Sharan Patil
(018327049)

Instructor: Prof. Keeyong Han

Homework 8: Running Pinecone Job as an Airflow Pipeline

GIT- https://github.com/sharan9219790/HW8_DATAWAREHOUSE

Introduction

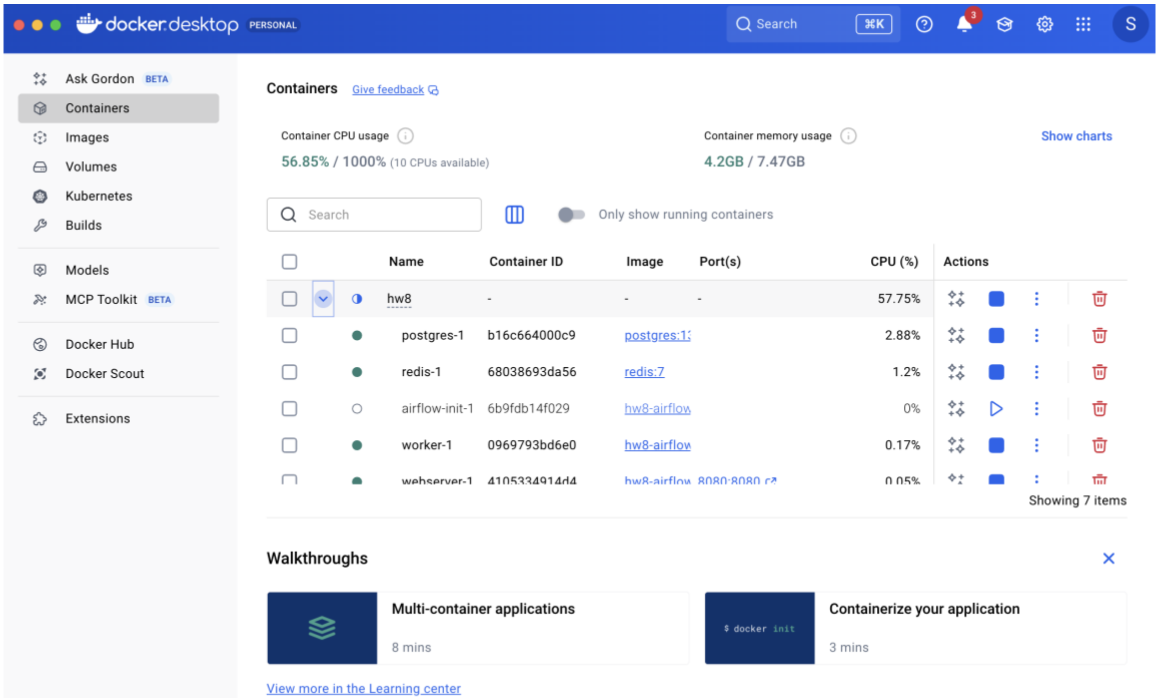
This report documents the complete workflow and implementation steps carried out for Homework 8 of the DATA 226 course. The task involved setting up an Apache Airflow pipeline integrated with Pinecone to perform end-to-end semantic search operations. The pipeline automated the process of downloading a dataset, preprocessing it, creating a Pinecone index, converting text into embeddings, and running a semantic search query against the index. The setup was deployed using Docker and verified through successful Airflow task executions.

1. Environment Setup

The first step was to set up the development environment using Docker Compose. We created a complete Airflow environment that included the webserver, scheduler, worker, Redis, and Postgres containers. A custom Dockerfile was written to pre-install the required Python libraries (sentence-transformers and pinecone), ensuring that Airflow could seamlessly run tasks involving these dependencies without restart errors. The docker-compose file was configured to expose Airflow on port 8080, map the required volumes, and initialize the Airflow database with an admin account.

We also created an .env file to store environment variables such as AIRFLOW_UID and AIRFLOW_GID to ensure proper file permissions on macOS. Once everything was configured, the setup was initialized and verified using the commands:

```
docker compose build --no-cache
docker compose up airflow-init
docker compose up -d
```



Containers

Container CPU usage
56.85% / 1000% (10 CPUs available)

Container memory usage
4.2GB / 7.47GB

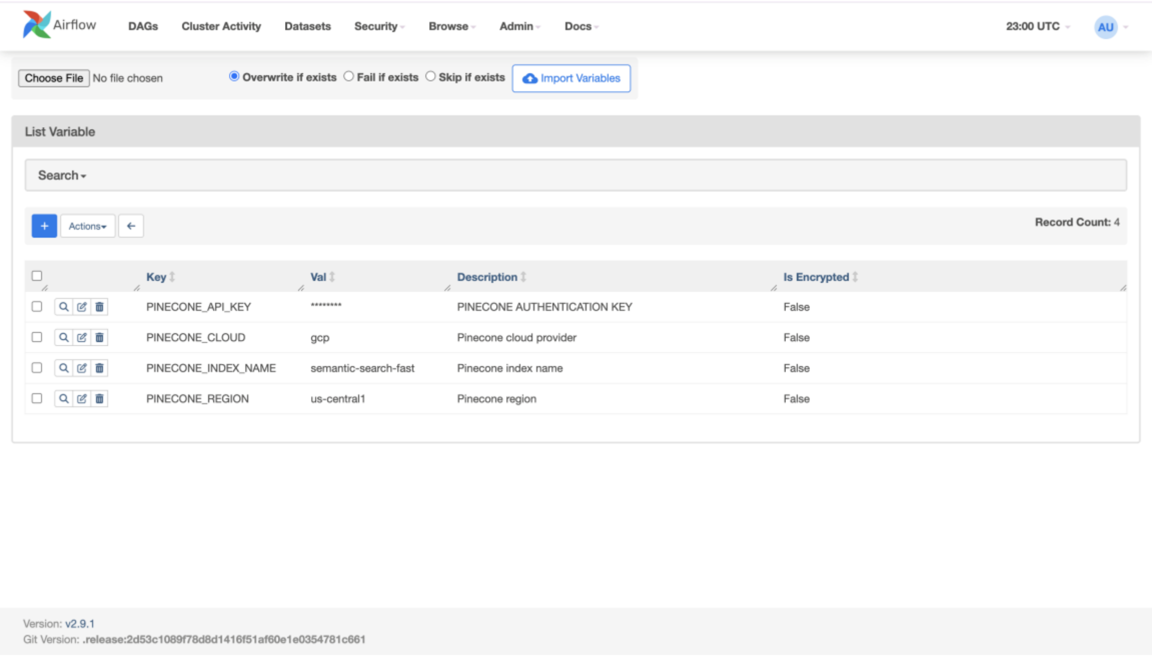
□			Name	Container ID	Image	Port(s)	CPU (%)	Actions
□		◑	hw8	-	-	-	57.75%	
□		●	postgres-1	b16c664000c9	postgres:1:		2.88%	
□		●	redis-1	68038693da56	redis:7		1.2%	
□		○	airflow-init-1	6b9fdb14f029	hw8-airflow		0%	
□		●	worker-1	0969793bd6e0	hw8-airflow		0.17%	
□		●	webserver-1	4105334914d4	hw8-airflow	8080:8080	0.05%	

