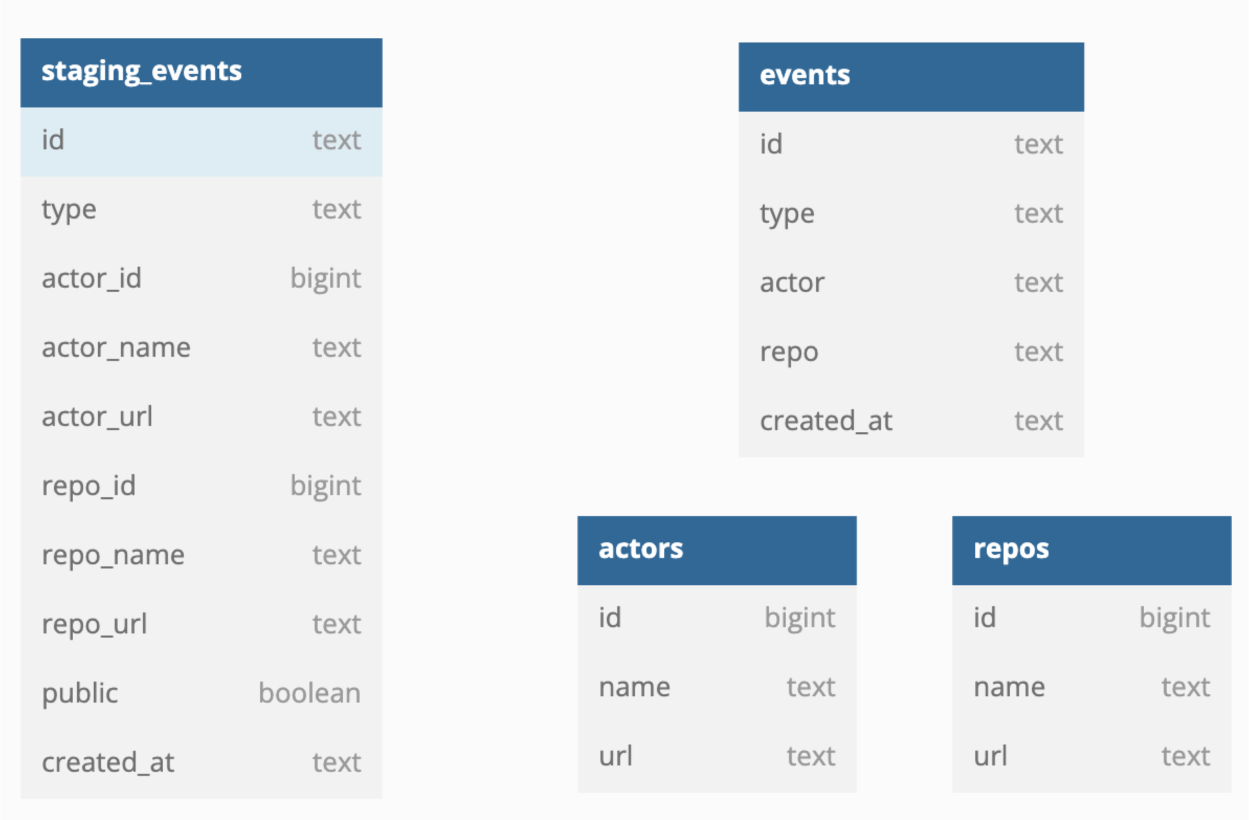
Week 3 – Project Building DWH

Data model



staging_events	
id	text
type	text
actor_id	bigint
actor_name	text
actor_url	text
repo_id	bigint
repo_name	text
repo_url	text
public	boolean
created_at	text

events	
id	text
type	text
actor	text
repo	text
created_at	text

actors	
id	bigint
name	text
url	text

repos	
id	bigint
name	text
url	text

Tables

staging_events:

This table is served a bulk batch loading which inject the data from json file in AWS S3 into AWS Redshift.

The table appears those mandatory columns for each data instance.

events:

This table collects the information related to event and only loads for non-existing data (validate on event id).

actors:

This table collects the information related to actor and only loads for non-existing data (validate on actor id).

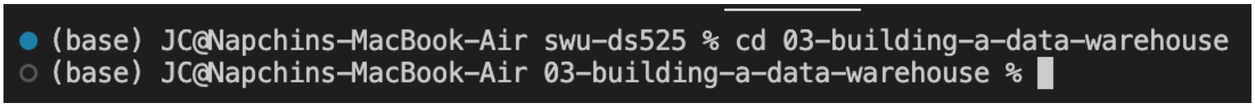
repos:

This table collects the information related to repository and only loads for non-existing data (validate on repo id).

