

# Capstone Project – Natcha Chinnapan 64199130042

## Movies on Streaming Platforms

### Problem

ปัจจุบัน Streaming Platforms มีจากหลากหลาย platforms ผู้ใช้งานบางรายอาจต้องเสียเงินสมัครทุก platform เพื่อเข้าชมภาพยนตร์ที่ตนเองสนใจ หรือบางรายอาจจะต้องลองสมัครทุก platforms เพื่อจะได้ทราบว่าภาพยนตร์ใน platforms นั้นๆ ตรงกับความชอบของตนเองไหม ทำให้เกิดปัญหาการสมัคร platforms มากเกินไปจนดูไม่ทัน หรือประเภทหนึ่งใน platforms นั้นๆ ไม่ตรงกับความต้องการของตนเอง ทำให้เกิดการเสียเงินโดยเปล่าประโยชน์

จากปัญหาที่กล่าวมาข้างต้น เราจึงอยากแก้ไขปัญหานี้โดยการนำ Data มาวิเคราะห์และหาข้อมูลเชิงลึกว่าภาพยนตร์ในแต่ละ platforms เป็นอย่างไร เพื่อเป็นข้อมูลประกอบการตัดสินใจให้กับผู้ใช้งานที่กำลังมองหา streaming platforms ที่เหมาะกับตนเองได้

### Dataset

ข้อมูลภาพยนตร์จาก 4 streaming platforms ตั้งแต่ปี 1914-2021 ใช้การรวบรวมข้อมูลโดยการ scraping <https://www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney>

### Data model

เนื่องจาก Dataset ที่ได้มา ข้อมูลมีความสมบูรณ์ และ Column ทั้งหมดสามารถนำมาใช้ได้และเชื่อมกันอยู่แล้วจึงจัดทำเป็นแบบ NoSQL

movieplatforms	
movie_id	int
title	text
year	int
age_rate	text
tomatoes_rate	text
netflix	int
hulu	int
prime_video	int
disney_plus	int

## Data Pipeline

**Collection** : รวบรวมข้อมูลโดยการ scraping จาก 4 platforms ให้ออกมาเป็น Raw Data

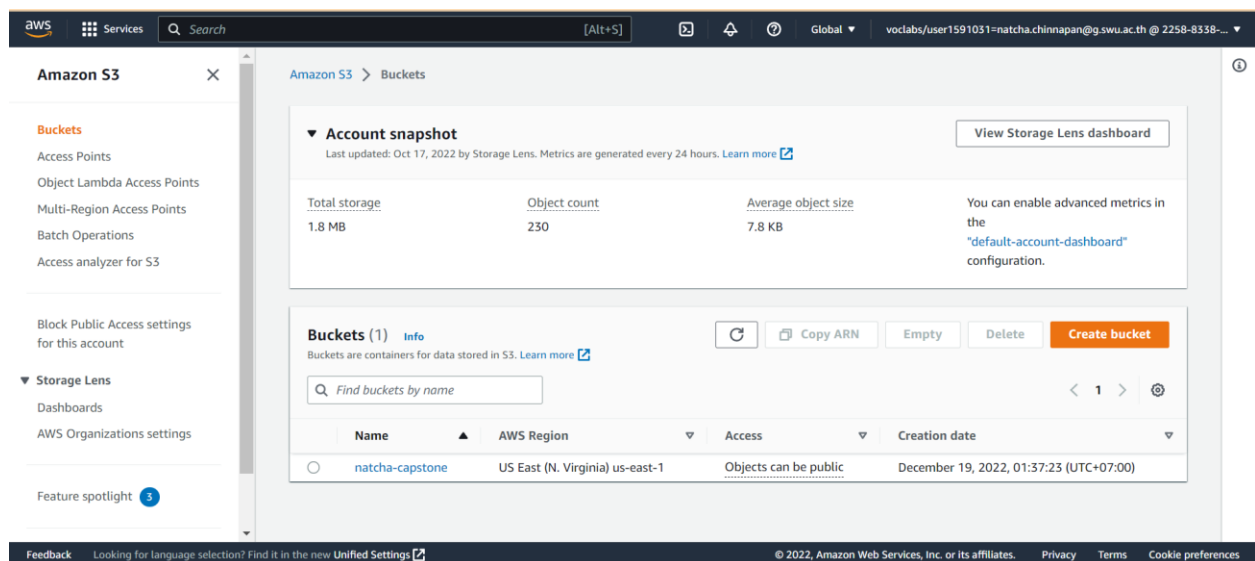
**Ingestion** : ใช้ S3 ของ AWS เป็น Data Lake ในการเก็บ Raw Data

**Integration** : ใช้ Redshift ของ AWS เป็น Data Warehouse โดยใช้ Airflow และใช้ dbt ในการสร้าง Schedule เพื่อดึงข้อมูลจาก S3 และสร้างตารางที่ Redshift และ Transform ข้อมูลไปเก็บที่ Redshift