Showing 7 items

2. Airflow and Pinecone Configuration

Once the environment was live, we accessed the Airflow UI via http://localhost:8080 and created four Airflow Variables under Admin → Variables. These variables stored Pinecone credentials and configuration details:
- PINECONE_API_KEY
- PINECONE_INDEX_NAME
- PINECONE_CLOUD
- PINECONE_REGION

This approach allowed us to keep sensitive API credentials secure while making them accessible to all Airflow tasks via the Variable.get() method.
A free Pinecone account was used, and when the default region was unsupported, we updated the variables to use gcp/us-central1,
which is compatible with free-tier accounts.



No file chosen

● Overwrite if exists ○ Fail if exists ○ Skip if exists

List Variable

Record Count: 4

□	Key	Val	Description	Is Encrypted
□	PINECONE_API_KEY	********	PINECONE AUTHENTICATION KEY	False
□	PINECONE_CLOUD	gcp	Pinecone cloud provider	False
□	PINECONE_INDEX_NAME	semantic-search-fast	Pinecone index name	False
□	PINECONE_REGION	us-central1	Pinecone region	False

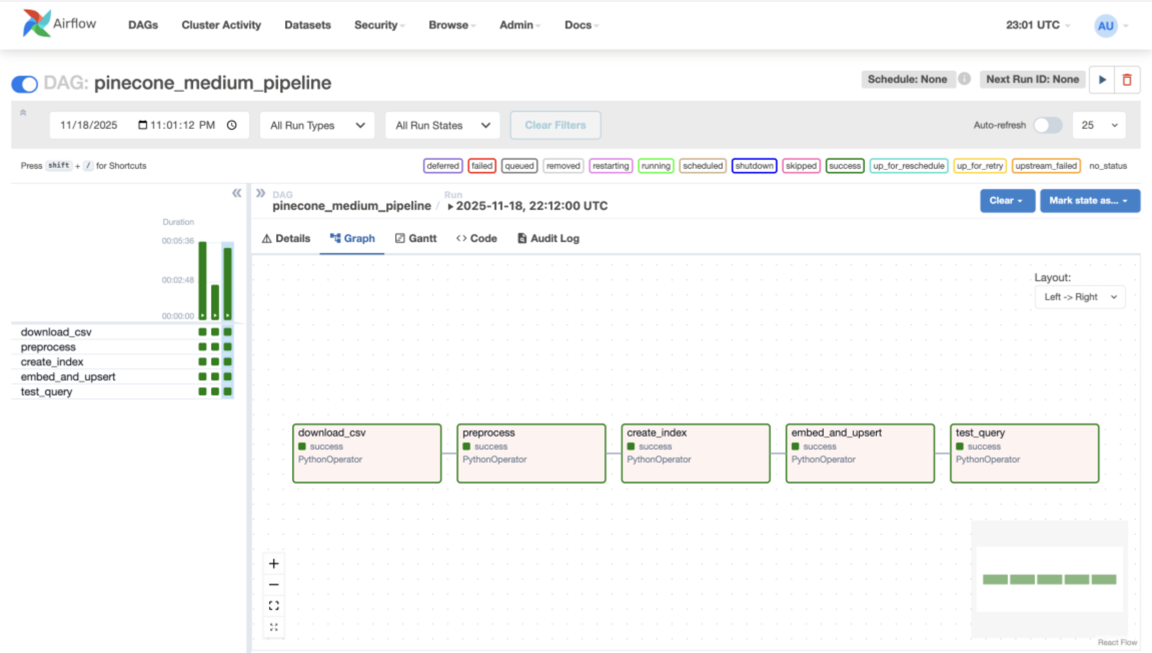
Version: v2.9.1
Git Version: .release:2d53c1089f78d8d1416f51af60e1e0354781c661

3. DAG Creation and Task Implementation

The Airflow DAG file pinecone_medium_pipeline.py was created to define the workflow. The DAG contained five PythonOperator tasks,
each performing a specific stage of the process:

• download_csv → Downloaded the dataset from a public URL (Medium articles dataset).
• preprocess → Processed the CSV file by cleaning, merging title and subtitle, and generating a metadata column.
• create_index → Created a Pinecone serverless index (dimension 384, dotproduct metric). It automatically retried with alternate
regions if the user's free plan didn't support the default region.
• embed_and_upsert → Loaded the SentenceTransformer model (all-MiniLM-L6-v2), generated text embeddings, and uploaded
them to Pinecone using upsert_from_dataframe().
• test_query → Queried Pinecone with a sample semantic search ("what is ethics in AI") and logged the top results.

The DAG was defined with a retry mechanism, two-minute retry delay, and a clear dependency chain ensuring linear execution.
Each task was implemented as a standalone Python function with proper logging for debugging and monitoring.