Project implementation instruction

1. Reach the project repository ‘swu-ds525/03-building-a-data-warehouse’

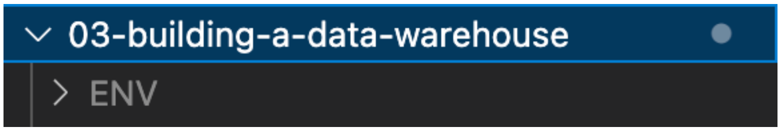
   : *$ cd 03-building-a-data-warehouse*



```
(base) JC@Napchins-MacBook-Air swu-ds525 % cd 03-building-a-data-warehouse
(base) JC@Napchins-MacBook-Air 03-building-a-data-warehouse %
```

2. Create visual environment for the project’s resources, named ‘ENV’ (only 1st time)

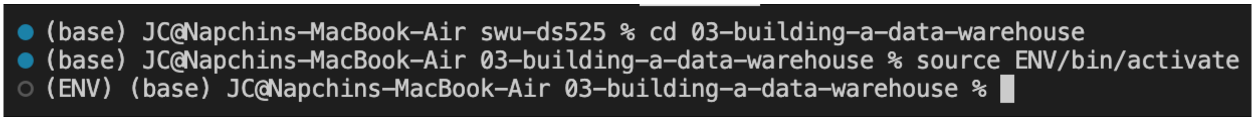
   : *$ python -m venv ENV*



```
⌄ 03-building-a-data-warehouse
  › ENV
```

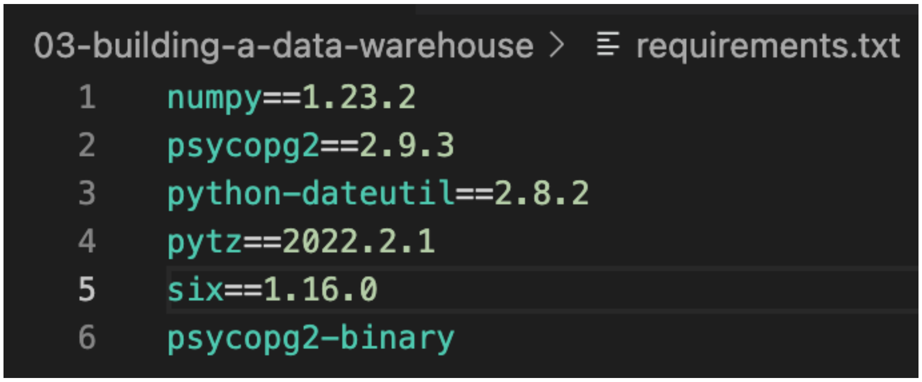
3. Activate visual environment ‘ENV’ to be used

   : *$ source ENV/bin/activate*



```
(base) JC@Napchins-MacBook-Air swu-ds525 % cd 03-building-a-data-warehouse
(base) JC@Napchins-MacBook-Air 03-building-a-data-warehouse % source ENV/bin/activate
(ENV) (base) JC@Napchins-MacBook-Air 03-building-a-data-warehouse %
```

4. Install mandatory libraries from configuration file, named ‘requirements.txt’ (only 1st time)



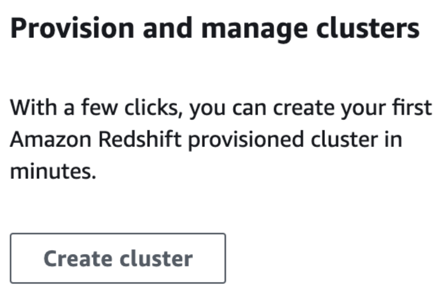
03-building-a-data-warehouse › requirements.txt

```
numpy==1.23.2
psycopg2==2.9.3
python-dateutil==2.8.2
pytz==2022.2.1
six==1.16.0
psycopg2-binary
```