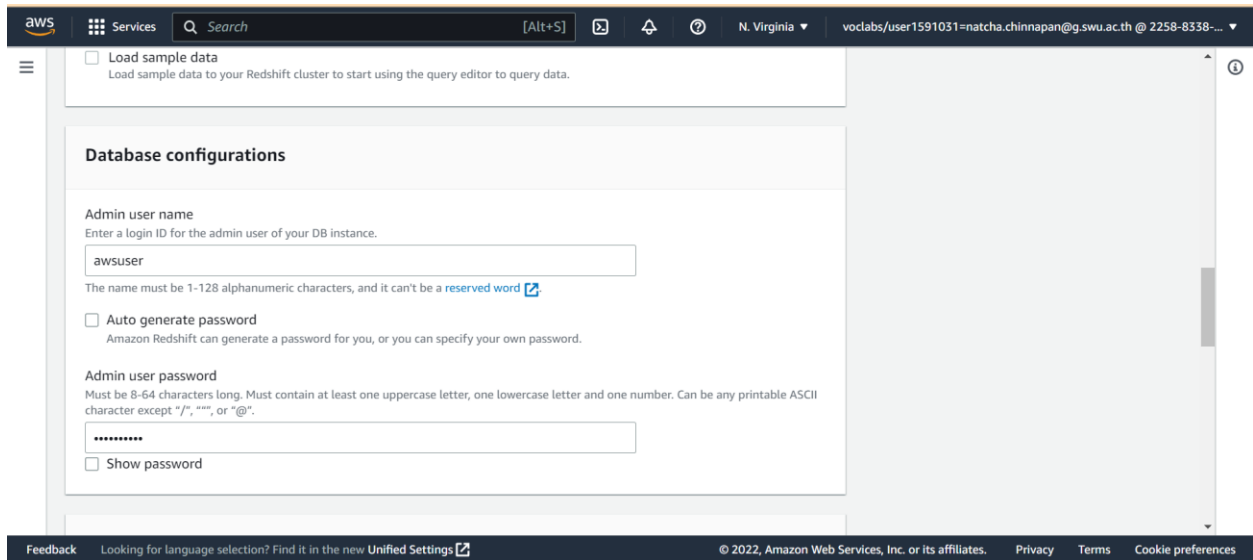
**Presentation** : จัดทำ Data Visualization และรวบรวมเป็น Dashboard เพื่อนำเสนอ โดยใช้ Tableau

### ขั้นตอนการดำเนินงาน

1. สร้าง Bucket ที่ S3 ใน AWS สำหรับเก็บ Raw data และทำการ Unblock all public เพื่อให้สามารถเข้าถึงข้อมูลใน bucket ได้



2. สร้าง Cluster ที่ Redshift ใน AWS สำหรับสร้าง Data Warehouse และ กำหนด username และ password สำหรับการเข้าถึง Cluster



☐ Load sample data  
Load sample data to your Redshift cluster to start using the query editor to query data.

### Database configurations

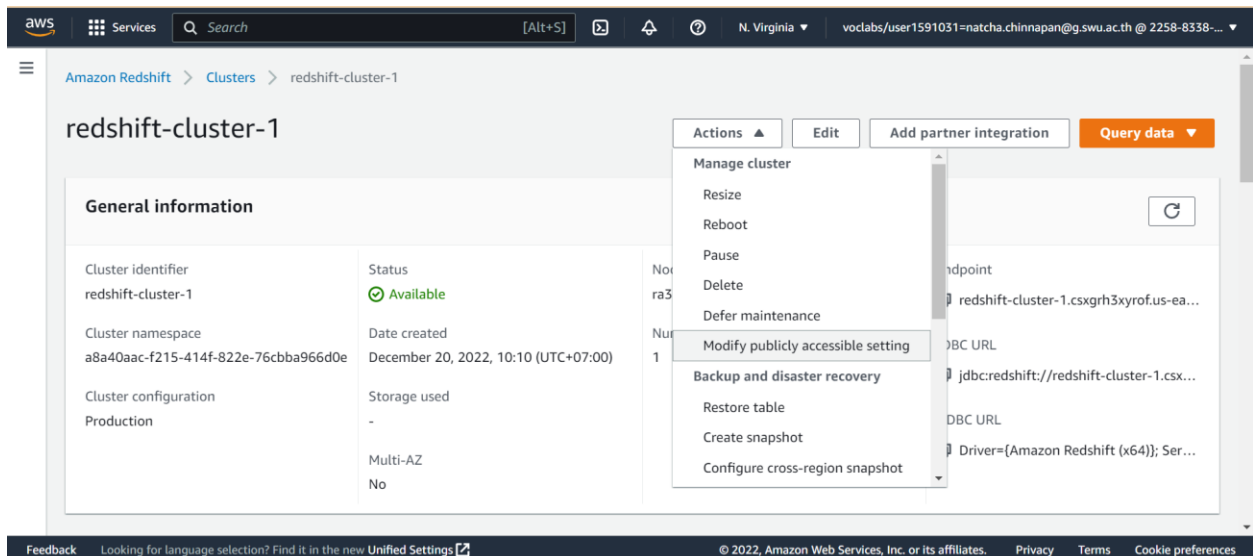
**Admin user name**  
Enter a login ID for the admin user of your DB instance.  
  
The name must be 1-128 alphanumeric characters, and it can't be a [reserved word](#).

☐ Auto generate password  
Amazon Redshift can generate a password for you, or you can specify your own password.

**Admin user password**  
Must be 8-64 characters long. Must contain at least one uppercase letter, one lowercase letter and one number. Can be any printable ASCII character except "/", "", or "@".  
  
