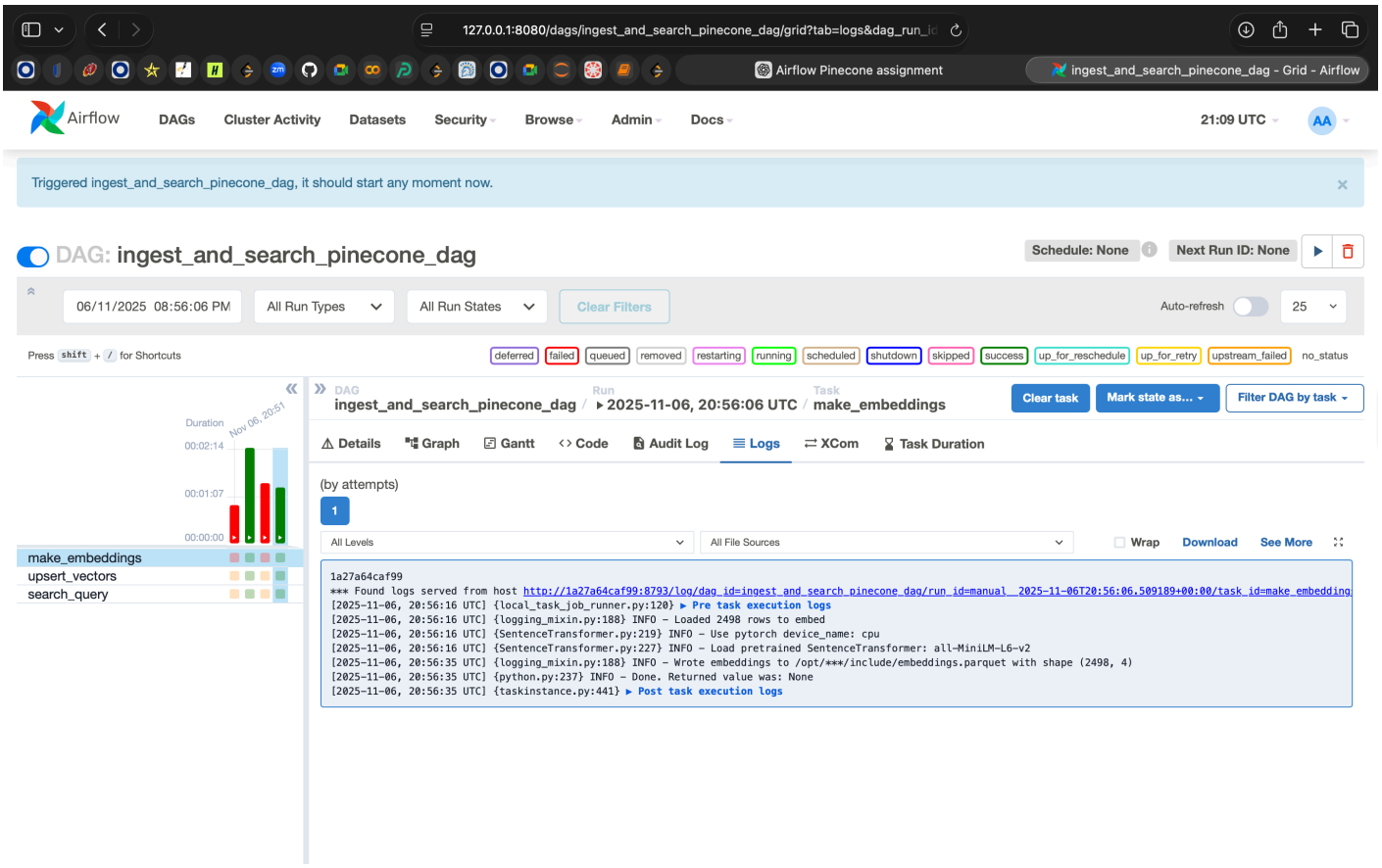

```
025-11-06T14:41:09.517+0000] {taskinstance.py:2648} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='***' AIRFLOW_CTX_DAG_ID='build_input_file_dag' AIRFLOW_CTX_TASK_ID='download_data' AIRFLOW_CTX_
TION_DATE='2025-11-06T14:41:08.226194+00:00' AIRFLOW_CTX_TRY_NUMBER='1' AIRFLOW_CTX_DAG_RUN_ID='manual__2025-11-06T14:41:08.226194+00:00'
025-11-06T14:41:09.517+0000] {taskinstance.py:430} INFO - ::endgroup::
025-11-06T14:41:10.151+0000] {logging_mixin.py:188} INFO - Downloaded to /opt/***/include/medium_data.csv