   : *$ pip install -r requirements.txt*

DS525 - Chin Lertvipada - 64199130039

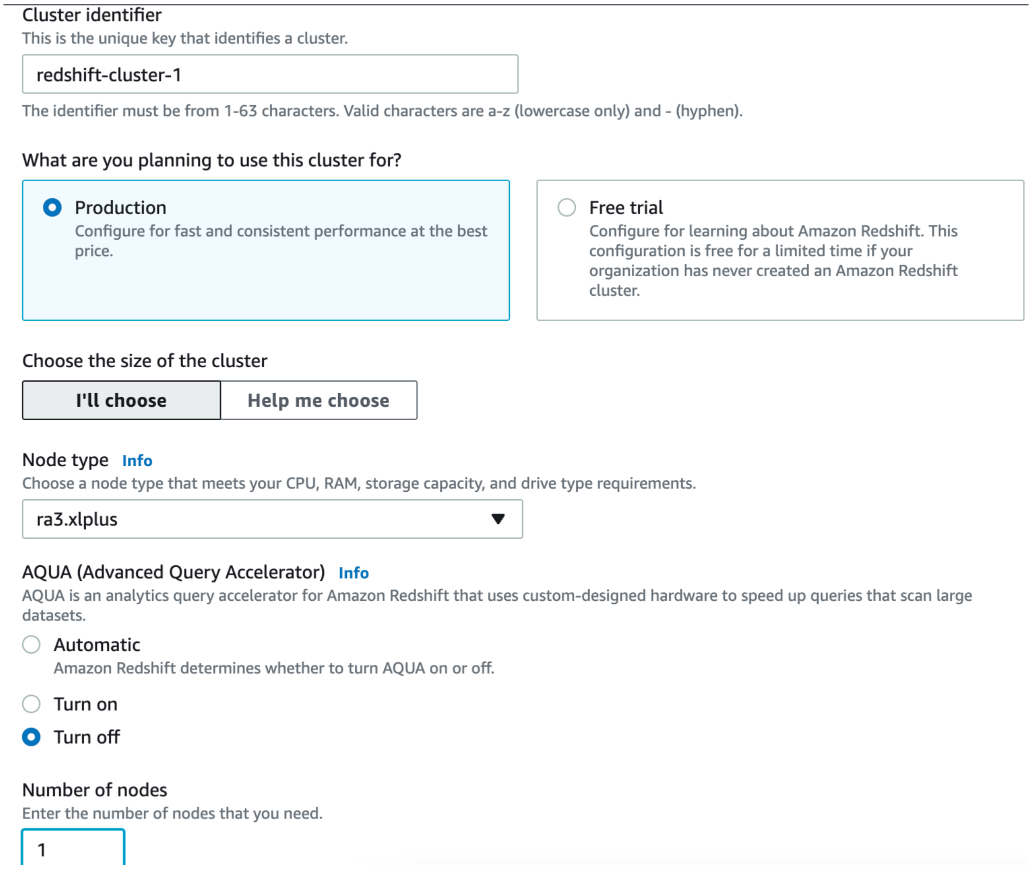
5. Create AWS Redshift cluster
    a. Create cluster



Provision and manage clusters

With a few clicks, you can create your first Amazon Redshift provisioned cluster in minutes.

    b. Fill information
        i. Cluster identification : redshift-cluster-1
        ii. Cluster for? : Production
        iii. Node type : ra3.xlplus
        iv. AQUA : Turn off
        v. Number of nodes : 1



Cluster identifier
This is the unique key that identifies a cluster.

redshift-cluster-1

The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

What are you planning to use this cluster for?

● Production
Configure for fast and consistent performance at the best price.

○ Free trial
Configure for learning about Amazon Redshift. This configuration is free for a limited time if your organization has never created an Amazon Redshift cluster.

Choose the size of the cluster

Node type Info
Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

ra3.xlplus ▼

AQUA (Advanced Query Accelerator) Info
AQUA is an analytics query accelerator for Amazon Redshift that uses custom-designed hardware to speed up queries that scan large datasets.

○ Automatic
Amazon Redshift determines whether to turn AQUA on or off.