☐ Show password

Feedback   Looking for language selection? Find it in the new [Unified Settings](#)   © 2022, Amazon Web Services, Inc. or its affiliates.   Privacy   Terms   Cookie preferences

3. กำหนดสิทธิ์ของ Cluster ให้เป็นแบบ public เพื่อให้สามารถเข้าถึงได้



Amazon Redshift > Clusters > redshift-cluster-1

## redshift-cluster-1

Actions   Edit   Add partner integration   Query data

### General information

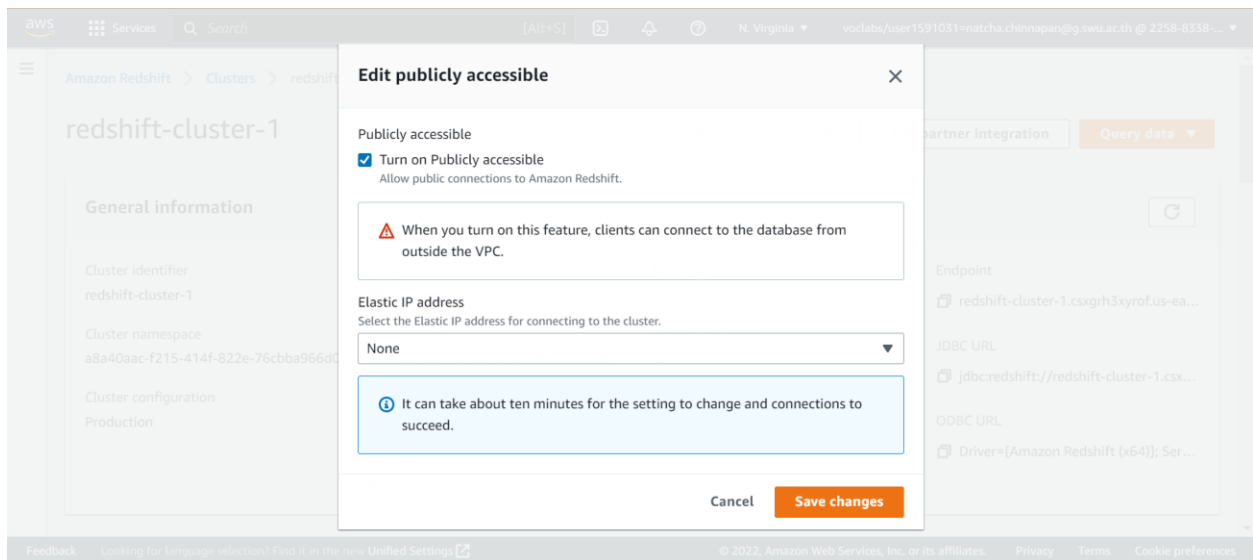
Cluster identifier	redshift-cluster-1	Status	Available
Cluster namespace	a8a40aac-f215-414f-822e-76cbba966d0e	Date created	December 20, 2022, 10:10 (UTC+07:00)
Cluster configuration	Production	Storage used	-
		Multi-AZ	No

**Actions**

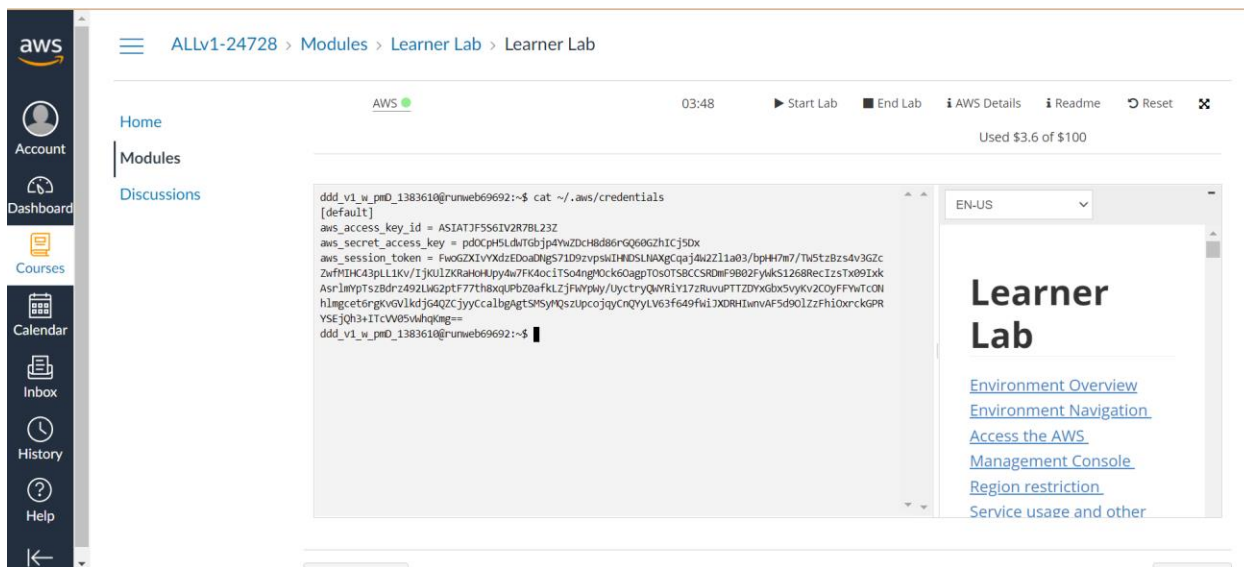
- Manage cluster
- Resize
- Reboot
- Pause
- Delete
- Defer maintenance
- Modify publicly accessible setting
- Backup and disaster recovery
- Restore table
- Create snapshot
- Configure cross-region snapshot

Endpoint: redshift-cluster-1.csxgrh3xyrof.us-east-1.amazonaws.com  
JDBC URL: jdbc:redshift://redshift-cluster-1.csxgrh3xyrof.us-east-1.amazonaws.com:5439/  
Driver={Amazon Redshift (x64)}; Server=redshift-cluster-1