025-11-06T14:41:10.154+0000] {python.py:237} INFO - Done. Returned value was: None
025-11-06T14:41:10.155+0000] {taskinstance.py:441} INFO - ::group::Post task execution logs
025-11-06T14:41:10.168+0000] {taskinstance.py:1206} INFO - Marking task as SUCCESS. dag_id=build_input_file_dag, task_id=download_data, run_id=manual__2025-11-06T14:41:08.226194+00:00, execution_da
251106T144108, start_date=20251106T144109, end_date=20251106T144110
025-11-06T14:41:10.201+0000] {local_task_job_runner.py:240} INFO - Task exited with return code 0
025-11-06T14:41:10.223+0000] {taskinstance.py:3503} INFO - 1 downstream tasks scheduled from follow-on schedule check
025-11-06T14:41:10.224+0000] {local_task_job_runner.py:222} INFO - ::endgroup::
daf@Sadafs-MacBook-Air pinecone-airflow %
```

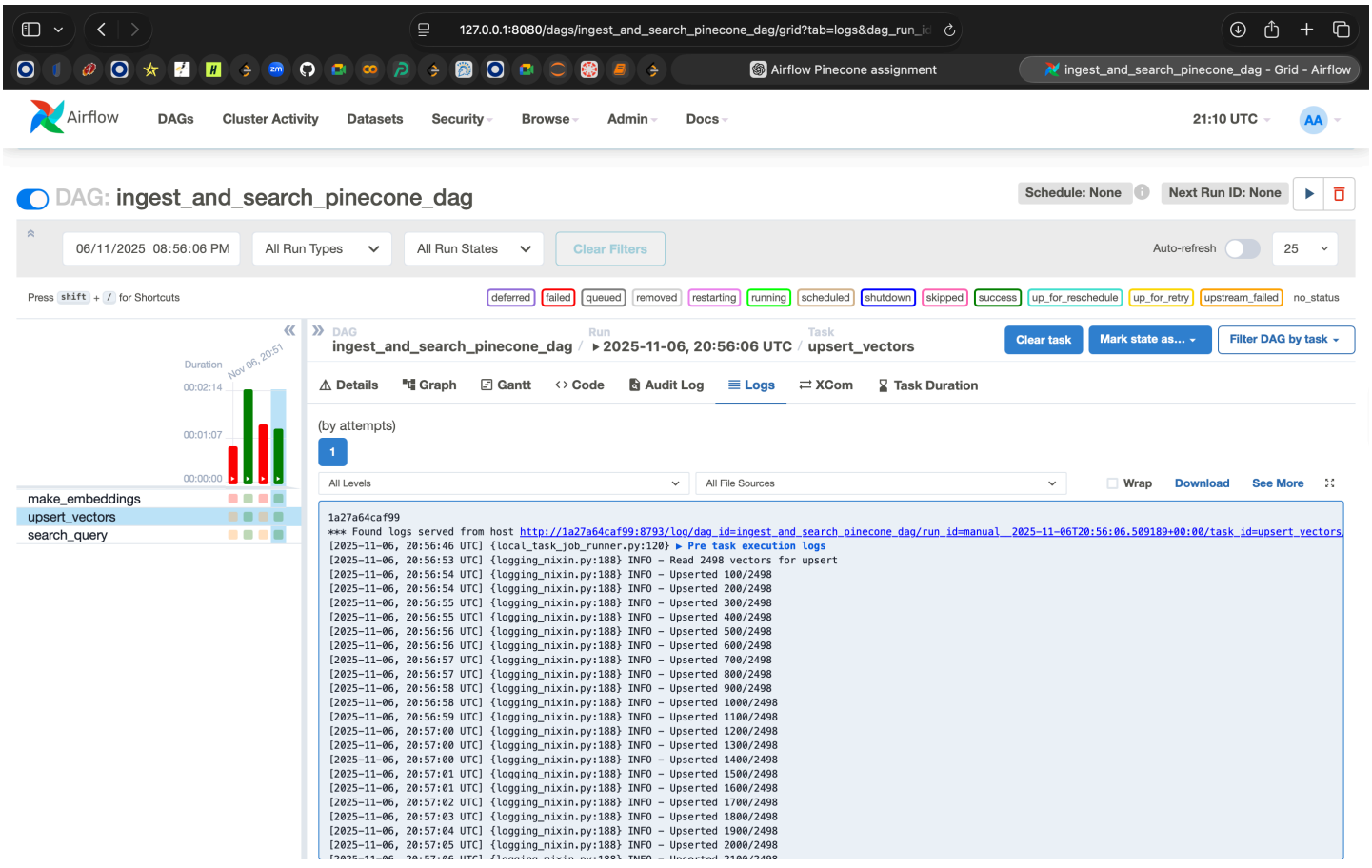
```
025-11-06T14:41:11.327+0000] {logging_mixin.py:188} INFO - Wrote 2498 rows to /opt/***/include/medium_data_prepped.csv
025-11-06T14:41:11.328+0000] {python.py:237} INFO - Done. Returned value was: None
025-11-06T14:41:11.328+0000] {taskinstance.py:441} INFO - ::group::Post task execution logs