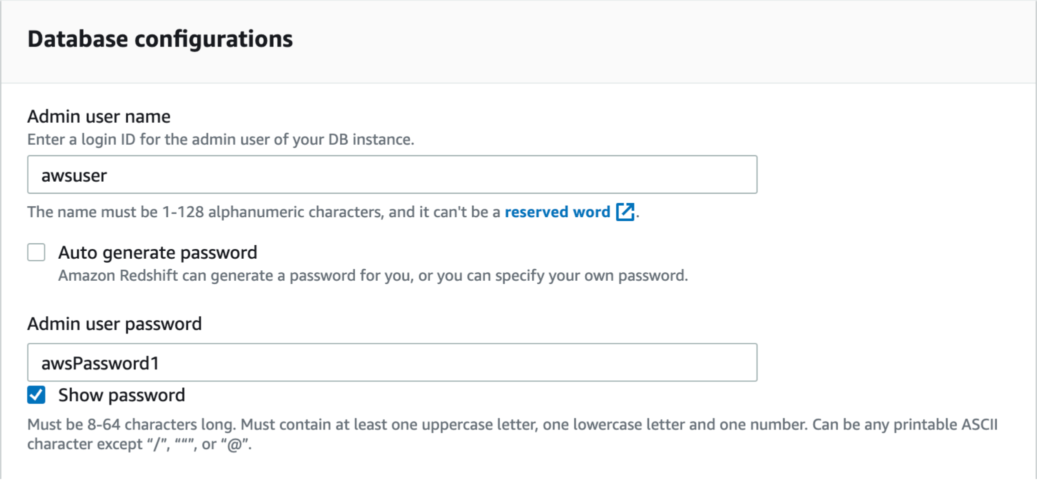
○ Turn on

● Turn off

Number of nodes
Enter the number of nodes that you need.

1

vi. Define Database username and password



Database configurations

Admin user name

Enter a login ID for the admin user of your DB instance.

awsuser

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.

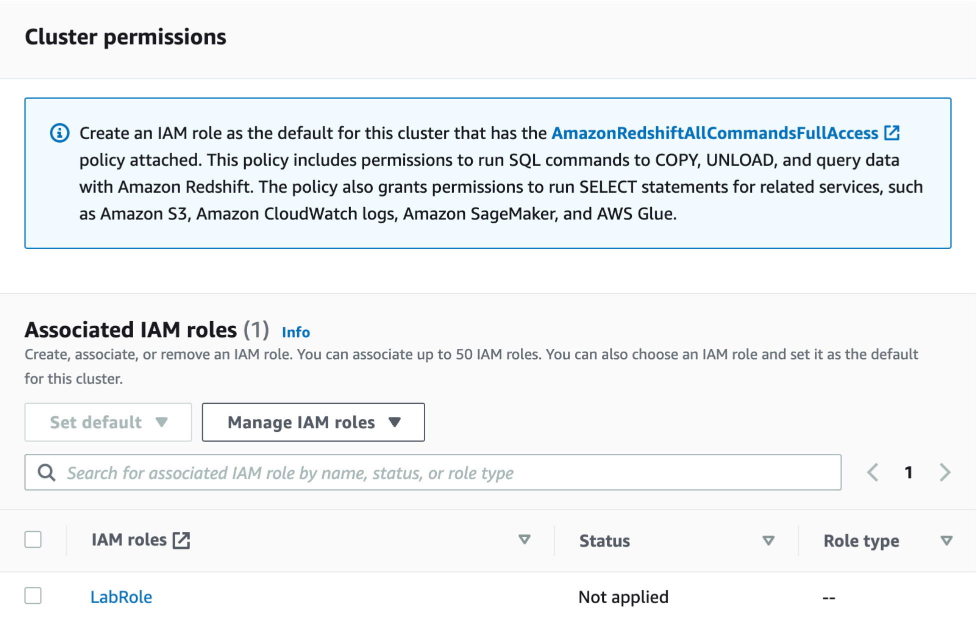
Admin user password

awsPassword1

☑ Show 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 “@”.

vii. Cluster permission : LabRole



Cluster permissions

Create an IAM role as the default for this cluster that has the **AmazonRedshiftAllCommandsFullAccess** policy attached. This policy includes permissions to run SQL commands to COPY, UNLOAD, and query data with Amazon Redshift. The policy also grants permissions to run SELECT statements for related services, such as Amazon S3, Amazon CloudWatch logs, Amazon SageMaker, and AWS Glue.

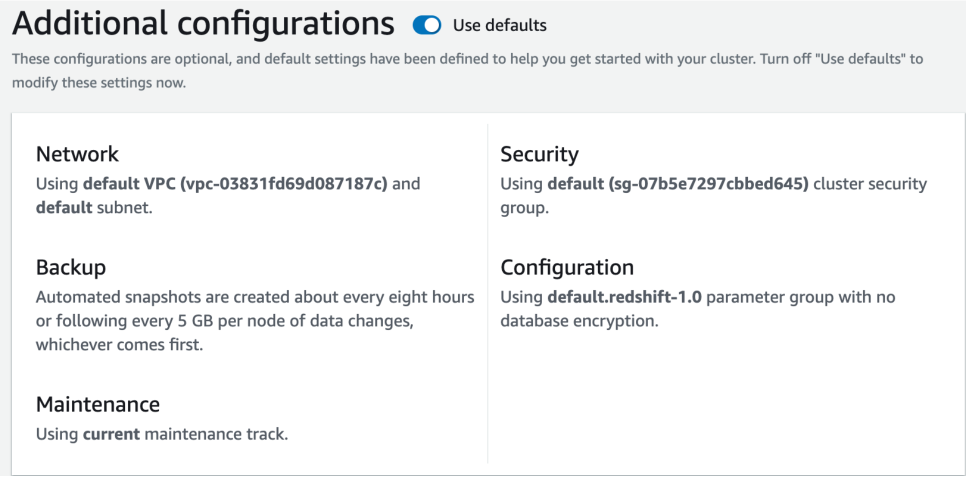
Associated IAM roles (1) Info

Create, associate, or remove an IAM role. You can associate up to 50 IAM roles. You can also choose an IAM role and set it as the default for this cluster.

