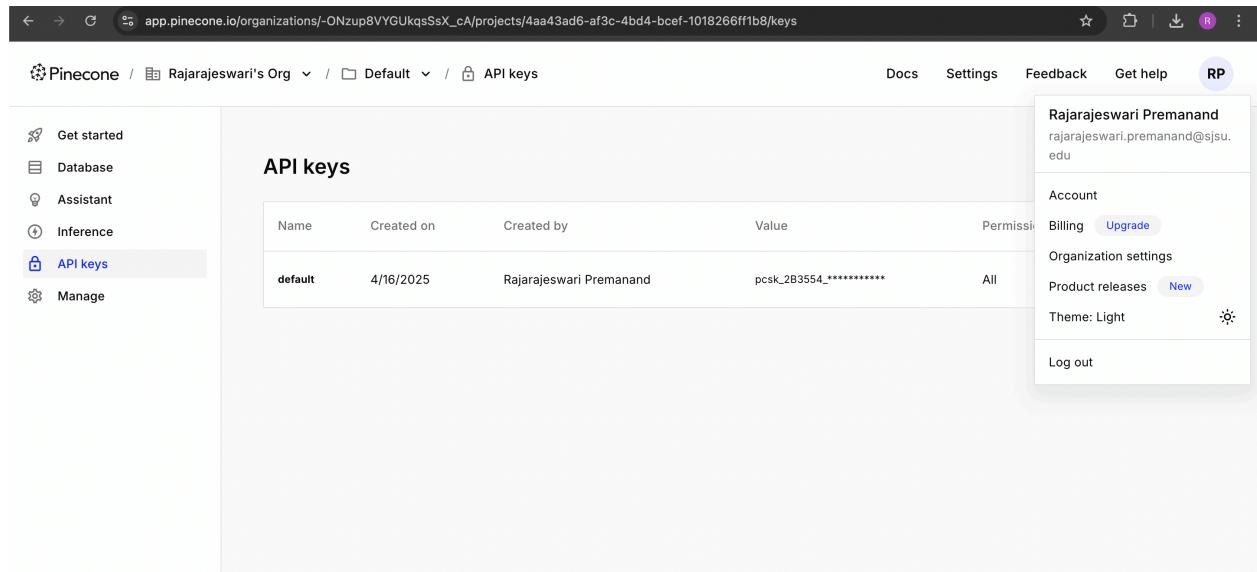


HOMEWORK 8

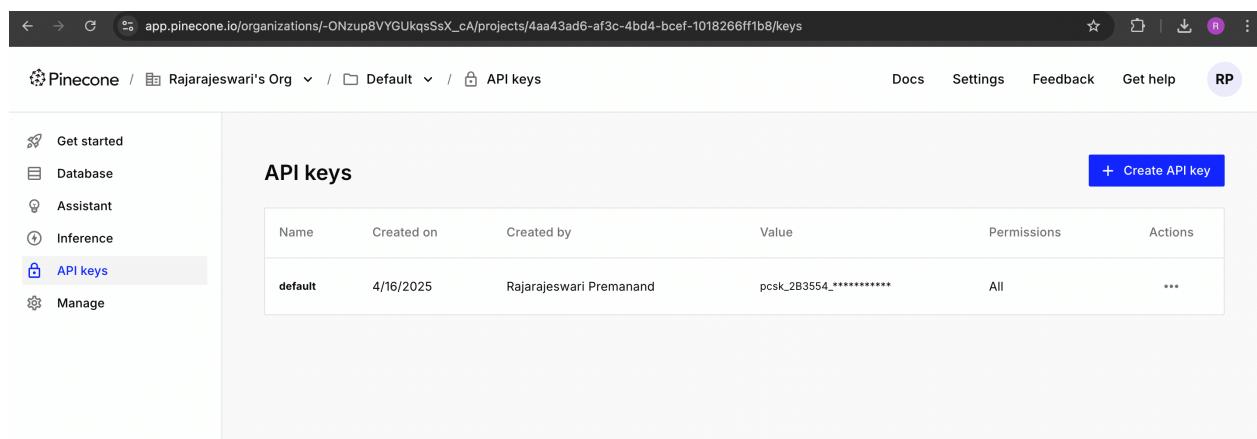
Github link is : https://github.com/Rajeswari195/Data-226_Assignments

1. Question 2

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



The screenshot shows the Pinecone web interface. The URL is https://app.pinecone.io/organizations/-ONzup8VYGuqssX_cA/projects/4aa43ad6-af3c-4bd4-bcef-1018266ff1b8/keys. The sidebar on the left has links for Get started, Database, Assistant, Inference, API keys (which is selected and highlighted in blue), and Manage. The main content area is titled "API keys" and shows a table with one row. The table columns are Name, Created on, Created by, Value, Permissions, and Actions. The single row contains: default, 4/16/2025, Rajarajeswari Premanand, pcsk_2B3554_***** (redacted), All, and three dots (...). On the right side, there is a sidebar with account information (Rajarajeswari Premanand, rajarajeswari.premanand@sjtu.edu), account settings (Billing, Organization settings, Product releases), theme selection (Theme: Light), and a log out button.



This screenshot is similar to the previous one, showing the Pinecone API keys page. The URL is the same: https://app.pinecone.io/organizations/-ONzup8VYGuqssX_cA/projects/4aa43ad6-af3c-4bd4-bcef-1018266ff1b8/keys. The sidebar and table structure are identical to the first screenshot. However, there is a prominent blue button labeled "+ Create API key" located at the top right of the table header. The rest of the interface and sidebar are identical to the first screenshot.