025-11-06T14:41:11.334+0000] {taskinstance.py:1206} INFO - Marking task as SUCCESS. dag_id=build_input_file_dag, task_id=preprocess_data, run_id=manual__2025-11-06T14:41:08.226194+00:00, execu
=20251106T144108, start_date=20251106T144111, end_date=20251106T144111
025-11-06T14:41:11.357+0000] {local_task_job_runner.py:240} INFO - Task exited with return code 0
025-11-06T14:41:11.370+0000] {taskinstance.py:3503} INFO - 0 downstream tasks scheduled from follow-on schedule check
025-11-06T14:41:11.371+0000] {local_task_job_runner.py:222} INFO - ::endgroup::
daf@Sadafs-MacBook-Air pinecone-airflow %
```

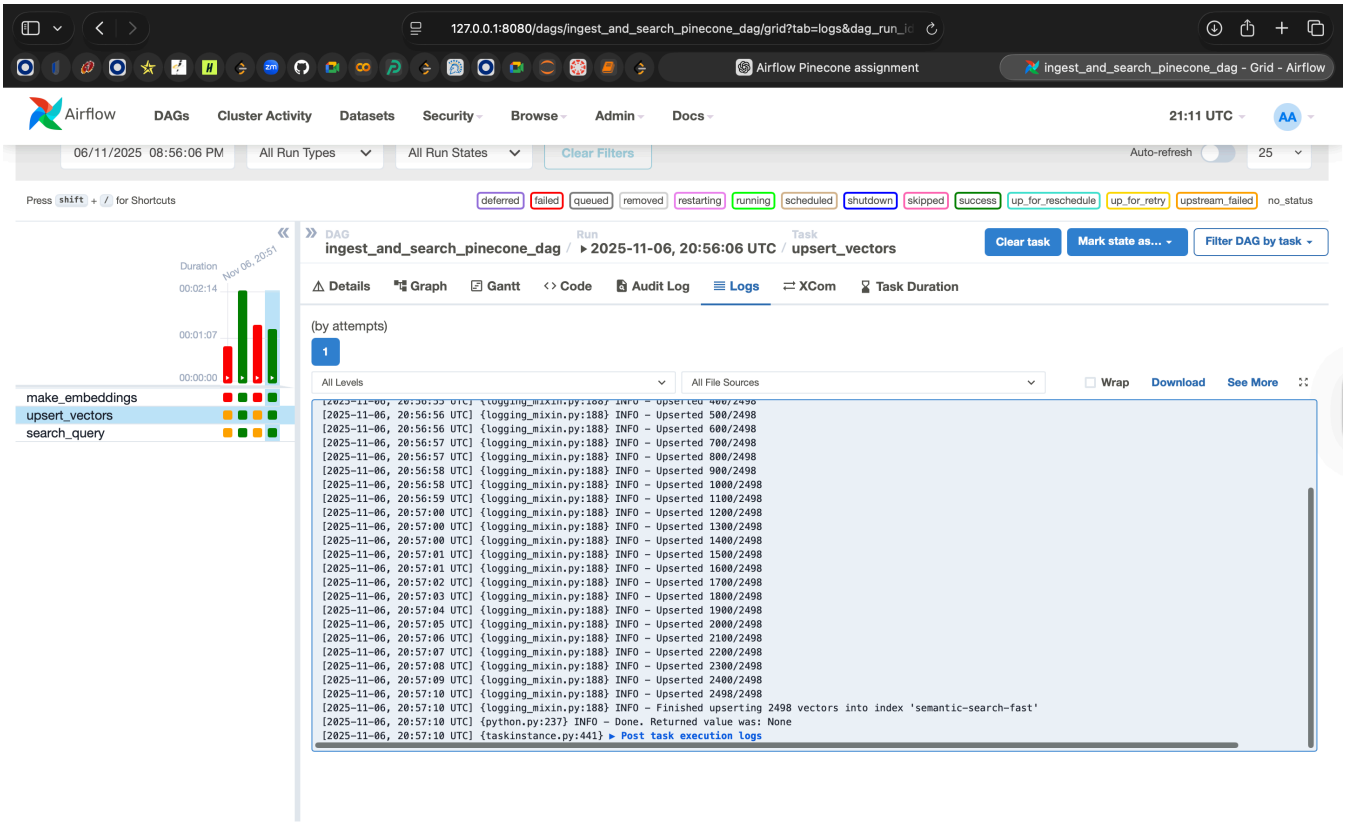
4. Create Pinecone index (1tp)

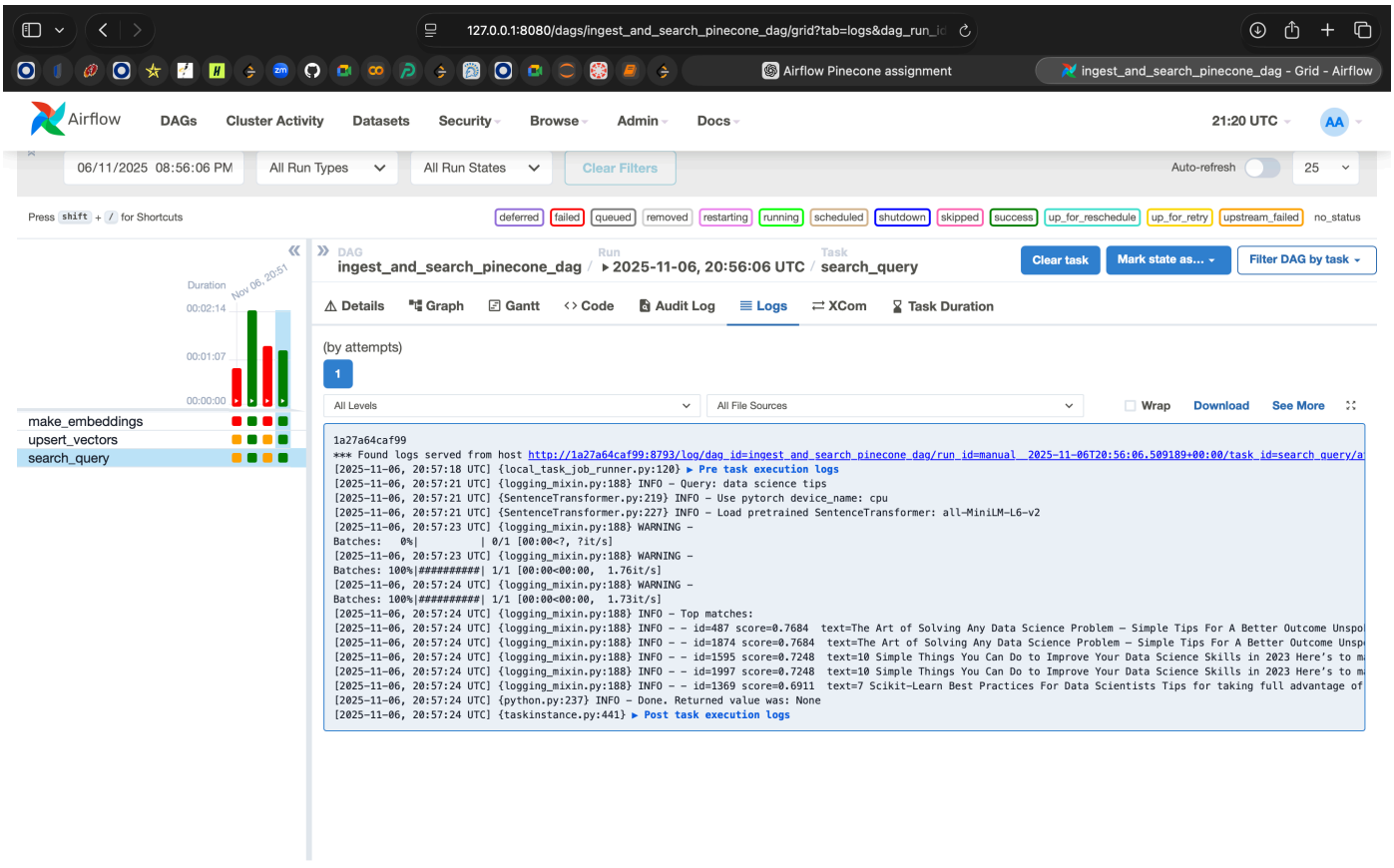


5. Convert the input file into embeddings and ingest them into Pinecone (2pt)

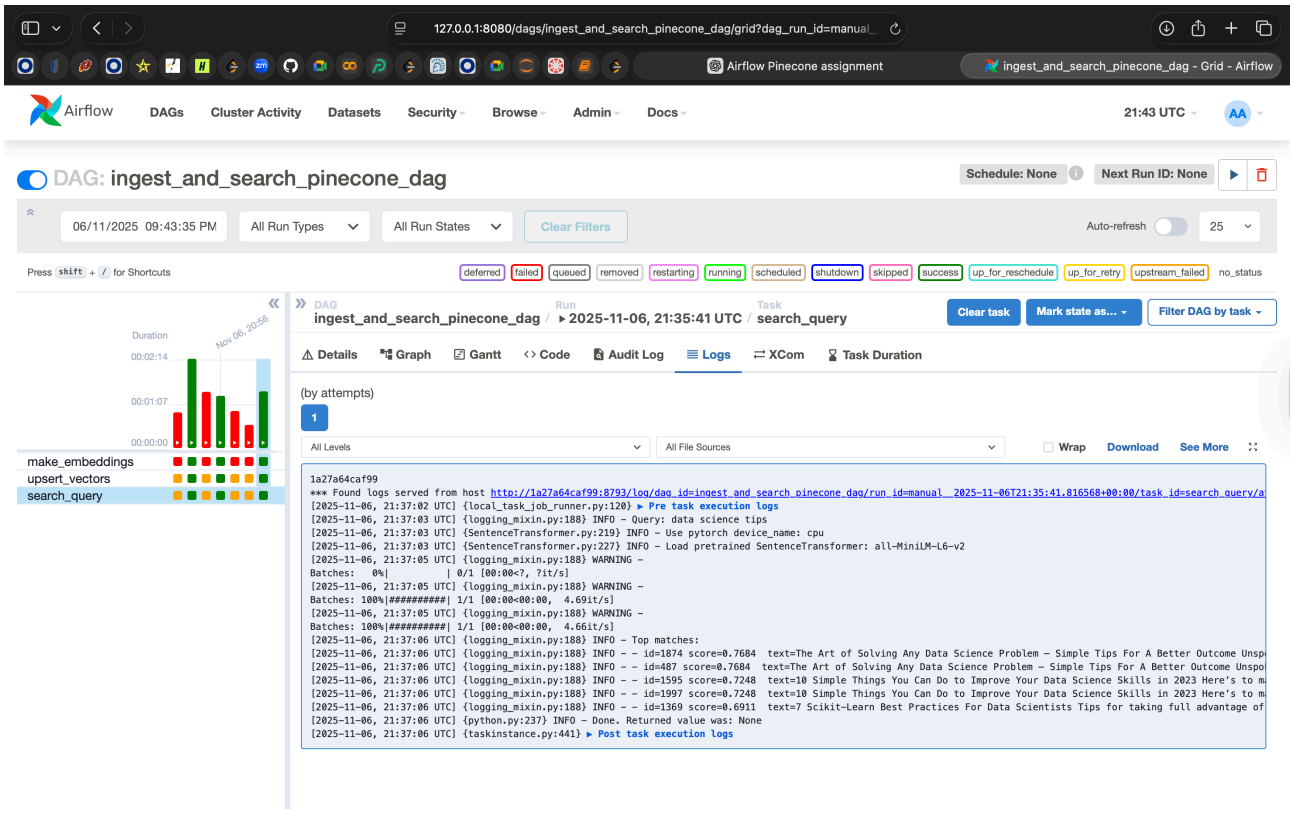

127.0.0.1:8080/dags/ingest_and_search_pinecone_dag/grid?tab=logs&dag_run_id
Airflow Pinecone assignment
ingest_and_search_pinecone_dag - Grid - Airflow
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
21:09 UTC
AA
Triggered ingest_and_search_pinecone_dag, it should start any moment now.