Set default ▼
Manage IAM roles ▼
Search for associated IAM role by name, status, or role type
1

☐	IAM roles	Status	Role type
☐	LabRole	Not applied	--

viii. Remaining can keep the ‘default’



Additional configurations

● Use defaults

These configurations are optional, and default settings have been defined to help you get started with your cluster. Turn off "Use defaults" to modify these settings now.

Network

Using **default VPC (vpc-03831fd69d087187c)** and **default** subnet.

Security

Using **default (sg-07b5e7297cbbed645)** cluster security group.

Backup

Automated snapshots are created about every eight hours or following every 5 GB per node of data changes, whichever comes first.

Configuration

Using **default.redshift-1.0** parameter group with no database encryption.

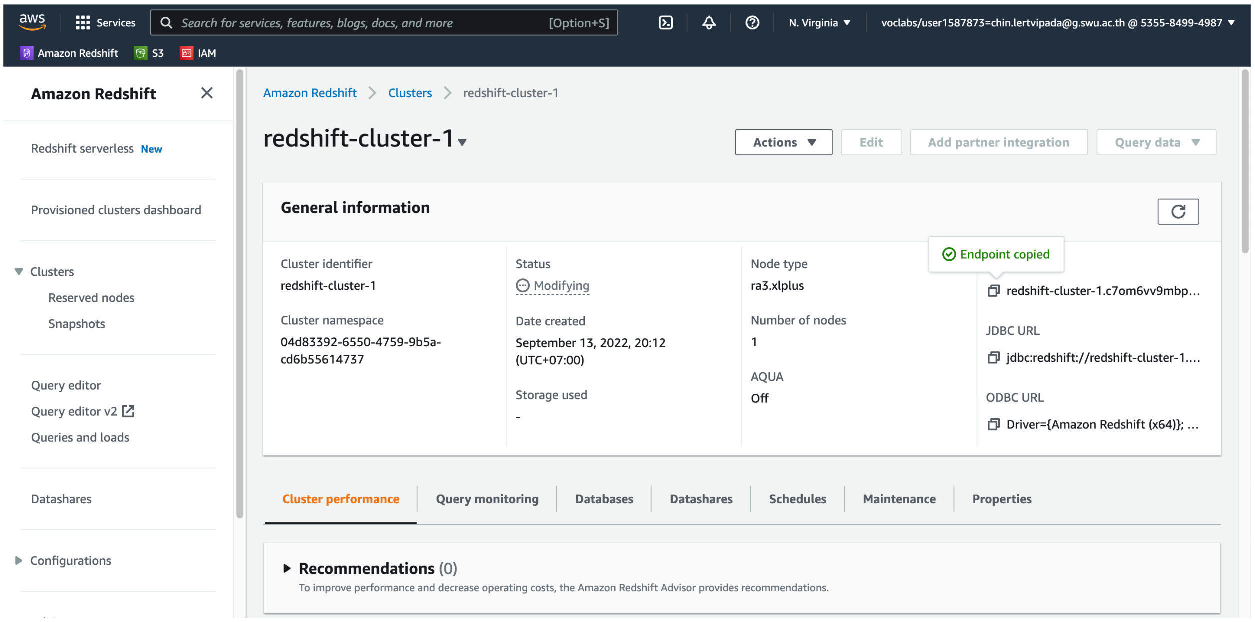
Maintenance

Using **current** maintenance track.

c. Wait until cluster is ready

d. Config cluster to enable public access

e. Wait until cluster is ready to use



redshift-cluster-1

General information

Endpoint copied

Cluster identifier	Status	Node type	
redshift-cluster-1	Modifying	ra3.xlplus	redshift-cluster-1.c7om6vv9mbp…
Cluster namespace	Date created	Number of nodes	JDBC URL
04d83392-6550-4759-9b5a-cd6b55614737	September 13, 2022, 20:12 (UTC+07:00)	1	jdbc:redshift://redshift-cluster-1.…
	Storage used	AQUA	ODBC URL
	-	Off	Driver={Amazon Redshift (x64)}; …

Recommendations (0)

To improve performance and decrease operating costs, the Amazon Redshift Advisor provides recommendations.

