

ASSIGNMENT – 8

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

```
docker-compose.yaml
47
54
59   RE_PAUSED_AT_CREATION: 'true'
60   XAMPLES: 'false'
61   CKENDS: 'airflow.api.auth.backend.basic_auth,airflow.api.auth.backend.session'
62   le:line-length
63   ver on scheduler for health checks
64   l.apache.org/docs/apache-airflow/stable/administration-and-deployment/logging-monitoring/check-health.html#scheduler-health-check-server
65   e:line-length
66   NABLE_HEALTH_CHECK: 'true'
67   DDITIONAL_REQUIREMENTS option ONLY for a quick checks
68   development, test and especially production usage) build/extend Airflow image.
69   REMENTS: ${_PIP_ADDITIONAL_REQUIREMENTS:- yfinance apache-airflow-providers-snowflake snowflake-connector-python sentence-transformers pinecone}
70   can be used to set a custom config file, stored in the local config folder
71   it, uncomment it and replace airflow.cfg with the name of your config file
72   /airflow/config/airflow.cfg'
73
74   - ./dags:/opt/airflow/dags
75   - ./logs:/opt/airflow/logs
76   - ./config:/opt/airflow/config
77   - ./plugins:/opt/airflow/plugins
78   ports:
79     - 8080->8080
```

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

Edit Variable

Key *

pinecone_api_key

Val

Description

Description

Save

<

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

The screenshot displays the Airflow DAG interface for the 'Medium_to_Pinecone' DAG. The task 'download_data' is selected, and its logs are visible. The logs show the following content:

```
1b0476403e9b
*** Found local files:
*** * /opt/airflow/logs/dag_id=Medium_to_Pinecone/run_id=manual__2025-04-17T16:12:52.908445+00:00/task_id=download_data/attempt-1.log
[2025-04-17, 16:13:12 UTC] [local_task_job_runner.py:123] ▶ Pre task execution logs
[2025-04-17, 16:13:14 UTC] [logging_mixin.py:190] INFO - Downloaded file has 2499 lines
[2025-04-17, 16:13:14 UTC] [python.py:240] INFO - Done. Returned value was: /tmp/medium_data/medium_data.csv
[2025-04-17, 16:13:14 UTC] [taskinstance.py:340] ▶ Post task execution logs
```

The DAG graph on the left shows the sequence of tasks: download_data, preprocess_data, create_pinecone_index, generate_embeddings_and_upsert, and test_search_query. The 'download_data' task is currently running.

4. Create Pinecone index (1tp)

The screenshot displays the Airflow DAG interface for the 'Medium_to_Pinecone' DAG. The task 'create_pinecone_index' is selected, and its logs are visible. The logs show the following content:

```
1b0476403e9b
*** Found local files:
*** * /opt/airflow/logs/dag_id=Medium_to_Pinecone/run_id=manual__2025-04-17T16:12:52.908445+00:00/task_id=create_pinecone_index/attempt-1.log
[2025-04-17, 16:13:12 UTC] [local_task_job_runner.py:123] ▶ Pre task execution logs
[2025-04-17, 16:13:21 UTC] [logging_mixin.py:190] INFO - Pinecone index 'semantic-search-fast' created successfully
[2025-04-17, 16:13:21 UTC] [python.py:240] INFO - Done. Returned value was: semantic-search-fast
[2025-04-17, 16:13:21 UTC] [taskinstance.py:340] ▶ Post task execution logs
```

The DAG graph on the left shows the sequence of tasks: download_data, preprocess_data, create_pinecone_index, generate_embeddings_and_upsert, and test_search_query. The 'create_pinecone_index' task is currently running.