DAG: ingest_and_search_pinecone_dag
Schedule: None
Next Run ID: None
06/11/2025 08:56:06 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Nov 06, 20:51
00:02:14
00:01:07
00:00:00
make_embeddings
upsert_vectors
search_query
DAG
ingest_and_search_pinecone_dag
Run
2025-11-06, 20:56:06 UTC
Task
make_embeddings
Clear task
Mark state as...
Filter DAG by task
Details
Graph
Gantt
Code
Audit Log
Logs
XCom
Task Duration
(by attempts)
1
All Levels
All File Sources
Wrap
Download
See More
1a27a64caf99
*** Found logs served from host http://1a27a64caf99:8793/log/dag_id=ingest_and_search_pinecone_dag/run_id=manual__2025-11-06T20:56:06.509189+00:00/task_id=make_embedding
[2025-11-06, 20:56:16 UTC] {local_task_job_runner.py:120} ▶ Pre task execution logs
[2025-11-06, 20:56:16 UTC] {logging_mixin.py:188} INFO - Loaded 2498 rows to embed
[2025-11-06, 20:56:16 UTC] {SentenceTransformer.py:219} INFO - Use pytorch device_name: cpu
[2025-11-06, 20:56:16 UTC] {SentenceTransformer.py:227} INFO - Load pretrained SentenceTransformer: all-MiniLM-L6-v2
[2025-11-06, 20:56:35 UTC] {logging_mixin.py:188} INFO - Wrote embeddings to /opt/***/include/embeddings.parquet with shape (2498, 4)
[2025-11-06, 20:56:35 UTC] {python.py:237} INFO - Done. Returned value was: None
[2025-11-06, 20:56:35 UTC] {taskinstance.py:441} ▶ Post task execution logs


127.0.0.1:8080/dags/ingest_and_search_pinecone_dag/grid?tab=logs&dag_run_id
Airflow Pinecone assignment
ingest_and_search_pinecone_dag - Grid - Airflow
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
21:10 UTC
AA
DAG: ingest_and_search_pinecone_dag
Schedule: None
Next Run ID: None
06/11/2025 08:56:06 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Nov 06, 20:51
00:02:14
00:01:07
00:00:00
make_embeddings
upsert_vectors
search_query
DAG
ingest_and_search_pinecone_dag
Run
2025-11-06, 20:56:06 UTC
Task
upsert_vectors
Clear task
Mark state as...
Filter DAG by task
Details
Graph
Gantt
Code
Audit Log
Logs
XCom
Task Duration
(by attempts)
1
All Levels
All File Sources
Wrap
Download
See More
1a27a64caf99
*** Found logs served from host http://1a27a64caf99:8793/log/dag_id=ingest_and_search_pinecone_dag/run_id=manual__2025-11-06T20:56:06.509189+00:00/task_id=upsert_vectors/
[2025-11-06, 20:56:46 UTC] {local_task_job_runner.py:120} ▶ Pre task execution logs
[2025-11-06, 20:56:53 UTC] {logging_mixin.py:188} INFO - Read 2498 vectors for upsert
[2025-11-06, 20:56:54 UTC] {logging_mixin.py:188} INFO - Upserted 100/2498
[2025-11-06, 20:56:54 UTC] {logging_mixin.py:188} INFO - Upserted 200/2498
[2025-11-06, 20:56:55 UTC] {logging_mixin.py:188} INFO - Upserted 300/2498
[2025-11-06, 20:56:55 UTC] {logging_mixin.py:188} INFO - Upserted 400/2498
[2025-11-06, 20:56:56 UTC] {logging_mixin.py:188} INFO - Upserted 500/2498
[2025-11-06, 20:56:56 UTC] {logging_mixin.py:188} INFO - Upserted 600/2498
[2025-11-06, 20:56:57 UTC] {logging_mixin.py:188} INFO - Upserted 700/2498
[2025-11-06, 20:56:57 UTC] {logging_mixin.py:188} INFO - Upserted 800/2498
[2025-11-06, 20:56:58 UTC] {logging_mixin.py:188} INFO - Upserted 900/2498