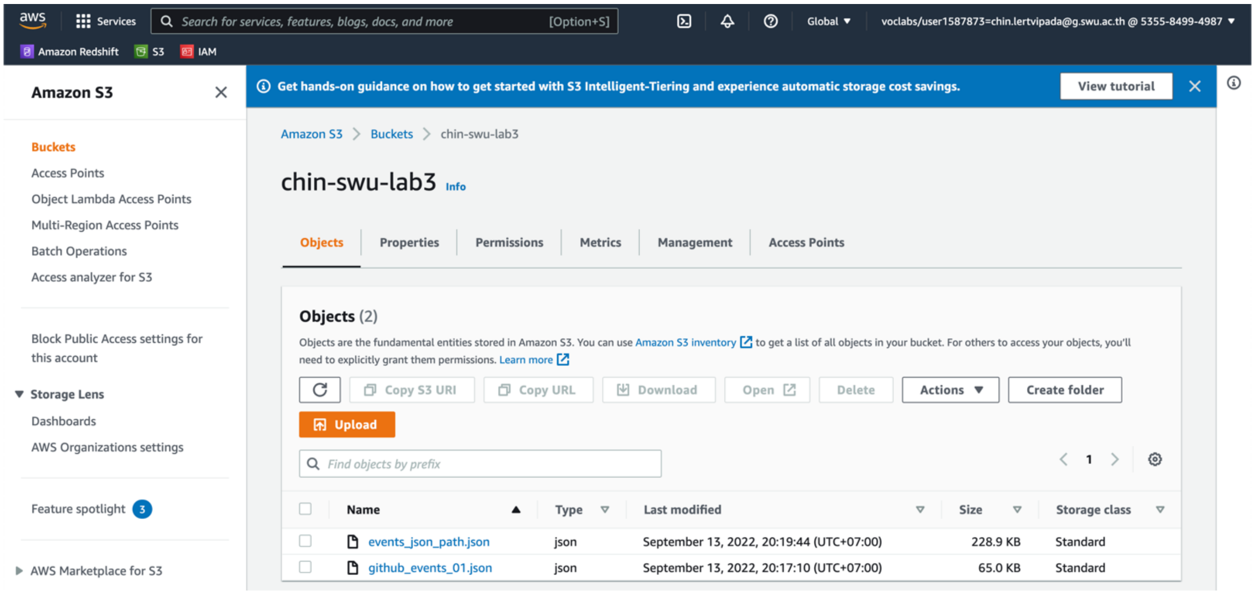
6. Upload data file and manifest file to AWS S3

   a. Create AWS S3 bucket with 'Full public access'

   b. Upload files

      i. Manifest file : events_json_path.json

      ii. Data file : github_events_01.json



Get hands-on guidance on how to get started with S3 Intelligent-Tiering and experience automatic storage cost savings.

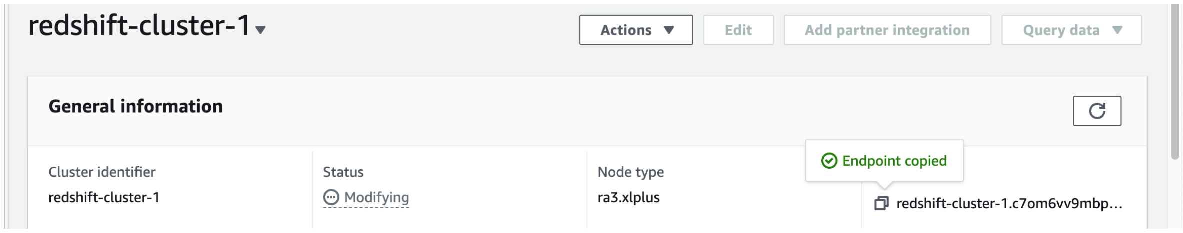
chin-swu-lab3

Objects (2)

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

	Name	Type	Last modified	Size	Storage class
□	events_json_path.json	json	September 13, 2022, 20:19:44 (UTC+07:00)	228.9 KB	Standard
□	github_events_01.json	json	September 13, 2022, 20:17:10 (UTC+07:00)	65.0 KB	Standard