List Variable				
<input type="button" value="Choose File"/> No file chosen <input checked="" type="radio"/> Overwrite if exists <input type="radio"/> Fail if exists <input type="radio"/> Skip if exists <input type="button" value="Import Variables"/>				
<input type="text" value="Search"/>				
Actions	Key	Val	Description	Is Encrypted
<input type="checkbox"/>	alpha_vantage_api_key	*****		False
<input type="checkbox"/>	pinecone_api_key	*****		False
<input type="checkbox"/>	snowflake_account	sfedu02-kab65579		False
<input type="checkbox"/>	snowflake_password	*****		False
<input type="checkbox"/>	snowflake_userid	GATOR		False

localhost:8081/variable/edit/5

Edit Variable	
Key *	pinecone_api_key
Val	*****
Description	Description
<input type="button" value="Save"/> <input type="button" value="Cancel"/>	

2. Question 6

Run search against Pinecone (1pt)

download_data

The screenshot shows the Airflow web interface for the Medium_to_Pinecone DAG. The top navigation bar includes links for Airflow, DAGs, Cluster Activity, Datasets, Security, Browse, Admin, and Docs. The top right corner shows the current time as 05:59 UTC and a refresh button. The main page displays the Medium_to_Pinecone DAG, which is currently scheduled to run from 2025-04-10T00:00:00 UTC. The DAG has a single task, download_data, which is currently running. The task duration chart shows the task took approximately 00:00:01. The logs tab is open, displaying the execution log for the task. The log output includes:

```
2025-04-17 05:55:36 UTC [FileDownloader] INFO - Found local file: /tmp/medium_data/medium_data.csv
2025-04-17 05:55:36 UTC [FileDownloader] INFO - > /opt/airflow/dags/04_dag_id-Medium_to_Pinecone/runs/2025-04-17T00:00:00Z/task_id-download_data/attempt=1.log
2025-04-17 05:55:36 UTC [LocalTaskJobRunner] INFO - Pre task execution log:
2025-04-17 05:55:36 UTC [logging_main.py:198] INFO - Downloaded file has 2499 lines
2025-04-17 05:55:36 UTC [python.py:148] INFO - Done. Returned value was: /tmp/medium_data/medium_data.csv
2025-04-17 05:55:37 UTC [TaskInstance] INFO - Post task execution log:
2025-04-17 05:55:37 UTC [TaskInstance] INFO - Post task execution log
```

Pre-process_data

The screenshot shows the Airflow web interface for a DAG named 'Medium_to_Pinecone'. The DAG has one task, 'preprocess_data', which was run on April 10, 2025, at 00:00:00 UTC. The task duration was 00:00:00. The task status is 'success'. The logs for this task show the following output:

```
1ebd52d4-1e1c-45a1-b1a1-*** + cogitflowlogs/dag_id/Medium_to_Pinecone/run_id=scheduled_2025-04-10T00%3A00%2B00%3A00/task_id=preprocess_data/attemp=1.log
[2025-04-17, 05:55:16 UTC] [local_task_job_runner.py:223] * Pre task execution logs
[2025-04-17, 05:55:16 UTC] [local_task_job_runner.py:223] * Task preprocess_data has been moved to /tmp/medium_data/medium_preprocessed.csv
[2025-04-17, 05:55:48 UTC] [ipykernel.py:248] INFO - Done. Returned wsc: /tmp/medium_data/medium_preprocessed.csv
[2025-04-17, 05:55:48 UTC] [taskinstance.py:340] * Post task execution logs
```

Create_pinecone_index

The screenshot shows the Airflow web interface for the Medium_to_Pinecone DAG. The top navigation bar includes links for Airflow, DAGs, Cluster Activity, Datasets, Security, Browse, Admin, and Docs. The status bar at the top right shows "06:00 UTC". The main dashboard displays the DAG's progress: it was scheduled 7 days ago and has a next run ID of 2025-04-17, 00:00:00 UTC. The DAG itself is titled "Medium_to_Pinecone" and is described as "Build a Medium Posting Search Engine using Pinecone". It has a single task named "create_pinecone_index" which is currently running. The DAG visualization shows a green bar indicating the task's duration. Below the DAG view, there are tabs for Details, Graph, Gantt, Code, Event Log, Logs (selected), XCom, and Task Duration. The Logs tab shows log entries from 2025-04-17, 05:55:52 UTC, detailing the execution of the Python code for creating a Pinecone index. The bottom of the page includes buttons for Clear task, Mark state as..., Filter DAG by task, Wrap, Download, and See More.

Generate_embeddings_and_upsert

The screenshot shows the Airflow interface for the Medium_to_Pinecone DAG. The top navigation bar includes links for Airflow, DAGs, Cluster Activity, Datasets, Security, Browse, Admin, and Docs. The title bar displays the URL as localhost:8081/dags/Medium_to_Pinecone/grid?dag_run_id=scheduled__2025-04-10T00%3A00%2B00%3A00&task_id=generate_embeddings_a... and the current time as 06:01 UTC. The main content area shows the DAG structure with tasks: download_data, preprocess_data, create_pinecone_index, generate_embeddings_and_upsert, and test_search_query. The 'generate_embeddings_and_upsert' task is currently running, with a duration of 00:00:00 and an end time of Apr 10, 2025, 04:57:49. The task log shows multiple batches of embeddings being processed and successfully upserted into Pinecone. The bottom right corner has a 'Clear task' button and a 'Mark state as...' dropdown.

Test_search_query