DAG: pinecone_medium_pipeline

Schedule: None

Next Run ID: None

11/18/2025 11:01:12 PM

All Run Types

All Run States

Auto-refresh 25

download_csv
preprocess
create_index
embed_and_upsert
test_query

DAG pinecone_medium_pipeline / Run 2025-11-18, 22:12:00 UTC

Layout: Left -> Right

download_csv success PythonOperator

preprocess success PythonOperator

create_index success PythonOperator

embed_and_upsert success PythonOperator

test_query success PythonOperator

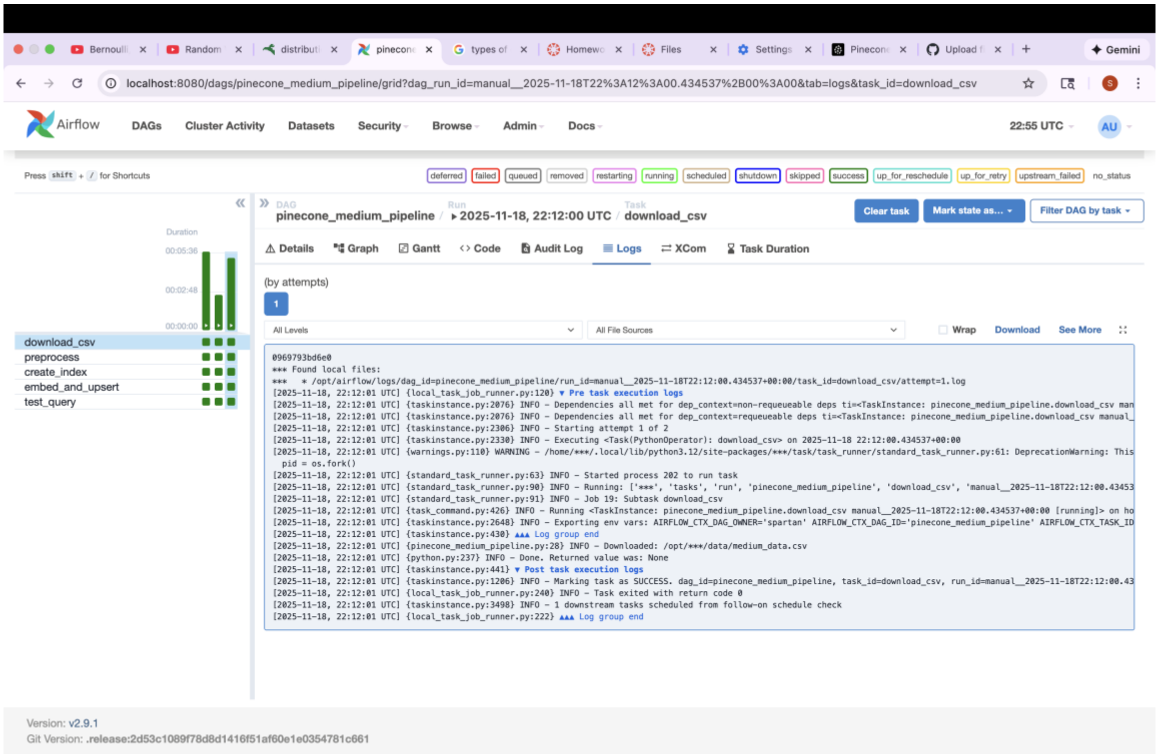
4. Execution and Results

After configuring everything, we triggered the DAG manually from the Airflow UI. Each task executed successfully in sequence.
Initially, we encountered an error when creating the Pinecone index due to unsupported region configuration, which was resolved
by modifying the region to us-central1. Once the pipeline was rerun, all tasks turned green (success).

The logs from the test_query task confirmed that the embeddings and index were functioning correctly. The log entries showed
top-ranked results for the semantic search query, validating that the pipeline successfully encoded, ingested, and queried data
from Pinecone. The final DAG graph view showed all tasks completed successfully.

LOGS SCREENSHOT FOR EVERY DAG TASK:

(Next page)



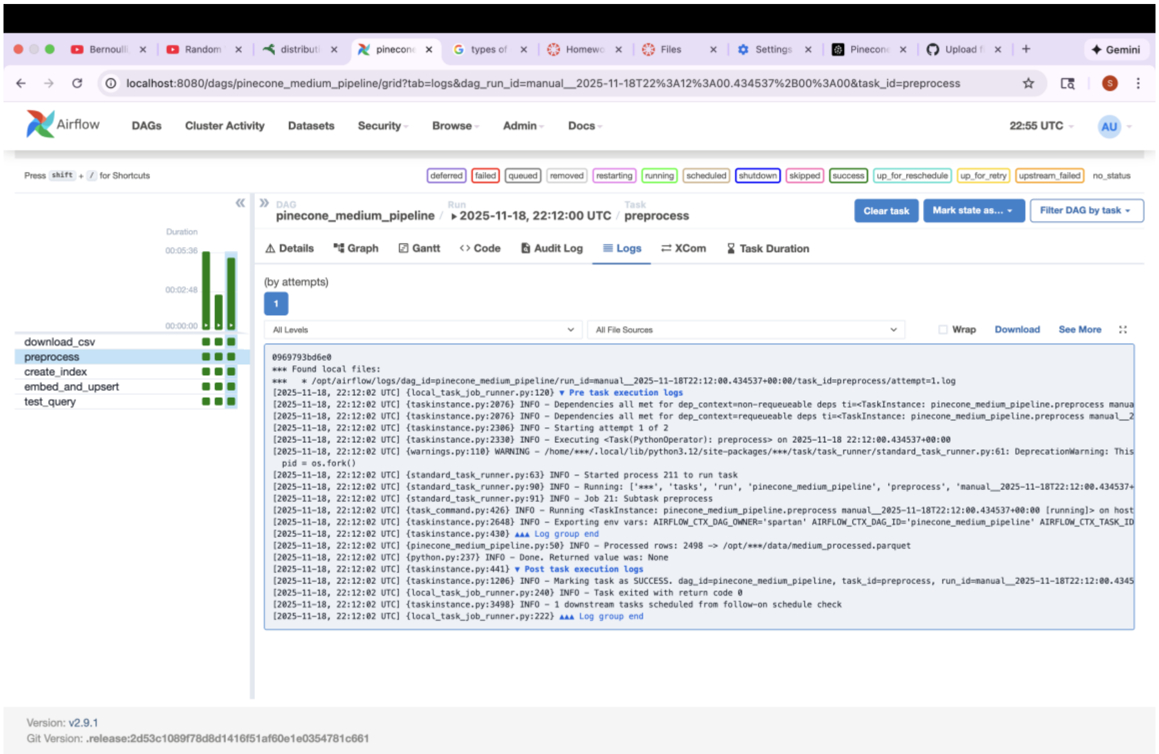
localhost:8080/dags/pinecone_medium_pipeline/grid?dag_run_id=manual__2025-11-18T22%3A12%3A00.434537%2B00%3A00&tab=logs&task_id=download_csv

