DATA SCIENCE PROJECT

Presented by: ขอพักแปป

OVERVIEW

01

About Project

02

Overview

03

Airflow

04

API Scrapping

05

Visulize

ABOUT PROJECT

In this project, we wil use DE tools to prepare data for visulization

OVERVIEW

API SCRAPING

use Scopus API

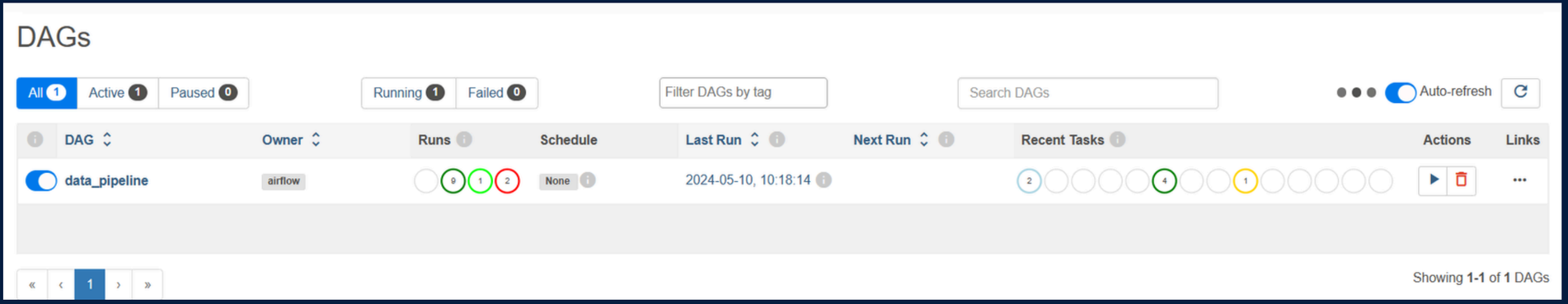
```
def fetch_data(idx, apiKey, last_page, pages):
    dois = []
    public_year = 2018
    affiliation = "chulalongkorn%20niversity"
    country = "thailand"
    for i in range(last_page,pages):
        URL = f"https://api.elsevier.com/content/search/scopus?query=AFFILCOUNTRY({country})%20AND%20AFFILORG
        ({affiliation})%20AND%20PUBYEAR%20<%20{public_year}&apiKey={apiKey}&start={str(idx * i)}&
        httpAccept=application/json"
        response = requests.get(URL)
```

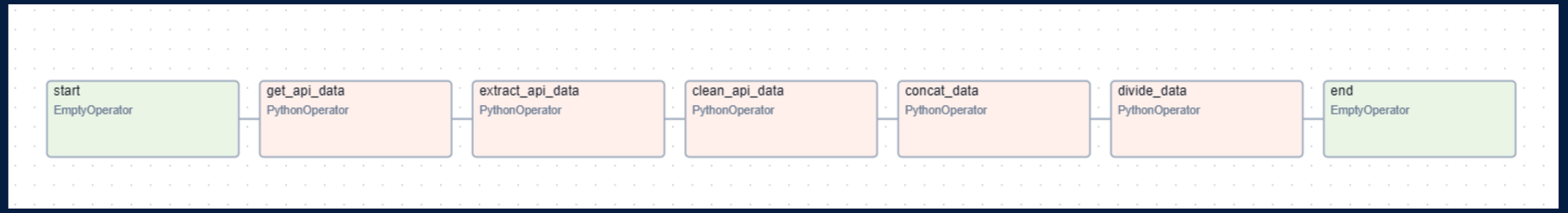
API SCRAPING

use Scopus API

```python
def get_api_data():
    idx = 25
    first_page = 0
    num_pages = 40
    dois = fetch_data(idx, apiKey, first_page, first_page + num_pages)
    bucket_name = "datapipeline"

    for i in range(len(dois)):
        research_data = requests.get('https://api.elsevier.com/content/abstract/DOI:' + dois[i] + '?apiKey=' +
        apiKey + '&view=FULL&httpAccept=application/json')

        if not os.path.exists('/opt/airflow/data/api_data_json'):
            os.makedirs('/opt/airflow/data/api_data_json')

        file_path = f"/opt/airflow/data/api_data_json/research_{i}.json"
        object_name = f"data/api_data_json/research_{i}.json"
        if research_data.status_code == 200:
            with open(file_path, 'w', encoding='utf-8') as f:
                json.dump(research_data.json(), f)

            upload_to_minio(minio_client, bucket_name, object_name, file_path)
        else:
            print('Failed to retrieve data:', research_data.status_code)
```

