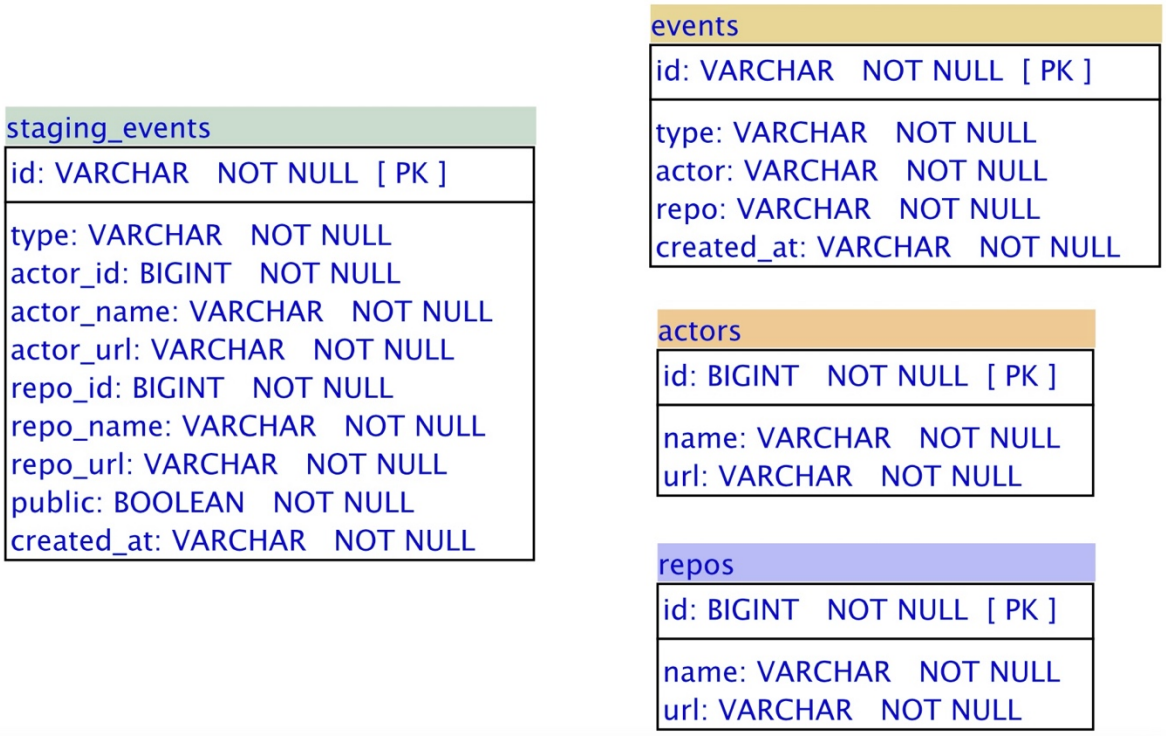
Week 3 : 03-building-a-data-warehouse

Data model



Project Processing

1. change directory to project 03-data-warehouse:

$ cd 03-building-a-data-warehouse

2. create visual environment named 'ENV' (only 1st time):

$ python -m venv ENV

3. activate the visual environment:

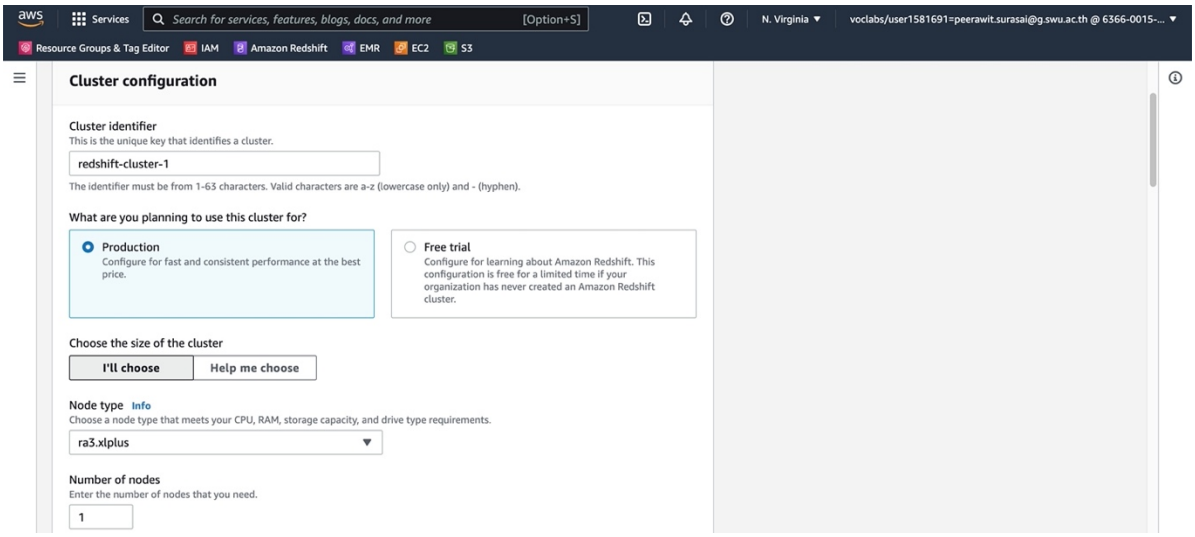
$ source ENV/bin/activate

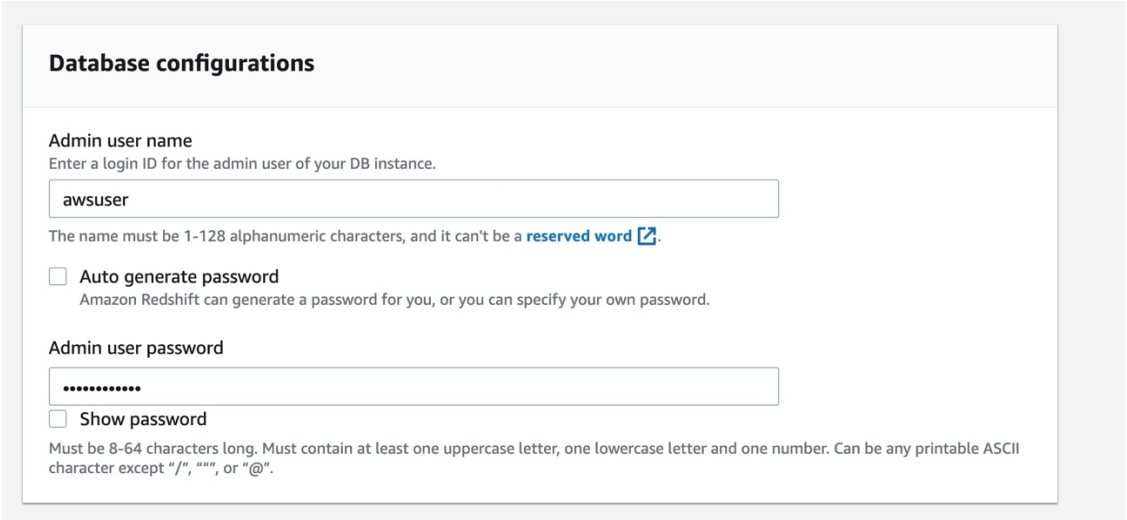
4. install required libraries from config file (only 1st time):

$ pip install -r requirements.txt

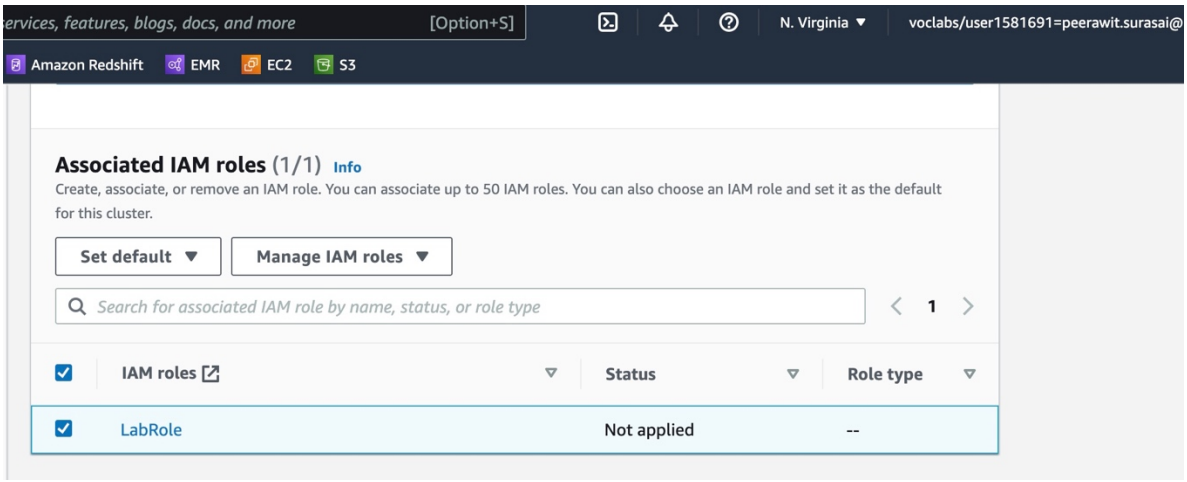
5. Create AWS Redshift cluster (with following config):