Feedback   Looking for language selection? Find it in the new [Unified Settings](#)   © 2022, Amazon Web Services, Inc. or its affiliates.   Privacy   Terms   Cookie preferences



4. ดู secret keys ใน AWS console เพื่อใช้ในการเข้าถึง S3 โดยใช้คำสั่ง `cat ~/.aws/credentials` ซึ่ง secret keys ที่จะได้ จะมี 3 key คือ `aws_access_key_id`, `aws_secret_access_key` และ `aws_session_token`




## 5. run คำสั่ง docker-compose up เพื่อเปิดใช้งาน Apache Airflow เพื่อสร้าง Automating data pipelines

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
docker-compose - capstone-project + v [] [] ^ x

tureWarning: The auth_backends setting in [api] has had airflow.api.auth.backend.session added in the running config, which is needed by the UI. Please update your config before Apache Airflow 3.0.
capstone-project-airflow-scheduler-1 | FutureWarning,
capstone-project-airflow-triggerer-1 |
capstone-project-airflow-triggerer-1 |
capstone-project-airflow-triggerer-1 |
capstone-project-airflow-triggerer-1 |
capstone-project-airflow-triggerer-1 |
capstone-project-airflow-triggerer-1 | [2022-12-20 05:14:02,262] {triggerer_job.py:101} INFO - Starting the triggerer
capstone-project-airflow-webserver-1 | [2022-12-20 05:14:03 +0000] [60] [INFO] Starting gunicorn 20.1.0
capstone-project-airflow-scheduler-1 |
capstone-project-airflow-scheduler-1 |
capstone-project-airflow-scheduler-1 |
capstone-project-airflow-scheduler-1 |
capstone-project-airflow-scheduler-1 |
capstone-project-airflow-scheduler-1 | [2022-12-20 05:14:03 +0000] [63] [INFO] Starting gunicorn 20.1.0
capstone-project-airflow-scheduler-1 | [2022-12-20 05:14:03 +0000] [63] [INFO] Listening at: http://0.0.0.0:8793 (63)
capstone-project-airflow-scheduler-1 | [2022-12-20 05:14:03 +0000] [63] [INFO] Using worker: sync
capstone-project-airflow-scheduler-1 | [2022-12-20 05:14:03,671] {scheduler_job.py:701} INFO - Starting the scheduler
capstone-project-airflow-scheduler-1 | [2022-12-20 05:14:03 +0000] [64] [INFO] Booting worker with pid: 64
capstone-project-airflow-scheduler-1 | [2022-12-20 05:14:03,671] {scheduler_job.py:706} INFO - Processing each file at most -1 times
```

## 6. ใช้งาน Airflow ที่ port 8080 โดยใช้ username และ password คือ airflow และ สร้าง connection เพื่อเชื่อมต่อกับ Redshift โดยใส่รายละเอียดตามภาพ และทำการ test การเชื่อมต่อ

 Airflow