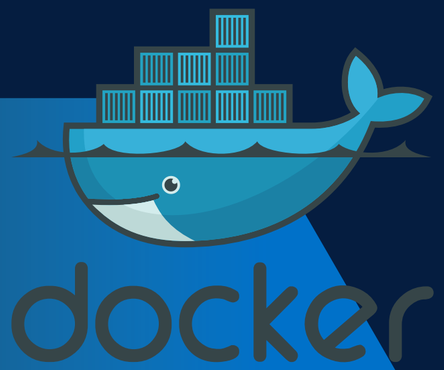
AIRFLOW


DAGs
All 1
Active 1
Paused 0
Running 1
Failed 0
Filter DAGs by tag
Search DAGs
Auto-refresh
DAG
Owner
Runs
Schedule
Last Run
Next Run
Recent Tasks
Actions
Links
data_pipeline
airflow
None
2024-05-10, 10:18:14
Showing 1-1 of 1 DAGs

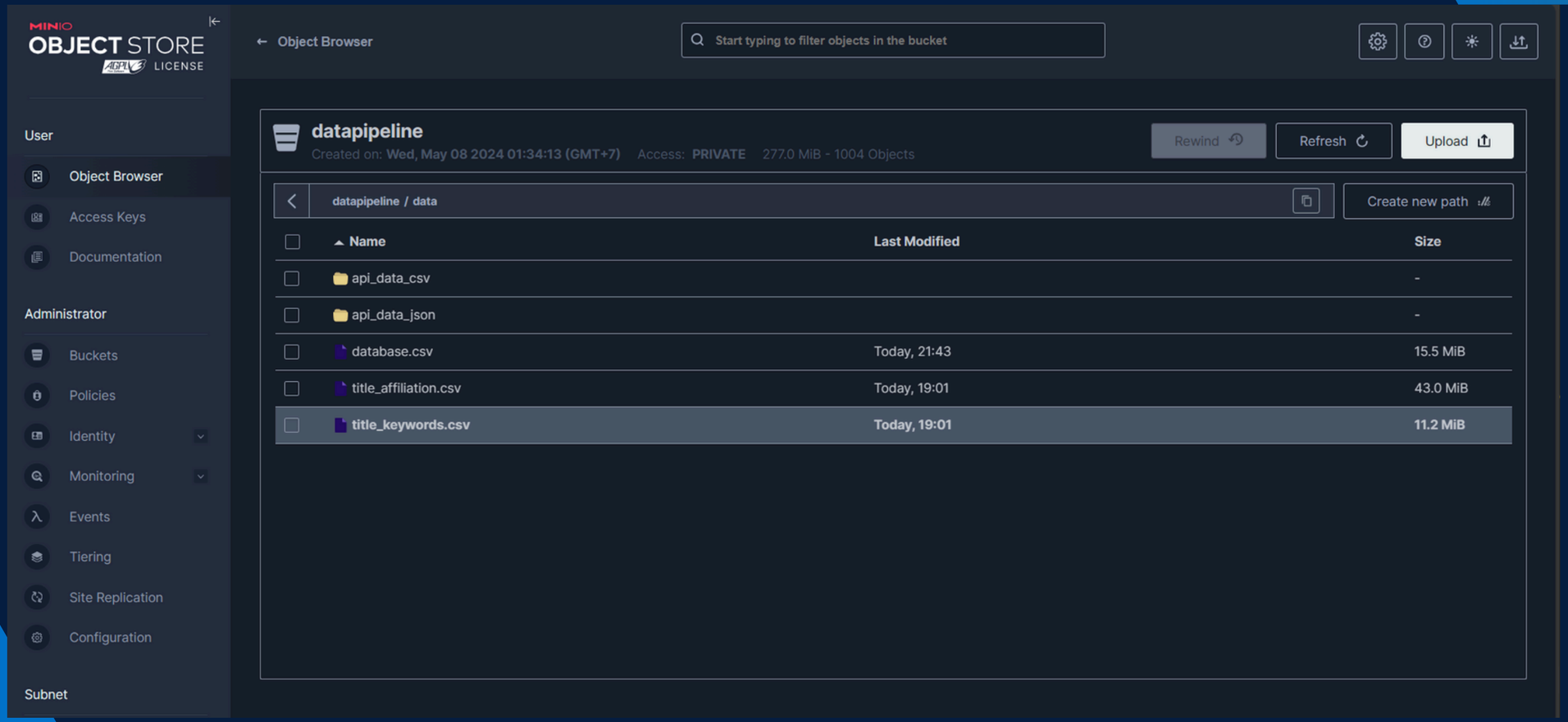

start
EmptyOperator
get_api_data
PythonOperator
extract_api_data
PythonOperator
clean_api_data
PythonOperator
concat_data
PythonOperator
divide_data
PythonOperator
end
EmptyOperator