DAG pinecone_medium_pipeline / Run 2025-11-18, 22:12:00 UTC / Task download_csv

- download_csv
- preprocess
- create_index
- embed_and_upsert
- test_query

```
0969793bd6e0
*** Found local files:
***   * /opt/airflow/logs/dag_id=pinecone_medium_pipeline/run_id=manual__2025-11-18T22:12:00.434537+00:00/task_id=download_csv/attempt=1.log
[2025-11-18, 22:12:01 UTC] {local_task_job_runner.py:120} ▼ Pre task execution logs
[2025-11-18, 22:12:01 UTC] {taskinstance.py:2076} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: pinecone_medium_pipeline.download_csv man
[2025-11-18, 22:12:01 UTC] {taskinstance.py:2076} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: pinecone_medium_pipeline.download_csv manual_
[2025-11-18, 22:12:01 UTC] {taskinstance.py:2306} INFO - Starting attempt 1 of 2
[2025-11-18, 22:12:01 UTC] {taskinstance.py:2330} INFO - Executing <Task(PythonOperator): download_csv> on 2025-11-18 22:12:00.434537+00:00
[2025-11-18, 22:12:01 UTC] {warnings.py:110} WARNING - /home/***/.local/lib/python3.12/site-packages/***/task/task_runner/standard_task_runner.py:61: DeprecationWarning: This
  pid = os.fork()
[2025-11-18, 22:12:01 UTC] {standard_task_runner.py:63} INFO - Started process 202 to run task
[2025-11-18, 22:12:01 UTC] {standard_task_runner.py:90} INFO - Running: ['***', 'tasks', 'run', 'pinecone_medium_pipeline', 'download_csv', 'manual__2025-11-18T22:12:00.43453
[2025-11-18, 22:12:01 UTC] {standard_task_runner.py:91} INFO - Job 19: Subtask download_csv
[2025-11-18, 22:12:01 UTC] {task_command.py:426} INFO - Running <TaskInstance: pinecone_medium_pipeline.download_csv manual__2025-11-18T22:12:00.434537+00:00 [running]> on ho
[2025-11-18, 22:12:01 UTC] {taskinstance.py:2648} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='spartan' AIRFLOW_CTX_DAG_ID='pinecone_medium_pipeline' AIRFLOW_CTX_TASK_ID
[2025-11-18, 22:12:01 UTC] {taskinstance.py:430} ▲▲▲ Log group end
[2025-11-18, 22:12:01 UTC] {pinecone_medium_pipeline.py:28} INFO - Downloaded: /opt/***/data/medium_data.csv
[2025-11-18, 22:12:01 UTC] {python.py:237} INFO - Done. Returned value was: None
[2025-11-18, 22:12:01 UTC] {taskinstance.py:441} ▼ Post task execution logs
[2025-11-18, 22:12:01 UTC] {taskinstance.py:1206} INFO - Marking task as SUCCESS. dag_id=pinecone_medium_pipeline, task_id=download_csv, run_id=manual__2025-11-18T22:12:00.43
[2025-11-18, 22:12:01 UTC] {local_task_job_runner.py:240} INFO - Task exited with return code 0
[2025-11-18, 22:12:01 UTC] {taskinstance.py:3498} INFO - 1 downstream tasks scheduled from follow-on schedule check
[2025-11-18, 22:12:01 UTC] {local_task_job_runner.py:222} ▲▲▲ Log group end
```

Version: v2.9.1
Git Version: .release:2d53c1089f78d8d1416f51af60e1e0354781c661



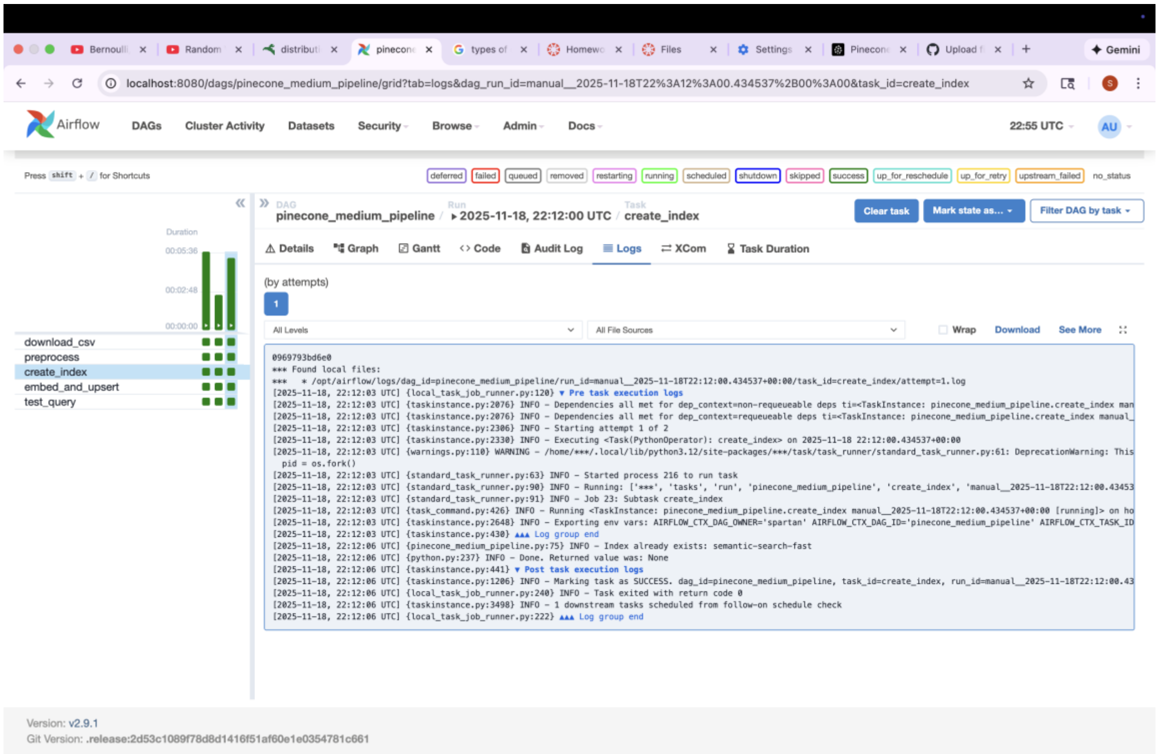
localhost:8080/dags/pinecone_medium_pipeline/grid?tab=logs&dag_run_id=manual__2025-11-18T22%3A12%3A00.434537%2B00%3A00&task_id=preprocess