DAGs Datasets Security Browse Admin Docs

05:27 UTC AA

Connection Id \*

redshift

Connection Type \*

Postgres

Connection Type missing? Make sure you've installed the corresponding Airflow Provider Package.

Description

Host

redshift-cluster-1.csxgrh3xyrof.us-east-1.redshift.amazonaws.com

Schema

dev

Login


awsuser

Password

\*\*\*\*\*

Port

5439


Airflow
DAGs
Datasets
Security
Browse
Admin
Docs
05:38 UTC
AA

Connection successfully tested

Edit Connection

**Connection Id \***

**Connection Type \***

Postgres


Connection Type missing? Make sure you've installed the corresponding Airflow Provider Package.

**Description**

**Host**

**Schema**

7. Run คำสั่ง python etl\_s3.py เพื่อ upload raw data ไปที่ S3


Services

[Alt+S]
Global
voclabs/user1591031=natcha.chinnapan@g.swu.ac.th @ 2258-8338-...

**Amazon S3**

Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3

Block Public Access settings for this account

Storage Lens
Dashboards
AWS Organizations settings

Feature spotlight
3


Amazon S3
>
Buckets
>
natcha-capstone

**natcha-capstone**
Info


Objects
Properties
Permissions
Metrics
Management
Access Points

**Objects (1)**


Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)




Copy S3 URI




Copy URL



Download



Open




Delete

Actions

▼

Create folder




Upload

<

1

>

⌂

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	 moviesonstreamingplatforms.csv	csv	December 20, 2022, 00:48:17 (UTC+07:00)	421.6 KB	Standard

Feedback

Looking for language selection? Find it in the new [Unified Settings](#)

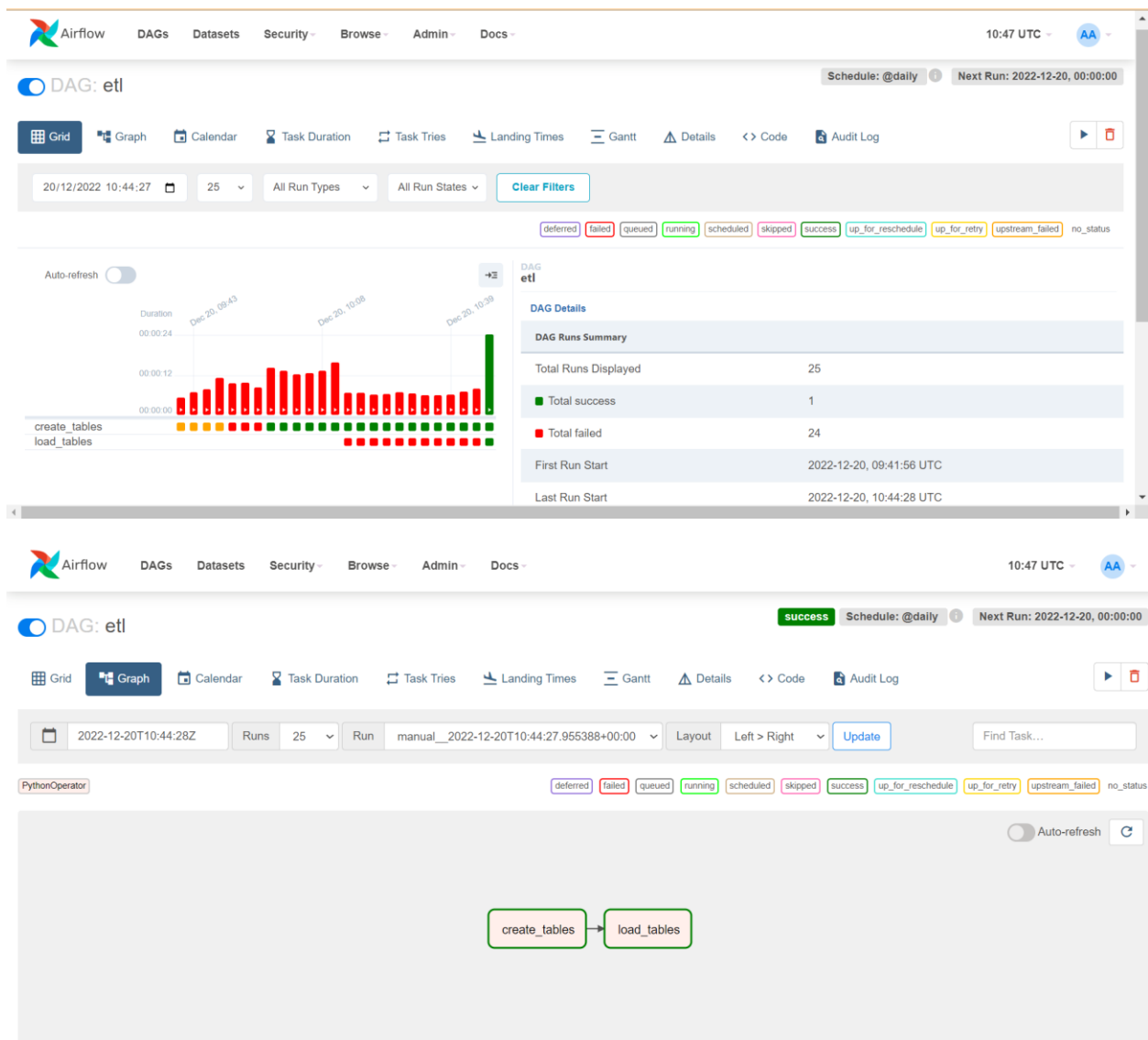
© 2022, Amazon Web Services, Inc. or its affiliates.

Privacy

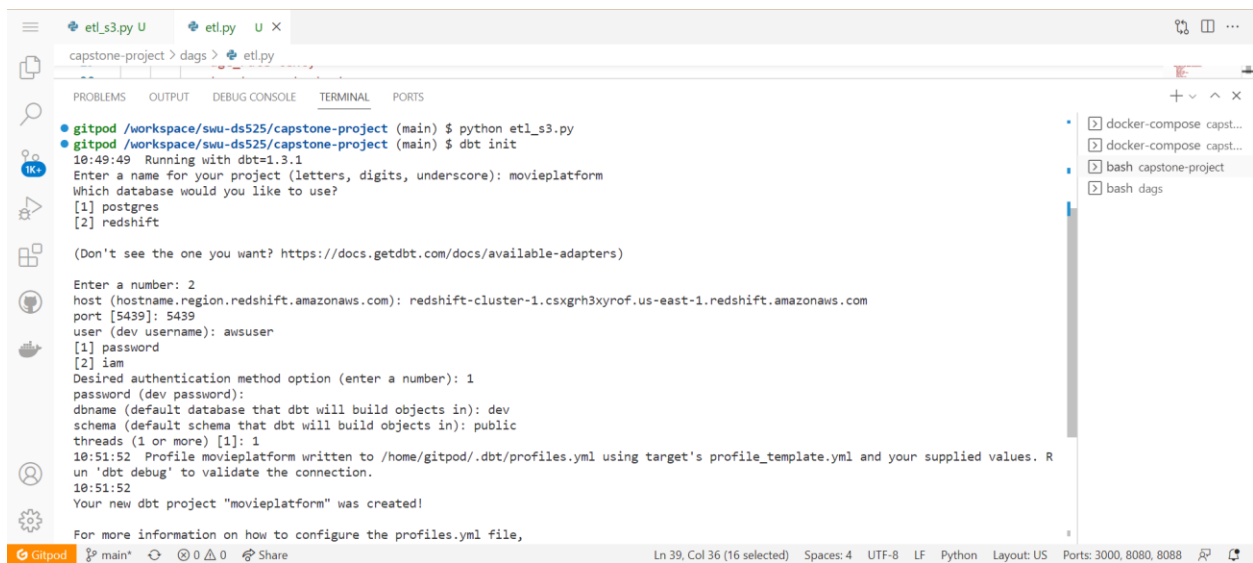
Terms

Cookie preferences

8. Run คำสั่ง python etl.py เพื่อสร้าง table และ load data จาก S3 ไปยัง Redshift หากดำเนินการสำเร็จ ใน Airflow จะแสดงสถานะ success ทุก task



9. ใช้ dbt ในการสร้าง data model และ transform data โดย run คำสั่ง dbt init และใส่ข้อมูลที่กำหนด



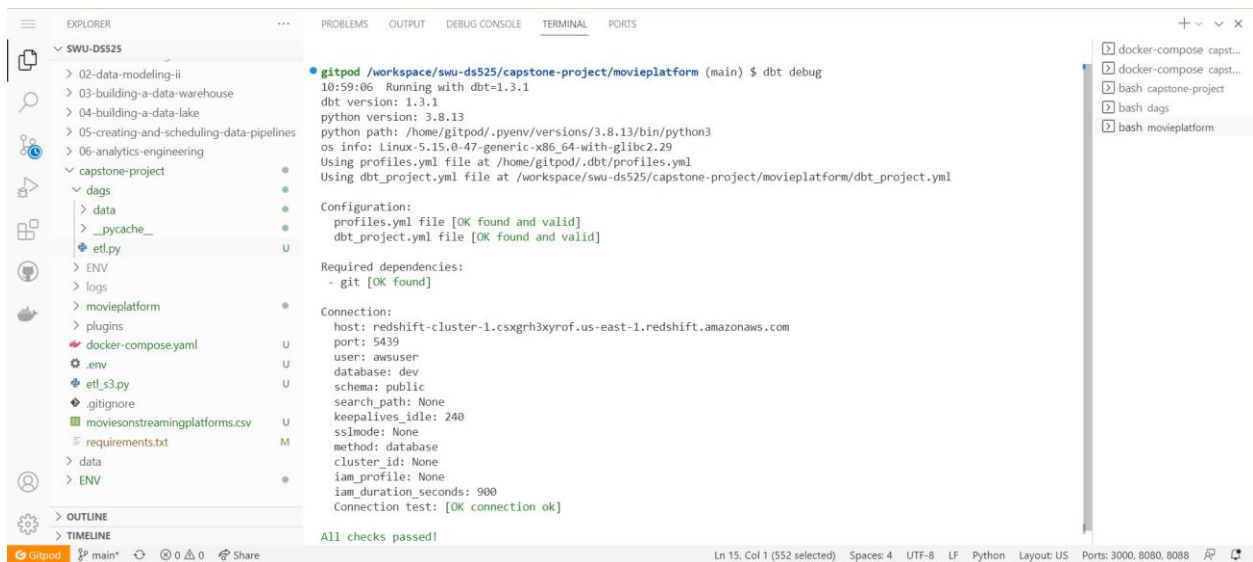
```
gitpod /workspace/swu-ds525/capstone-project (main) $ python etl_s3.py
gitpod /workspace/swu-ds525/capstone-project (main) $ dbt init
10:49:49 Running with dbt=1.3.1
Enter a name for your project (letters, digits, underscore): movieplatform
Which database would you like to use?
[1] postgres
[2] redshift