- 'Cluster identification' : redshift-cluster-1
- 'Cluster for' : Production
- 'Node type' : ra3.xlplus
- 'AQUA' : Turn off
- 'Number of nodes' : 1
- 'Database username' : awsuser
- 'Database password' : awsPassword1
- 'Cluster permission' : LabRole
- 'Remaining' : keep as default

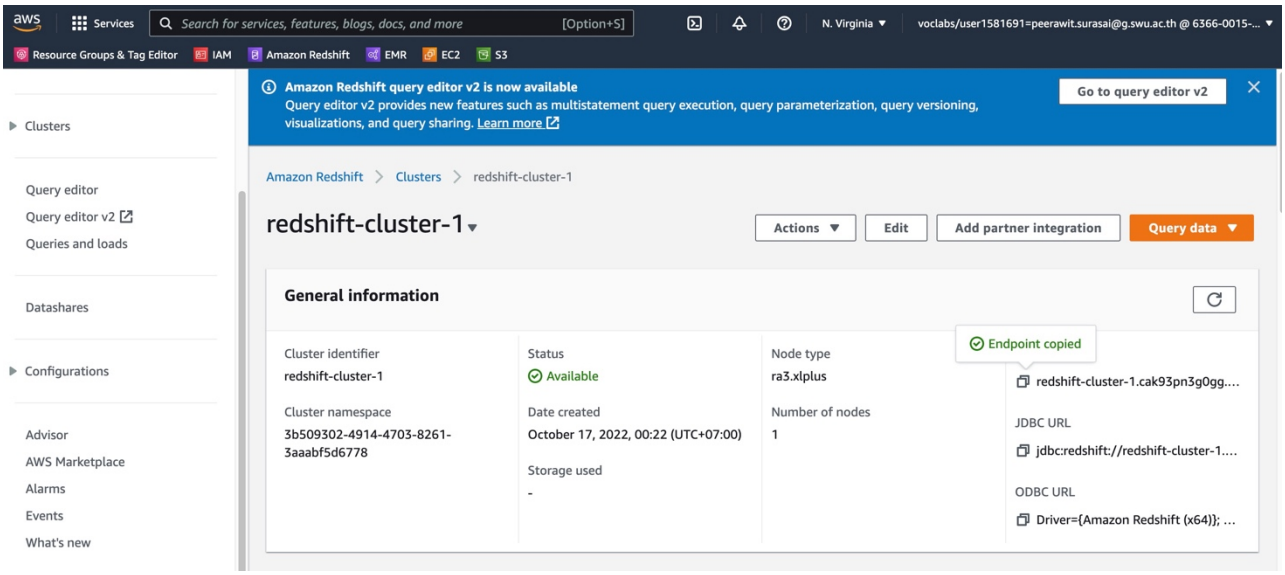




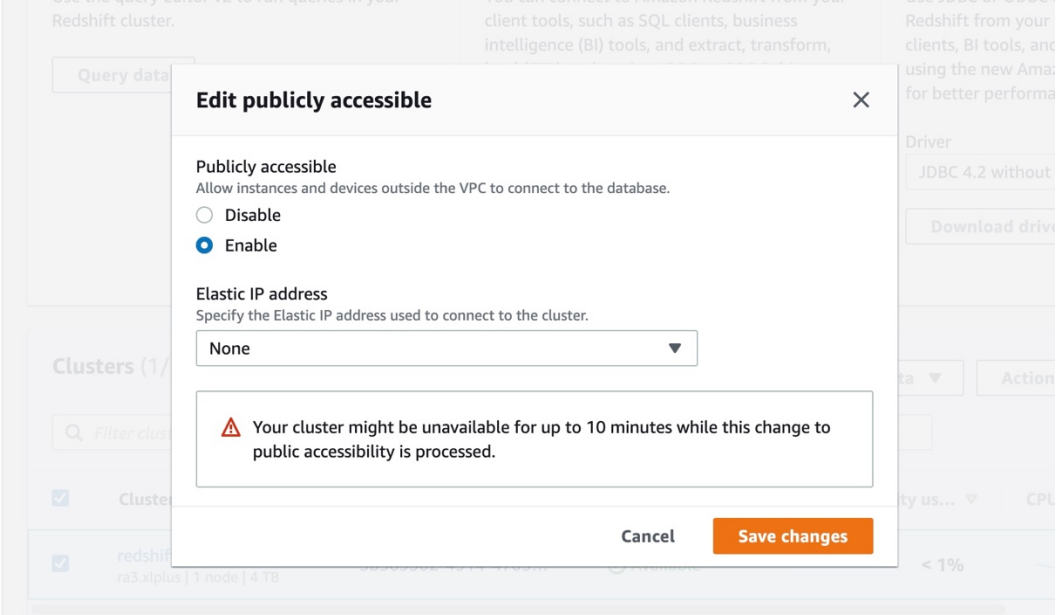
Set LabRole for Associated IAM roles



Redshift Cluster after creating



Set Enable publicly access for Redshift



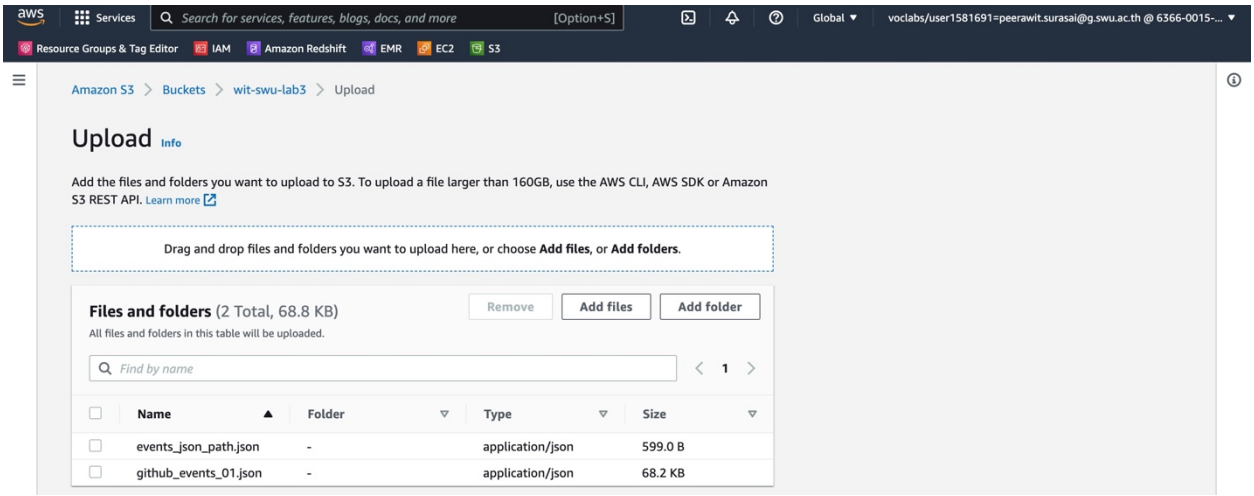
6. Upload data file and manifest file to AWS S3 :

   a. Create AWS S3 bucket with ‘Full public access’

   b. Upload files

   - Manifest file : events_json_path.json

   - Data file : github_events_01.json



7. Config ‘etl.py’ to connect to AWS Redshift:

   a. Host : copy from AWS Redshift endpoint

   b. Port : 5439

   c. Dbname : dev

   d. User/Password : as define when create the cluster

```
def main():
    host = "redshift-cluster-1.cak93pn3g0gg.us-east-1.redshift.amazonaws.com:5439/dev"
    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()
```