DAG pinecone_medium_pipeline / Run 2025-11-18, 22:12:00 UTC / Task preprocess

- download_csv
- preprocess
- create_index
- embed_and_upsert
- test_query

```
0969793bd6e0
*** Found local files:
***   * /opt/airflow/logs/dag_id=pinecone_medium_pipeline/run_id=manual__2025-11-18T22:12:00.434537+00:00/task_id=preprocess/attempt=1.log
[2025-11-18, 22:12:02 UTC] {local_task_job_runner.py:120} ▼ Pre task execution logs
[2025-11-18, 22:12:02 UTC] {taskinstance.py:2076} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: pinecone_medium_pipeline.preprocess manua
[2025-11-18, 22:12:02 UTC] {taskinstance.py:2076} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: pinecone_medium_pipeline.preprocess manual__2
[2025-11-18, 22:12:02 UTC] {taskinstance.py:2306} INFO - Starting attempt 1 of 2
[2025-11-18, 22:12:02 UTC] {taskinstance.py:2330} INFO - Executing <Task(PythonOperator): preprocess> on 2025-11-18 22:12:00.434537+00:00
[2025-11-18, 22:12:02 UTC] {warnings.py:110} WARNING - /home/***/.local/lib/python3.12/site-packages/***/task/task_runner/standard_task_runner.py:61: DeprecationWarning: This
  pid = os.fork()
[2025-11-18, 22:12:02 UTC] {standard_task_runner.py:63} INFO - Started process 211 to run task
[2025-11-18, 22:12:02 UTC] {standard_task_runner.py:90} INFO - Running: ['***', 'tasks', 'run', 'pinecone_medium_pipeline', 'preprocess', 'manual__2025-11-18T22:12:00.434537+
[2025-11-18, 22:12:02 UTC] {standard_task_runner.py:91} INFO - Job 21: Subtask preprocess
[2025-11-18, 22:12:02 UTC] {task_command.py:426} INFO - Running <TaskInstance: pinecone_medium_pipeline.preprocess manual__2025-11-18T22:12:00.434537+00:00 [running]> on host
[2025-11-18, 22:12:02 UTC] {taskinstance.py:2648} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='spartan' AIRFLOW_CTX_DAG_ID='pinecone_medium_pipeline' AIRFLOW_CTX_TASK_ID
[2025-11-18, 22:12:02 UTC] {taskinstance.py:430} ▲▲▲ Log group end
[2025-11-18, 22:12:02 UTC] {pinecone_medium_pipeline.py:50} INFO - Processed rows: 2498 -> /opt/***/data/medium_processed.parquet
[2025-11-18, 22:12:02 UTC] {python.py:237} INFO - Done. Returned value was: None
[2025-11-18, 22:12:02 UTC] {taskinstance.py:441} ▼ Post task execution logs
[2025-11-18, 22:12:02 UTC] {taskinstance.py:1206} INFO - Marking task as SUCCESS. dag_id=pinecone_medium_pipeline, task_id=preprocess, run_id=manual__2025-11-18T22:12:00.4345
[2025-11-18, 22:12:02 UTC] {local_task_job_runner.py:240} INFO - Task exited with return code 0
[2025-11-18, 22:12:02 UTC] {taskinstance.py:3498} INFO - 1 downstream tasks scheduled from follow-on schedule check
[2025-11-18, 22:12:02 UTC] {local_task_job_runner.py:222} ▲▲▲ Log group end
```

Version: v2.9.1
Git Version: .release:2d53c1089f78d8d1416f51af60e1e0354781c661



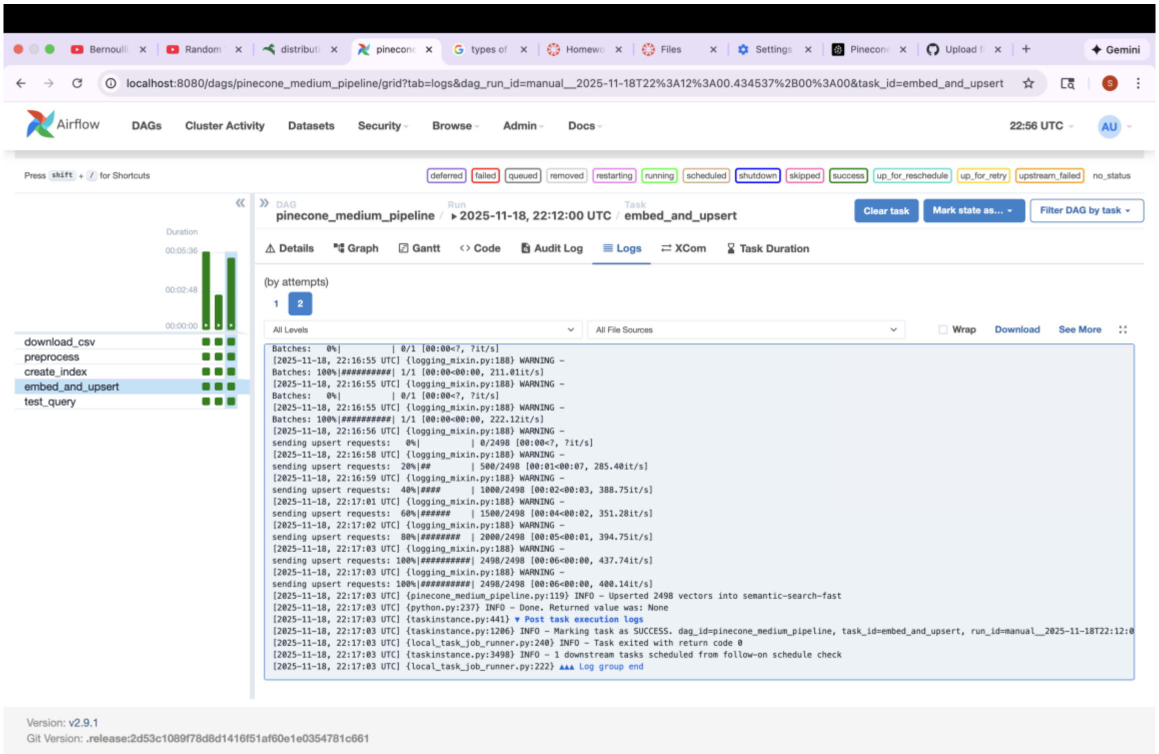
localhost:8080/dags/pinecone_medium_pipeline/grid?tab=logs&dag_run_id=manual__2025-11-18T22%3A12%3A00.434537%2B00%3A00&task_id=create_index