(Don't see the one you want? https://docs.getdbt.com/docs/available-adapters)

Enter a number: 2
host (hostname.region.redshift.amazonaws.com): redshift-cluster-1.csxgrh3xyrof.us-east-1.redshift.amazonaws.com
port [5439]: 5439
user (dev username): awsuser
[1] password
[2] iam
Desired authentication method option (enter a number): 1
password (dev password):
dbname (default database that dbt will build objects in): dev
schema (default schema that dbt will build objects in): public
threads (1 or more) [1]: 1
10:51:52 Profile movieplatform written to /home/gitpod/.dbt/profiles.yml using target's profile_template.yml and your supplied values. Run 'dbt debug' to validate the connection.
10:51:52
Your new dbt project "movieplatform" was created!

For more information on how to configure the profiles.yml file,
```

10. Run คำสั่ง dbt debug เพื่อตรวจสอบว่าสามารถ connection ได้หรือไม่



```
gitpod /workspace/swu-ds525/capstone-project/movieplatform (main) $ dbt debug
10:59:06 Running with dbt=1.3.1
dbt version: 1.3.1
python version: 3.8.13
python path: /home/gitpod/.pyenv/versions/3.8.13/bin/python3
os info: Linux-5.15.0-47-generic-x86_64-with-glibc2.29
Using profiles.yml file at /home/gitpod/.dbt/profiles.yml
Using dbt_project.yml file at /workspace/swu-ds525/capstone-project/movieplatform/dbt_project.yml

Configuration:
  profiles.yml file [OK found and valid]
  dbt_project.yml file [OK found and valid]