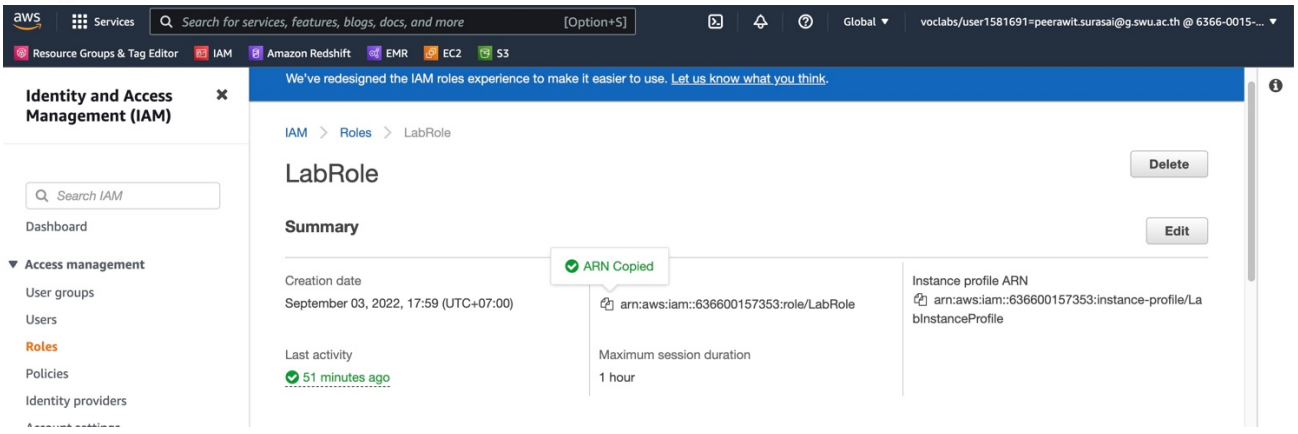
8. Config ‘etl.py’ to copy the data from AWS S3 to AWS Redshift:

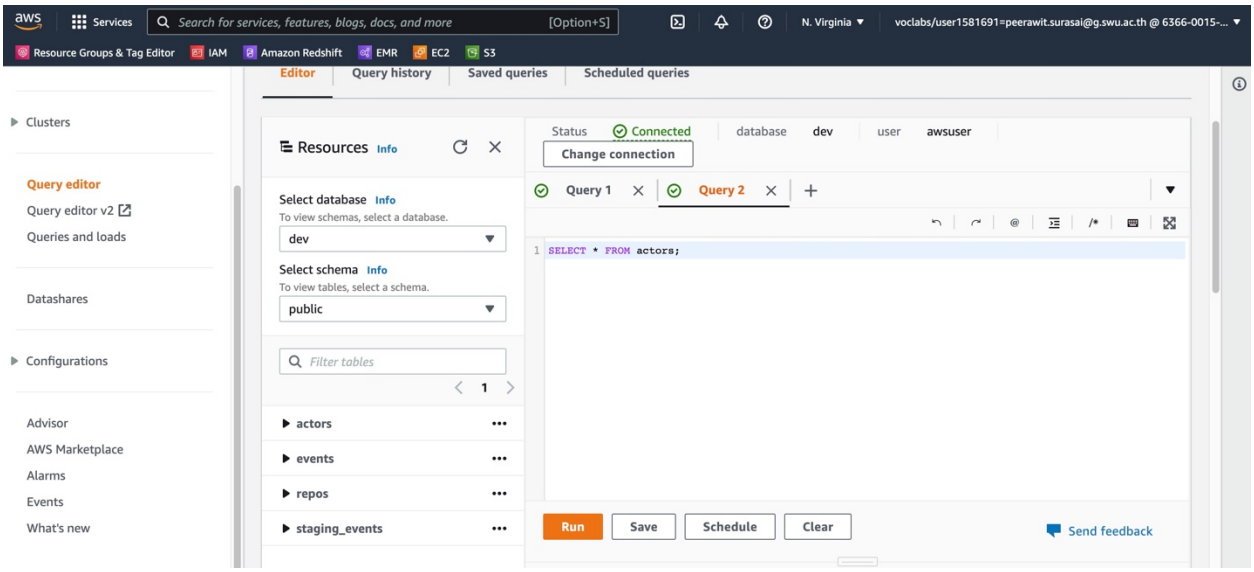
   a. From : the URI to data file

   b. Credentials : the ARN of LabRole

   c. Json : the URI to manifest file

```
copy_table_queries = [
    """
    COPY staging_events FROM 's3://wit-swu-lab3/github_events_01.json'
    CREDENTIALS 'aws_iam_role=arn:aws:iam::636600157353:role/LabRole'
    JSON 's3://wit-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’ (exported as csv):

[staging_events]

[events]

[actors]

[repos]

Shutdown steps

11. deactivate the visual environment:

$ deactivate

12. Delete the AWS Redshift cluster

13. Delete the files and bucket in AWS S3