pinecone_medium_pipeline / ▸2025-11-18, 22:12:00 UTC / create_index

download_csv
preprocess
create_index
embed_and_upsert
test_query

```
0969793bd6e0
*** Found local files:
***   * /opt/airflow/logs/dag_id=pinecone_medium_pipeline/run_id=manual__2025-11-18T22:12:00.434537+00:00/task_id=create_index/attempt=1.log
[2025-11-18, 22:12:03 UTC] {local_task_job_runner.py:120} ▼ Pre task execution logs
[2025-11-18, 22:12:03 UTC] {taskinstance.py:2076} INFO - Dependencies all met for dep_context=non-requeueable deps ti=<TaskInstance: pinecone_medium_pipeline.create_index man
[2025-11-18, 22:12:03 UTC] {taskinstance.py:2076} INFO - Dependencies all met for dep_context=requeueable deps ti=<TaskInstance: pinecone_medium_pipeline.create_index manual_
[2025-11-18, 22:12:03 UTC] {taskinstance.py:2306} INFO - Starting attempt 1 of 2
[2025-11-18, 22:12:03 UTC] {taskinstance.py:2330} INFO - Executing <Task(PythonOperator): create_index> on 2025-11-18 22:12:00.434537+00:00
[2025-11-18, 22:12:03 UTC] {warnings.py:110} WARNING - /home/***/.local/lib/python3.12/site-packages/***/task/task_runner/standard_task_runner.py:61: DeprecationWarning: This
  pid = os.fork()
[2025-11-18, 22:12:03 UTC] {standard_task_runner.py:63} INFO - Started process 216 to run task
[2025-11-18, 22:12:03 UTC] {standard_task_runner.py:90} INFO - Running: ['***', 'tasks', 'run', 'pinecone_medium_pipeline', 'create_index', 'manual__2025-11-18T22:12:00.43453
[2025-11-18, 22:12:03 UTC] {standard_task_runner.py:91} INFO - Job 23: Subtask create_index
[2025-11-18, 22:12:03 UTC] {task_command.py:426} INFO - Running <TaskInstance: pinecone_medium_pipeline.create_index manual__2025-11-18T22:12:00.434537+00:00 [running]> on ho
[2025-11-18, 22:12:03 UTC] {taskinstance.py:2648} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='spartan' AIRFLOW_CTX_DAG_ID='pinecone_medium_pipeline' AIRFLOW_CTX_TASK_ID
[2025-11-18, 22:12:03 UTC] {taskinstance.py:430} ▲▲▲ Log group end
[2025-11-18, 22:12:06 UTC] {pinecone_medium_pipeline.py:75} INFO - Index already exists: semantic-search-fast
[2025-11-18, 22:12:06 UTC] {python.py:237} INFO - Done. Returned value was: None
[2025-11-18, 22:12:06 UTC] {taskinstance.py:441} ▼ Post task execution logs
[2025-11-18, 22:12:06 UTC] {taskinstance.py:1206} INFO - Marking task as SUCCESS. dag_id=pinecone_medium_pipeline, task_id=create_index, run_id=manual__2025-11-18T22:12:00.43
[2025-11-18, 22:12:06 UTC] {local_task_job_runner.py:240} INFO - Task exited with return code 0
[2025-11-18, 22:12:06 UTC] {taskinstance.py:3498} INFO - 1 downstream tasks scheduled from follow-on schedule check
[2025-11-18, 22:12:06 UTC] {local_task_job_runner.py:222} ▲▲▲ Log group end
```

Version: v2.9.1
Git Version: .release:2d53c1089f78d8d1416f51af60e1e0354781c661



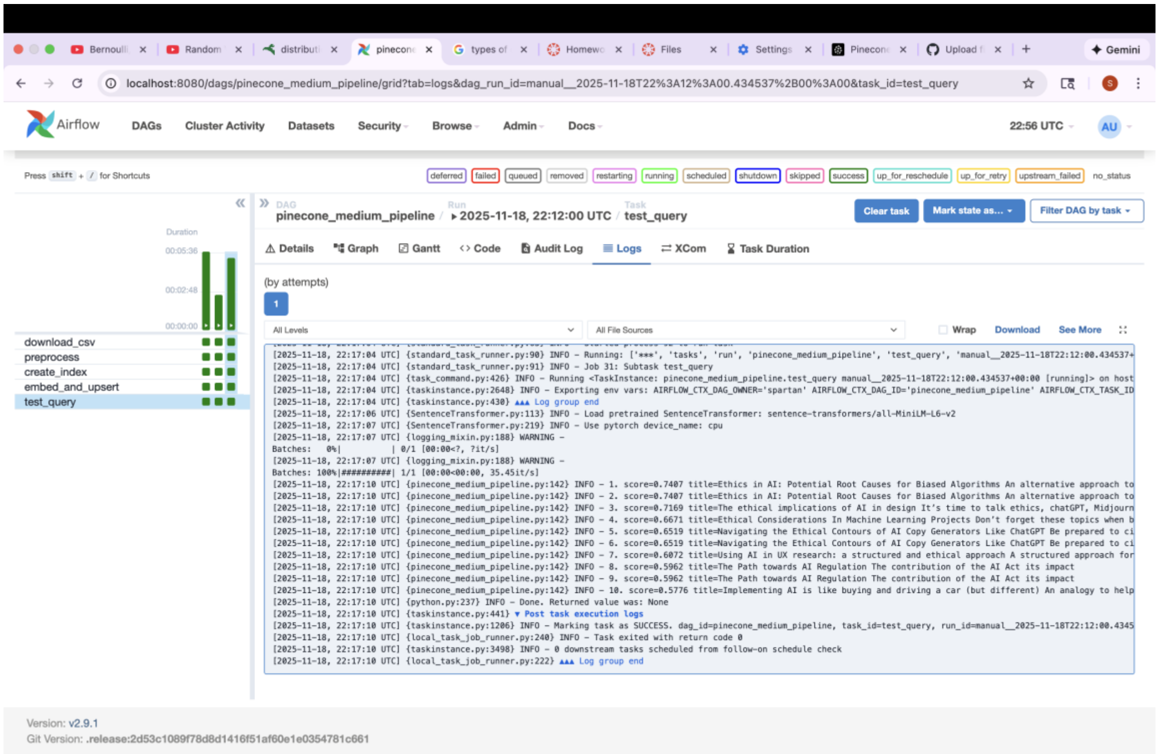
localhost:8080/dags/pinecone_medium_pipeline/grid?tab=logs&dag_run_id=manual__2025-11-18T22%3A12%3A00.434537%2B00%3A00&task_id=embed_and_upsert