7. Config ‘etl.py’ to connect to AWS Redshift
    a. Host : copy from AWS Redshift endpoint

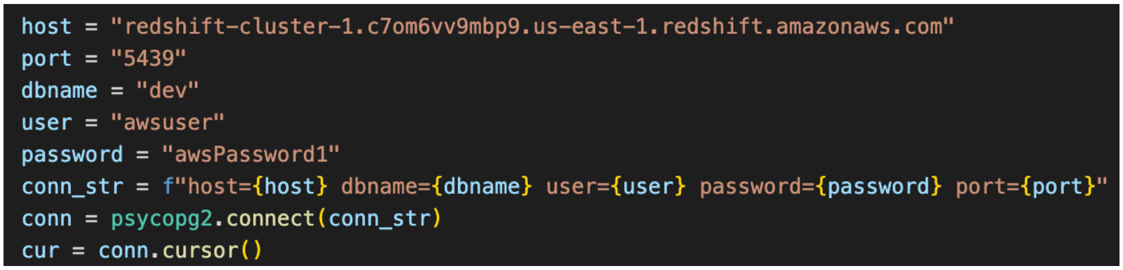


redshift-cluster-1

General information

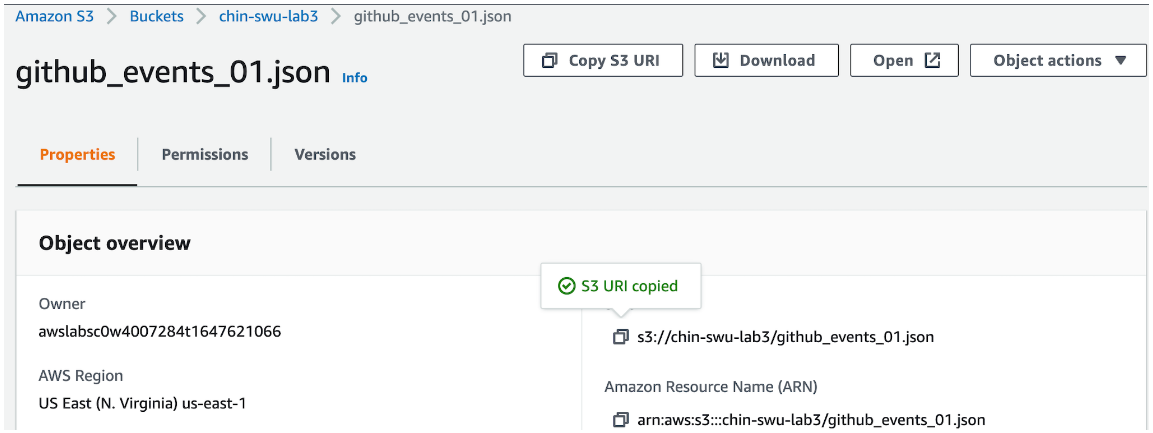
Cluster identifier: redshift-cluster-1
Status: Modifying
Node type: ra3.xlplus
redshift-cluster-1.c7om6vv9mbp...

    b. Port : 5439
    c. Dbname : dev
    d. User/Password : as define when create the cluster



```
host = "redshift-cluster-1.c7om6vv9mbp9.us-east-1.redshift.amazonaws.com"
port = "5439"
dbname = "dev"
user = "awsuser"
password = "awsPassword1"
conn_str = f"host={host} dbname={dbname} user={user} password={password} port={port}"
conn = psycopg2.connect(conn_str)
cur = conn.cursor()
```

8. Config ‘etl.py’ to copy the data from AWS S3 to AWS Redshift
    a. From : the URI to data file

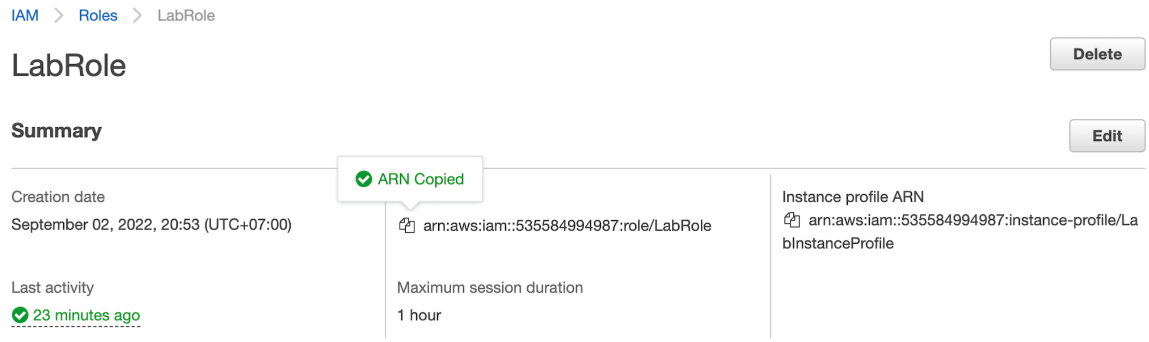


github_events_01.json

Object overview

Owner: awslabsc0w4007284t1647621066
AWS Region: US East (N. Virginia) us-east-1
s3://chin-swu-lab3/github_events_01.json
Amazon Resource Name (ARN): arn:aws:s3:::chin-swu-lab3/github_events_01.json