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

The screenshot shows the Airflow web interface with the 'Logs' tab selected for the 'test_search_query' task. The logs display the execution of the 'generate_embeddings_and_upsert' task, which successfully ingested 2498 records into Pinecone. The logs include timestamps, task IDs, and the output of the task.

```
2025-04-17, 16:17:18 UTC] [logging_mixin.py:190] WARNING -  
Batches: 100%##### 4/4 [00:05:00:00, 1.35s/it]  
[2025-04-17, 16:17:22 UTC] [logging_mixin.py:190] INFO - Processing batch 24/25  
[2025-04-17, 16:17:22 UTC] [logging_mixin.py:190] WARNING -  
Batches: 0% | 0/4 [00:00:00:00, 1.07s/it]  
[2025-04-17, 16:17:24 UTC] [logging_mixin.py:190] WARNING -  
Batches: 25%### 1/4 [00:01:00:00, 1.07s/it]  
[2025-04-17, 16:17:25 UTC] [logging_mixin.py:190] WARNING -  
Batches: 50%#### 2/4 [00:02:00:00, 1.45s/it]  
[2025-04-17, 16:17:26 UTC] [logging_mixin.py:190] WARNING -  
Batches: 75%##### 3/4 [00:04:00:00, 1.31s/it]  
[2025-04-17, 16:17:26 UTC] [logging_mixin.py:190] WARNING -  
Batches: 100%##### 4/4 [00:04:00:00, 1.17s/it]  
[2025-04-17, 16:17:26 UTC] [logging_mixin.py:190] WARNING -  
Batches: 100%##### 4/4 [00:04:00:00, 1.07s/it]  
[2025-04-17, 16:17:29 UTC] [logging_mixin.py:190] INFO - Processing batch 25/25  
[2025-04-17, 16:17:29 UTC] [logging_mixin.py:190] WARNING -  
Batches: 0% | 0/4 [00:00:00:00, 1.71s/it]  
[2025-04-17, 16:17:32 UTC] [logging_mixin.py:190] WARNING -  
Batches: 25%### 1/4 [00:02:00:00, 2.72s/it]  
[2025-04-17, 16:17:33 UTC] [logging_mixin.py:190] WARNING -  
Batches: 50%#### 2/4 [00:04:00:00, 2.10s/it]  
[2025-04-17, 16:17:35 UTC] [logging_mixin.py:190] WARNING -  
Batches: 75%##### 3/4 [00:05:00:00, 1.77s/it]  
[2025-04-17, 16:17:35 UTC] [logging_mixin.py:190] WARNING -  
Batches: 100%##### 4/4 [00:05:00:00, 1.12s/it]  
[2025-04-17, 16:17:35 UTC] [logging_mixin.py:190] WARNING -  
Batches: 100%##### 4/4 [00:05:00:00, 1.48s/it]  
[2025-04-17, 16:17:38 UTC] [logging_mixin.py:190] INFO - Successfully upserted 2498 records to Pinecone  
[2025-04-17, 16:17:38 UTC] [python.py:240] INFO - Done. Returned value was: semantic-search-fast  
[2025-04-17, 16:17:38 UTC] [taskinstance.py:340] Post task execution logs
```

6. Run search against Pinecone (1pt)

The screenshot shows the Airflow web interface with the 'Logs' tab selected for the 'test_search_query' task. The logs display the execution of the 'test_search_query' task, which successfully searched for 'what is ethics in AI' and returned results. The logs include timestamps, task IDs, and the output of the task.

```
1b0476403e9b  
***  
Found local files:  
***  
/opt/airflow/logs/dag_id=Medium_to_Pinecone/run_id=manual_2025-04-17:16:12:52.90845+00:00/task_id=test_search_query/attempt=1.log  
[2025-04-17, 16:18:04 UTC] [local_task_job_runner.py:122] Pre task execution logs  
[2025-04-17, 16:18:05 UTC] [SentenceTransformer.py:219] INFO - Load pretrained SentenceTransformer: all-MiniLM-L6-v2  
[2025-04-17, 16:18:09 UTC] [logging_mixin.py:190] WARNING -  
Batches: 0% | 0/1 [00:00:00:00, 7.23s/it]  
[2025-04-17, 16:18:09 UTC] [logging_mixin.py:190] WARNING -  
Batches: 100%##### 1/1 [00:00:00:00, 7.23s/it]  
[2025-04-17, 16:18:09 UTC] [logging_mixin.py:190] WARNING -  
Batches: 100%##### 1/1 [00:00:00:00, 7.13s/it]  
[2025-04-17, 16:18:11 UTC] [logging_mixin.py:190] INFO - Search results for query: 'what is ethics in AI'  
[2025-04-17, 16:18:11 UTC] [logging_mixin.py:190] INFO - ID: 1, Score: 0.748038627, Title: Ethics in AI: Potential Root Causes for Biased Alg...  
[2025-04-17, 16:18:11 UTC] [logging_mixin.py:190] INFO - ID: 1634, Score: 0.748038627, Title: Ethics in AI: Potential Root Causes for Biased Alg...  
[2025-04-17, 16:18:11 UTC] [logging_mixin.py:190] INFO - ID: 1326, Score: 0.718646314, Title: The ethical implications of AI in design It's time...  
[2025-04-17, 16:18:11 UTC] [logging_mixin.py:190] INFO - ID: 661, Score: 0.66682396, Title: Ethical Considerations In Machine Learning Project...  
[2025-04-17, 16:18:11 UTC] [logging_mixin.py:190] INFO - ID: 1627, Score: 0.651985758, Title: Navigating the Ethical Contours of AI Copy Generat...  
[2025-04-17, 16:18:11 UTC] [python.py:240] INFO - Done. Returned value was: None  
[2025-04-17, 16:18:11 UTC] [taskinstance.py:340] Post task execution logs
```