pinecone_medium_pipeline / ▸2025-11-18, 22:12:00 UTC / embed_and_upsert

download_csv
preprocess
create_index
embed_and_upsert
test_query

```
Batches:   0%|          | 0/1 [00:00<?, ?it/s]
[2025-11-18, 22:16:55 UTC] {logging_mixin.py:188} WARNING -
Batches: 100%|##########| 1/1 [00:00<00:00, 211.01it/s]
[2025-11-18, 22:16:55 UTC] {logging_mixin.py:188} WARNING -
Batches:   0%|          | 0/1 [00:00<?, ?it/s]
[2025-11-18, 22:16:55 UTC] {logging_mixin.py:188} WARNING -
Batches: 100%|##########| 1/1 [00:00<00:00, 222.12it/s]
[2025-11-18, 22:16:56 UTC] {logging_mixin.py:188} WARNING -
sending upsert requests:   0%|          | 0/2498 [00:00<?, ?it/s]
[2025-11-18, 22:16:58 UTC] {logging_mixin.py:188} WARNING -
sending upsert requests:  20%|##        | 500/2498 [00:01<00:07, 285.40it/s]
[2025-11-18, 22:16:59 UTC] {logging_mixin.py:188} WARNING -
sending upsert requests:  40%|####      | 1000/2498 [00:02<00:03, 388.75it/s]
[2025-11-18, 22:17:01 UTC] {logging_mixin.py:188} WARNING -
sending upsert requests:  60%|######    | 1500/2498 [00:04<00:02, 351.28it/s]
[2025-11-18, 22:17:02 UTC] {logging_mixin.py:188} WARNING -
sending upsert requests:  80%|########  | 2000/2498 [00:05<00:01, 394.75it/s]
[2025-11-18, 22:17:03 UTC] {logging_mixin.py:188} WARNING -
sending upsert requests: 100%|##########| 2498/2498 [00:06<00:00, 437.74it/s]
[2025-11-18, 22:17:03 UTC] {logging_mixin.py:188} WARNING -
sending upsert requests: 100%|##########| 2498/2498 [00:06<00:00, 400.14it/s]
[2025-11-18, 22:17:03 UTC] {pinecone_medium_pipeline.py:119} INFO - Upserted 2498 vectors into semantic-search-fast
[2025-11-18, 22:17:03 UTC] {python.py:237} INFO - Done. Returned value was: None
[2025-11-18, 22:17:03 UTC] {taskinstance.py:441} ▼ Post task execution logs
[2025-11-18, 22:17:03 UTC] {taskinstance.py:1206} INFO - Marking task as SUCCESS. dag_id=pinecone_medium_pipeline, task_id=embed_and_upsert, run_id=manual__2025-11-18T22:12:0
[2025-11-18, 22:17:03 UTC] {local_task_job_runner.py:240} INFO - Task exited with return code 0
[2025-11-18, 22:17:03 UTC] {taskinstance.py:3498} INFO - 1 downstream tasks scheduled from follow-on schedule check
[2025-11-18, 22:17:03 UTC] {local_task_job_runner.py:222} ▲▲▲ Log group end
```