WE RUN THIS ON DOCKER!!!

minio_server e0820c02347f	quay.io/minio/minio	Running	0.17%	9000:9000 Show all ports (2)	5 hours ago	
airflow-webserver-1 92c2a7aa2533	extending_airflow:latest	Running	1.41%	8080:8080	5 hours ago	
airflow-scheduler-1 dbfcb70dcca2	extending_airflow:latest	Running	2.82%		5 hours ago	
airflow-init-1 1278e9e328f6	extending_airflow:latest	Exited	0%		5 hours ago	



WE USE MINIO


MINIO
OBJECT STORE
AGPLv3 LICENSE
User
Object Browser
Access Keys
Documentation
Administrator
Buckets
Policies
Identity
Monitoring
Events
Tiering
Site Replication
Configuration
Subnet
Object Browser
Start typing to filter objects in the bucket
datapipeline
Created on: Wed, May 08 2024 01:34:13 (GMT+7) Access: PRIVATE 277.0 MiB - 1004 Objects
Rewind
Refresh
Upload
datapipeline / data
Create new path
Name
Last Modified
Size
api_data_csv -
api_data_json -
database.csv Today, 21:43 15.5 MiB
title_affiliation.csv Today, 19:01 43.0 MiB
title_keywords.csv Today, 19:01 11.2 MiB

STREAMLIT

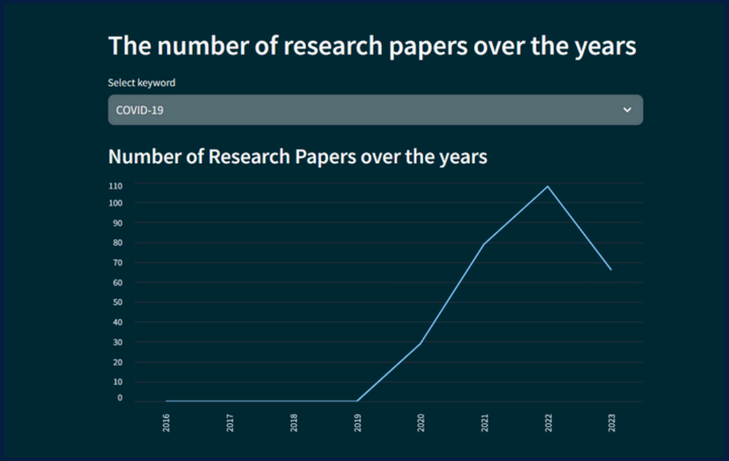
show top of Thailand university that collaborate with Chulalonkorn University (pie chart)

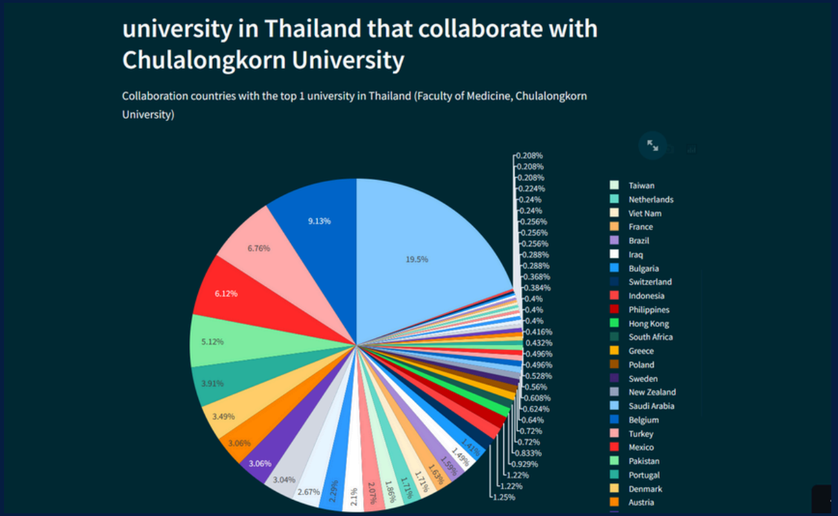
show top 10 keywords top of Thailand university that collaborate with Chulalonkorn University (pie chart)

show the collaboration countries with Chulalongkorn University (bar chart)

show the collaboration university in Top 1 collaboration countries(bar chart)

show the relationship between year and amount of research with specific keyword (line chart)





STREAMLIT

https://datasciproject-av4yb5cremwo6k6fmpffy6.streamlit.app

THANK YOU