[2025-11-06, 20:56:58 UTC] {logging_mixin.py:188} INFO - Upserted 1000/2498
[2025-11-06, 20:56:59 UTC] {logging_mixin.py:188} INFO - Upserted 1100/2498
[2025-11-06, 20:57:00 UTC] {logging_mixin.py:188} INFO - Upserted 1200/2498
[2025-11-06, 20:57:00 UTC] {logging_mixin.py:188} INFO - Upserted 1300/2498
[2025-11-06, 20:57:00 UTC] {logging_mixin.py:188} INFO - Upserted 1400/2498
[2025-11-06, 20:57:01 UTC] {logging_mixin.py:188} INFO - Upserted 1500/2498
[2025-11-06, 20:57:01 UTC] {logging_mixin.py:188} INFO - Upserted 1600/2498
[2025-11-06, 20:57:02 UTC] {logging_mixin.py:188} INFO - Upserted 1700/2498
[2025-11-06, 20:57:03 UTC] {logging_mixin.py:188} INFO - Upserted 1800/2498
[2025-11-06, 20:57:04 UTC] {logging_mixin.py:188} INFO - Upserted 1900/2498
[2025-11-06, 20:57:05 UTC] {logging_mixin.py:188} INFO - Upserted 2000/2498


127.0.0.1:8080/dags/ingest_and_search_pinecone_dag/grid?tab=logs&dag_run_id
Airflow Pinecone assignment
ingest_and_search_pinecone_dag - Grid - Airflow
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
21:11 UTC
AA
06/11/2025 08:56:06 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Nov 06, 20:51
00:02:14
00:01:07
00:00:00
make_embeddings
upsert_vectors
search_query
DAG
ingest_and_search_pinecone_dag
Run
2025-11-06, 20:56:06 UTC
Task
upsert_vectors
Clear task
Mark state as...
Filter DAG by task
Details
Graph
Gantt
Code
Audit Log
Logs
XCom
Task Duration
(by attempts)
1
All Levels
All File Sources
Wrap
Download
See More
[2025-11-06, 20:56:56 UTC] {logging_mixin.py:188} INFO - Upserted 500/2498
[2025-11-06, 20:56:56 UTC] {logging_mixin.py:188} INFO - Upserted 600/2498
[2025-11-06, 20:56:57 UTC] {logging_mixin.py:188} INFO - Upserted 700/2498
[2025-11-06, 20:56:57 UTC] {logging_mixin.py:188} INFO - Upserted 800/2498
[2025-11-06, 20:56:58 UTC] {logging_mixin.py:188} INFO - Upserted 900/2498
[2025-11-06, 20:56:58 UTC] {logging_mixin.py:188} INFO - Upserted 1000/2498
[2025-11-06, 20:56:59 UTC] {logging_mixin.py:188} INFO - Upserted 1100/2498
[2025-11-06, 20:57:00 UTC] {logging_mixin.py:188} INFO - Upserted 1200/2498
[2025-11-06, 20:57:00 UTC] {logging_mixin.py:188} INFO - Upserted 1300/2498
[2025-11-06, 20:57:00 UTC] {logging_mixin.py:188} INFO - Upserted 1400/2498
[2025-11-06, 20:57:01 UTC] {logging_mixin.py:188} INFO - Upserted 1500/2498
[2025-11-06, 20:57:01 UTC] {logging_mixin.py:188} INFO - Upserted 1600/2498
[2025-11-06, 20:57:02 UTC] {logging_mixin.py:188} INFO - Upserted 1700/2498
[2025-11-06, 20:57:03 UTC] {logging_mixin.py:188} INFO - Upserted 1800/2498
[2025-11-06, 20:57:04 UTC] {logging_mixin.py:188} INFO - Upserted 1900/2498
[2025-11-06, 20:57:05 UTC] {logging_mixin.py:188} INFO - Upserted 2000/2498
[2025-11-06, 20:57:06 UTC] {logging_mixin.py:188} INFO - Upserted 2100/2498
[2025-11-06, 20:57:07 UTC] {logging_mixin.py:188} INFO - Upserted 2200/2498
[2025-11-06, 20:57:08 UTC] {logging_mixin.py:188} INFO - Upserted 2300/2498
[2025-11-06, 20:57:09 UTC] {logging_mixin.py:188} INFO - Upserted 2400/2498