Version: v2.9.1
Git Version: .release:2d53c1089f78d8d1416f51af60e1e0354781c661



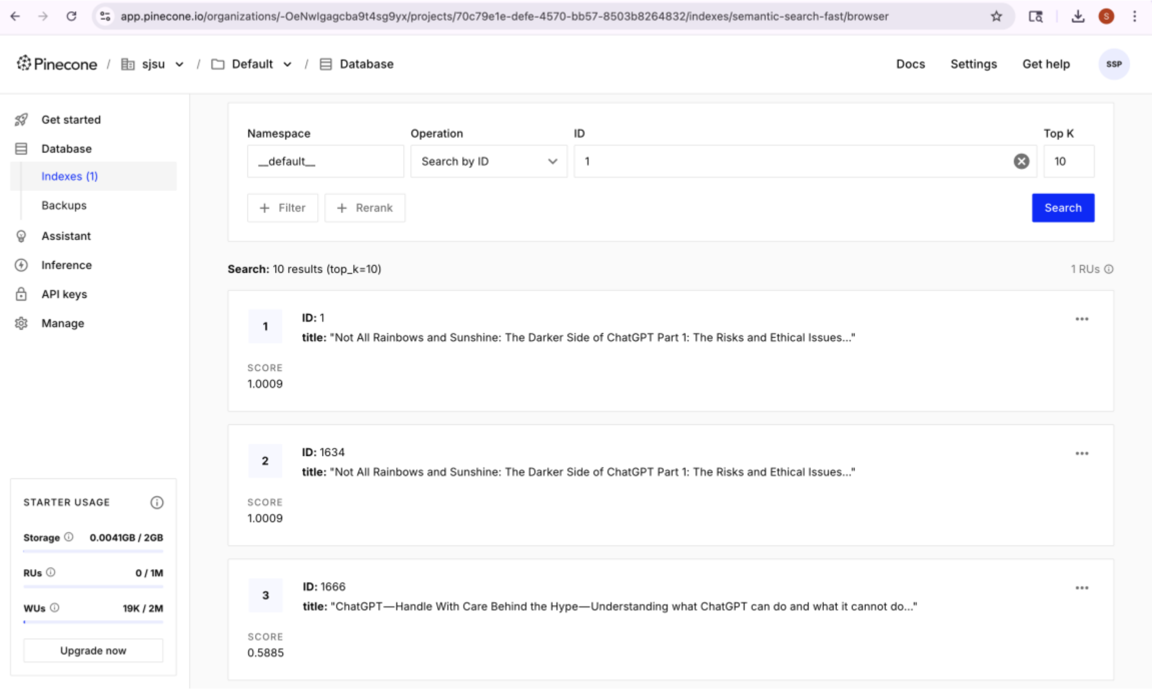
```
[2025-11-18, 22:17:04 UTC] {standard_task_runner.py:90} INFO - Running: ['***', 'tasks', 'run', 'pinecone_medium_pipeline', 'test_query', 'manual__2025-11-18T22:12:00.434537+
[2025-11-18, 22:17:04 UTC] {standard_task_runner.py:91} INFO - Job 31: Subtask test_query
[2025-11-18, 22:17:04 UTC] {task_command.py:426} INFO - Running <TaskInstance: pinecone_medium_pipeline.test_query manual__2025-11-18T22:12:00.434537+00:00 [running]> on host
[2025-11-18, 22:17:04 UTC] {taskinstance.py:2648} INFO - Exporting env vars: AIRFLOW_CTX_DAG_OWNER='spartan' AIRFLOW_CTX_DAG_ID='pinecone_medium_pipeline' AIRFLOW_CTX_TASK_ID
[2025-11-18, 22:17:04 UTC] {taskinstance.py:430} ▲▲▲ Log group end
[2025-11-18, 22:17:06 UTC] {SentenceTransformer.py:113} INFO - Load pretrained SentenceTransformer: sentence-transformers/all-MiniLM-L6-v2
[2025-11-18, 22:17:07 UTC] {SentenceTransformer.py:219} INFO - Use pytorch device_name: cpu
[2025-11-18, 22:17:07 UTC] {logging_mixin.py:188} WARNING -
Batches:   0%|          | 0/1 [00:00<?, ?it/s]
[2025-11-18, 22:17:07 UTC] {logging_mixin.py:188} WARNING -
Batches: 100%|##########| 1/1 [00:00<00:00, 35.45it/s]
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 1. score=0.7407 title=Ethics in AI: Potential Root Causes for Biased Algorithms An alternative approach to
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 2. score=0.7407 title=Ethics in AI: Potential Root Causes for Biased Algorithms An alternative approach to
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 3. score=0.7169 title=The ethical implications of AI in design It's time to talk ethics, chatGPT, Midjourn
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 4. score=0.6671 title=Ethical Considerations In Machine Learning Projects Don't forget these topics when b
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 5. score=0.6519 title=Navigating the Ethical Contours of AI Copy Generators Like ChatGPT Be prepared to ci
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 6. score=0.6519 title=Navigating the Ethical Contours of AI Copy Generators Like ChatGPT Be prepared to ci
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 7. score=0.6072 title=Using AI in UX research: a structured and ethical approach A structured approach for
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 8. score=0.5962 title=The Path towards AI Regulation The contribution of the AI Act its impact
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 9. score=0.5962 title=The Path towards AI Regulation The contribution of the AI Act its impact
[2025-11-18, 22:17:10 UTC] {pinecone_medium_pipeline.py:142} INFO - 10. score=0.5776 title=Implementing AI is like buying and driving a car (but different) An analogy to help
[2025-11-18, 22:17:10 UTC] {python.py:237} INFO - Done. Returned value was: None
[2025-11-18, 22:17:10 UTC] {taskinstance.py:441} ▼ Post task execution logs
[2025-11-18, 22:17:10 UTC] {taskinstance.py:1206} INFO - Marking task as SUCCESS. dag_id=pinecone_medium_pipeline, task_id=test_query, run_id=manual__2025-11-18T22:12:00.4345
[2025-11-18, 22:17:10 UTC] {local_task_job_runner.py:240} INFO - Task exited with return code 0
[2025-11-18, 22:17:10 UTC] {taskinstance.py:3498} INFO - 0 downstream tasks scheduled from follow-on schedule check
[2025-11-18, 22:17:10 UTC] {local_task_job_runner.py:222} ▲▲▲ Log group end
```

5. Conclusion

This project demonstrated the integration of Apache Airflow with Pinecone for managing and automating semantic search workflows.
By using Docker, the environment remained fully isolated and reproducible. Airflow ensured that each stage of the data pipeline
executed in the correct order and was fully traceable through logs and UI visualization.

Key learnings from this assignment include:
- Managing Python dependencies in Airflow via Dockerfiles.
- Handling external API integrations (Pinecone) using Airflow Variables.
- Debugging distributed task failures in Airflow.
- Designing a robust, idempotent data ingestion and search workflow.

The final output proved that the pipeline could autonomously process text data, generate embeddings, store them in Pinecone,
and perform vector-based semantic search efficiently. The workflow's modular design allows it to be extended for larger datasets
or other vector databases in the future.
(Pine cone verification screenshot on next page)


app.pinecone.io/organizations/-OeNwlgagcba9t4sg9yx/projects/70c79e1e-defe-4570-bb57-8503b8264832/indexes/semantic-search-fast/browser
Pinecone / sjsu / Default / Database
Docs
Settings
Get help
SSP
Get started
Database
Indexes (1)
Backups
Assistant
Inference
API keys
Manage
+ Filter
+ Rerank
Search
Upgrade now

STARTER USAGE

Storage 0.0041GB / 2GB

RUs 0 / 1M

WUs 19K / 2M

Namespace: __default__

Operation: Search by ID

ID: 1

Top K: 10

Search: 10 results (top_k=10)

1 RUs

1

ID: 1

title: "Not All Rainbows and Sunshine: The Darker Side of ChatGPT Part 1: The Risks and Ethical Issues..."

SCORE
1.0009

2

ID: 1634

title: "Not All Rainbows and Sunshine: The Darker Side of ChatGPT Part 1: The Risks and Ethical Issues..."

SCORE
1.0009

3

ID: 1666

title: "ChatGPT—Handle With Care Behind the Hype—Understanding what ChatGPT can do and what it cannot do..."

SCORE
0.5885