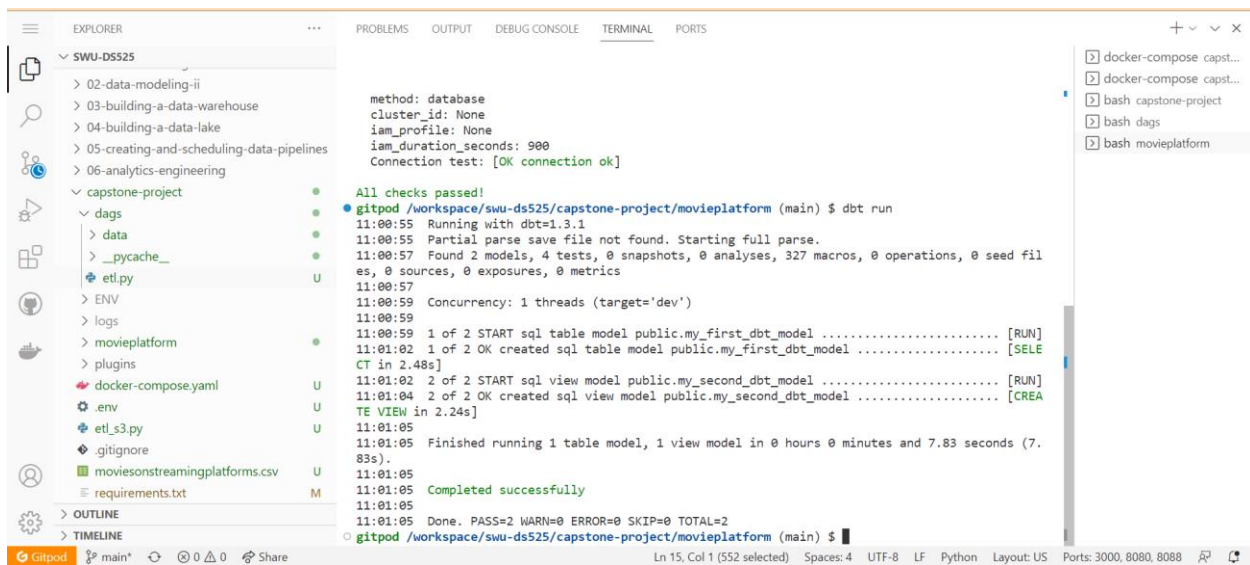
Required dependencies:
- git [OK found]

Connection:
  host: redshift-cluster-1.csxgrh3xyrof.us-east-1.redshift.amazonaws.com
  port: 5439
  user: awsuser
  database: dev
  schema: public
  search_path: None
  keepalives_idle: 240
  sslmode: None
  method: database
  cluster_id: None
  iam_profile: None
  iam_duration_seconds: 900
  Connection Test: [OK connection ok]

All checks passed!
```



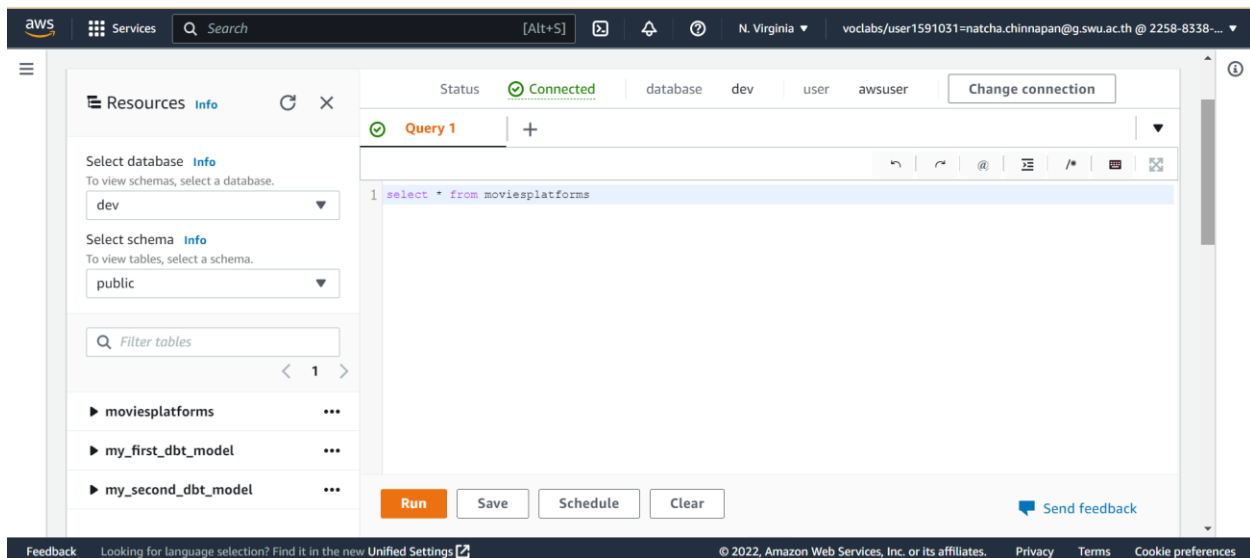
11. เมื่อสร้าง sql สำหรับ data model เรียบร้อยแล้ว ให้ run คำสั่ง dbt run เพื่อสร้าง data model และ transform data



```
method: database
cluster_id: None
iam_profile: None
iam_duration_seconds: 900
Connection test: [OK connection ok]