    b. Credentials : the ARN of LabRole

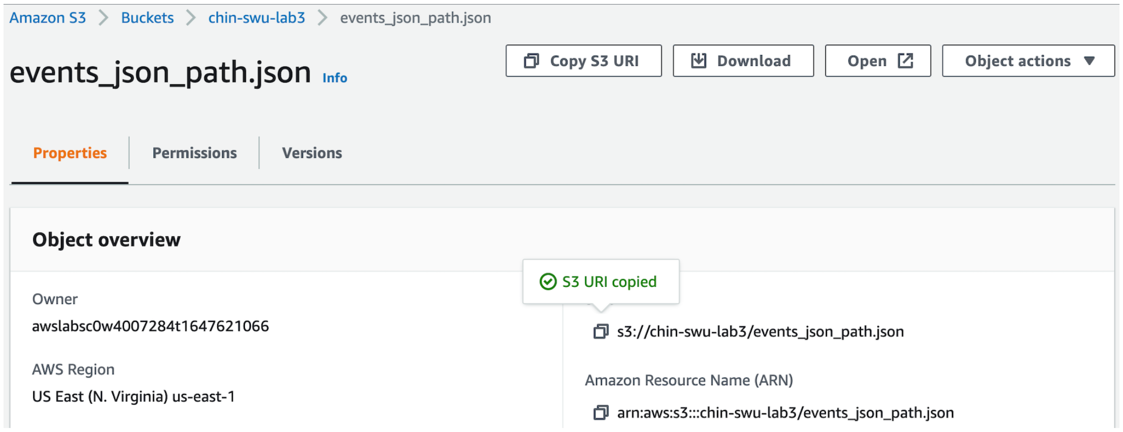


LabRole

Summary

Creation date: September 02, 2022, 20:53 (UTC+07:00)
arn:aws:iam::535584994987:role/LabRole
Instance profile ARN: arn:aws:iam::535584994987:instance-profile/LabInstanceProfile
Last activity: 23 minutes ago
Maximum session duration: 1 hour

c. Json : the URI to manifest file



events_json_path.json Info

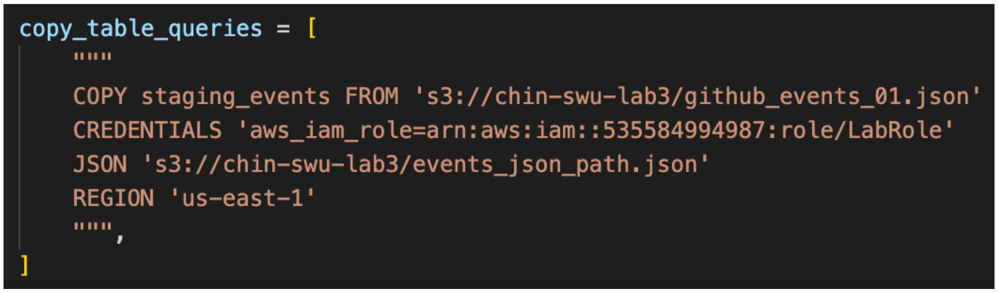
Object overview

Owner
awslabsc0w4007284t1647621066

AWS Region
US East (N. Virginia) us-east-1

s3://chin-swu-lab3/events_json_path.json

Amazon Resource Name (ARN)
arn:aws:s3:::chin-swu-lab3/events_json_path.json



```
copy_table_queries = [
    """
    COPY staging_events FROM 's3://chin-swu-lab3/github_events_01.json'
    CREDENTIALS 'aws_iam_role=arn:aws:iam::535584994987:role/LabRole'
    JSON 's3://chin-swu-lab3/events_json_path.json'
    REGION 'us-east-1'
    """,
]
```

9. Create tables, Inject data from S3 to Redshift, Insert data, Query data thru python script, named ‘etl.py’
   : *$ python etl.py*

10. Check the data in cluster by ‘query editor’

11. [optional] Shutdown the environment
    a. Deactivate the visual environment ‘ENV’
       : *$ deactivate*
    b. Delete the AWS Redshift cluster
    c. Delete the files and bucket in AWS S3