[2025-11-06, 20:57:10 UTC] {logging_mixin.py:188} INFO - Upserted 2498/2498
[2025-11-06, 20:57:10 UTC] {logging_mixin.py:188} INFO - Finished upserting 2498 vectors into index 'semantic-search-fast'
[2025-11-06, 20:57:10 UTC] {python.py:237} INFO - Done. Returned value was: None
[2025-11-06, 20:57:10 UTC] {taskinstance.py:441} ▶ Post task execution logs


127.0.0.1:8080/dags/ingest_and_search_pinecone_dag/grid?tab=logs&dag_run_id
Airflow Pinecone assignment
ingest_and_search_pinecone_dag - Grid - Airflow
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
21:20 UTC
AA
06/11/2025 08:56:06 PM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Nov 06, 20:51
00:02:14
00:01:07
00:00:00
make_embeddings
upsert_vectors
search_query
DAG
ingest_and_search_pinecone_dag
Run
2025-11-06, 20:56:06 UTC
Task
search_query
Clear task
Mark state as...
Filter DAG by task
Details
Graph
Gantt
Code
Audit Log
Logs
XCom
Task Duration
(by attempts)
1
All Levels
All File Sources
Wrap
Download
See More
1a27a64caf99
*** Found logs served from host http://1a27a64caf99:8793/log/dag_id=ingest_and_search_pinecone_dag/run_id=manual__2025-11-06T20:56:06.509189+00:00/task_id=search_query/a
[2025-11-06, 20:57:18 UTC] {local_task_job_runner.py:120} ▶ Pre task execution logs
[2025-11-06, 20:57:21 UTC] {logging_mixin.py:188} INFO - Query: data science tips
[2025-11-06, 20:57:21 UTC] {SentenceTransformer.py:219} INFO - Use pytorch device_name: cpu
[2025-11-06, 20:57:21 UTC] {SentenceTransformer.py:227} INFO - Load pretrained SentenceTransformer: all-MiniLM-L6-v2
[2025-11-06, 20:57:23 UTC] {logging_mixin.py:188} WARNING -
Batches: 0%| | 0/1 [00:00<?, ?it/s]
[2025-11-06, 20:57:23 UTC] {logging_mixin.py:188} WARNING -
Batches: 100%|##########| 1/1 [00:00<00:00, 1.76it/s]
[2025-11-06, 20:57:24 UTC] {logging_mixin.py:188} WARNING -
Batches: 100%|##########| 1/1 [00:00<00:00, 1.73it/s]
[2025-11-06, 20:57:24 UTC] {logging_mixin.py:188} INFO - Top matches:
[2025-11-06, 20:57:24 UTC] {logging_mixin.py:188} INFO - - id=487 score=0.7684 text=The Art of Solving Any Data Science Problem — Simple Tips For A Better Outcome Unspo
[2025-11-06, 20:57:24 UTC] {logging_mixin.py:188} INFO - - id=1874 score=0.7684 text=The Art of Solving Any Data Science Problem — Simple Tips For A Better Outcome Unsp
[2025-11-06, 20:57:24 UTC] {logging_mixin.py:188} INFO - - id=1595 score=0.7248 text=10 Simple Things You Can Do to Improve Your Data Science Skills in 2023 Here's to m
[2025-11-06, 20:57:24 UTC] {logging_mixin.py:188} INFO - - id=1997 score=0.7248 text=10 Simple Things You Can Do to Improve Your Data Science Skills in 2023 Here's to m
[2025-11-06, 20:57:24 UTC] {logging_mixin.py:188} INFO - - id=1369 score=0.6911 text=7 Scikit-Learn Best Practices For Data Scientists Tips for taking full advantage of
[2025-11-06, 20:57:24 UTC] {python.py:237} INFO - Done. Returned value was: None
[2025-11-06, 20:57:24 UTC] {taskinstance.py:441} ▶ Post task execution logs

6. Run search against Pinecone (1pt)



GITHUB REPO :
https://github.com/sadaffatimaee/pinecone-airflow