All checks passed!
gitpod /workspace/swu-ds525/capstone-project/movieplatform (main) $ dbt run
11:00:55 Running with dbt=1.3.1
11:00:55 Partial parse save file not found. Starting full parse.
11:00:57 Found 2 models, 4 tests, 0 snapshots, 0 analyses, 327 macros, 0 operations, 0 seed files, 0 sources, 0 exposures, 0 metrics
11:00:57
11:00:57 Concurrency: 1 threads (target='dev')
11:00:59
11:00:59 1 of 2 START sql table model public.my_first_dbt_model ..... [RUN]
11:01:02 1 of 2 OK created sql table model public.my_first_dbt_model ..... [SELECT]
CT in 2.48s]
11:01:02 2 of 2 START sql view model public.my_second_dbt_model ..... [RUN]
11:01:04 2 of 2 OK created sql view model public.my_second_dbt_model ..... [CREATE]
TE VIEW in 2.24s]
11:01:05
11:01:05 Finished running 1 table model, 1 view model in 0 hours 0 minutes and 7.83 seconds (7.83s).
11:01:05 Completed successfully
11:01:05 Done. PASS=2 WARN=0 ERROR=0 SKIP=0 TOTAL=2
gitpod /workspace/swu-ds525/capstone-project/movieplatform (main) $
```

12. ตรวจสอบข้อมูลใน Redshift หากดำเนินการสำเร็จจะสามารถ query ข้อมูลได้



aws Services Search [Alt+S] N. Virginia voclabs/user1591031=natcha.chinnapan@g.swu.ac.th @ 2258-8338-...

Query results Table details

Query 5071 [Execution](#) [Data](#) [Visualize](#)

Completed, started on December 20, 2022 at 18:23:28  
ELAPSED TIME: 00 m 11 s

Rows returned (9515) [Export](#)

Search rows

movie_id	title	year	age_rate	tomatoes_rate	netflix	hulu
1	The Irishman	2019	18 plus	98	1	0
2	Dangal	2016	13 plus	97	1	0
3	David Attenborough: A Life on Our Planet	2020	13 plus	95	1	0

Feedback Looking for language selection? Find it in the new Unified Settings © 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

### 13. เชื่อมต่อ Redshift กับ Tableau สำหรับการทำให้ Data Visualization

Tableau - Book1 - Tableau license expires in 14 days

File Data Server Help

Connect

Search for Data

Tableau Server

To a File

Microsoft Excel

Text file

JSON file

Microsoft Access

PDF file

Spatial file

Statistical file

More...

To a Server

Vertica

Web Data Connector

Other Databases (JDBC)

Other Databases (ODBC)

More...

Amazon Redshift

General Initial SQL Advanced

Server

redshift-cluster-1.csxgrh3xyrof.us-east-1.redshift.amazonaws.com

Port

5439

Database

dev

Username

awsuser

Password

\*\*\*\*\*

☐ Require SSL

Sign In

Services

SAP NetWeaver Business Warehouse

SAP Sybase ASE

SAP Sybase IQ

ServiceNow ITSM

SharePoint Lists

SingleStore

Snowflake

Spark SQL

Splunk

Teradata

Teradata OLAP Connector

TIBCO Data Virtualization

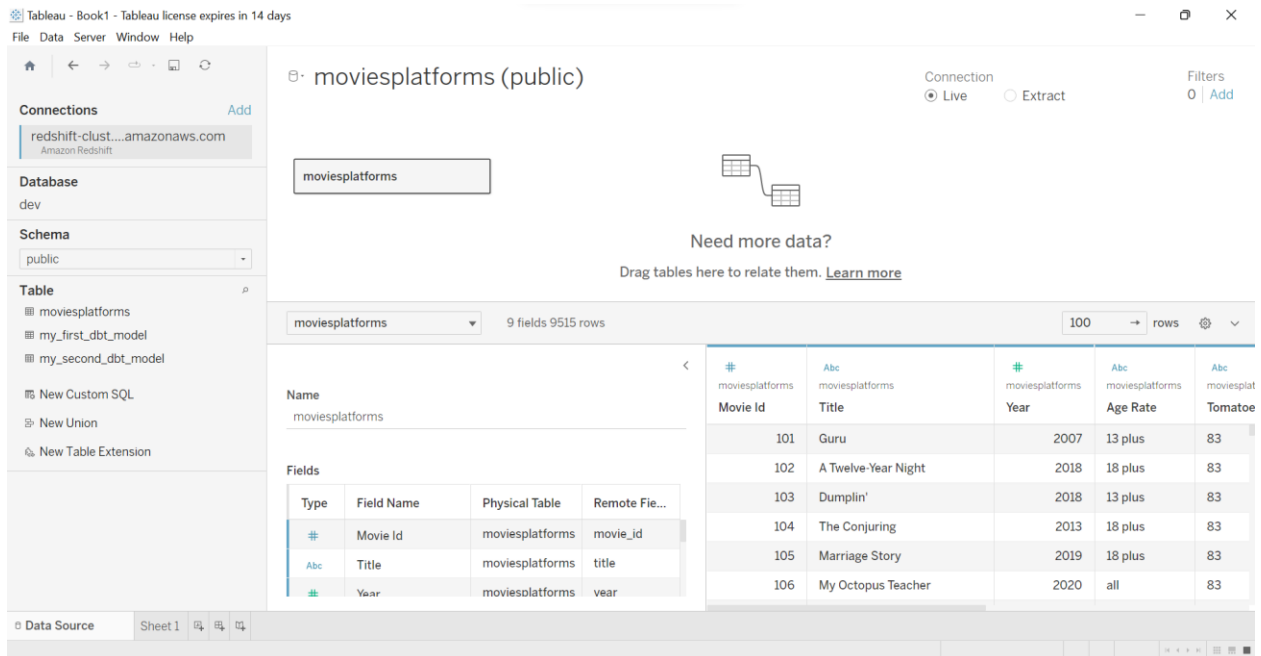
Vertica

Web Data Connector

Other Databases (JDBC)

Other Databases (ODBC)

Additional Connectors (24)



14. จัดทำ Data Visualization และรวบรวมเป็น Dashboard เพื่อตอบคำถามจากปัญหาข้างต้น

