

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

### Data model

#### staging\_events

id: VARCHAR NOT NULL [ PK ]
type: VARCHAR NOT NULL actor_id: BIGINT NOT NULL actor_name: VARCHAR NOT NULL actor_url: VARCHAR NOT NULL repo_id: BIGINT NOT NULL repo_name: VARCHAR NOT NULL repo_url: VARCHAR NOT NULL public: BOOLEAN NOT NULL created_at: VARCHAR NOT NULL

#### events

id: VARCHAR NOT NULL [ PK ]
type: VARCHAR NOT NULL actor: VARCHAR NOT NULL repo: VARCHAR NOT NULL created_at: VARCHAR NOT NULL

#### actors

id: BIGINT NOT NULL [ PK ]
name: VARCHAR NOT NULL url: VARCHAR NOT NULL

#### repos

id: BIGINT NOT NULL [ PK ]
name: VARCHAR NOT NULL url: VARCHAR NOT NULL

### Project Processing

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

```
$ cd 03-building-a-data-warehouse
```

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

```
$ python -m venv ENV
```

3. activate the virtual 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

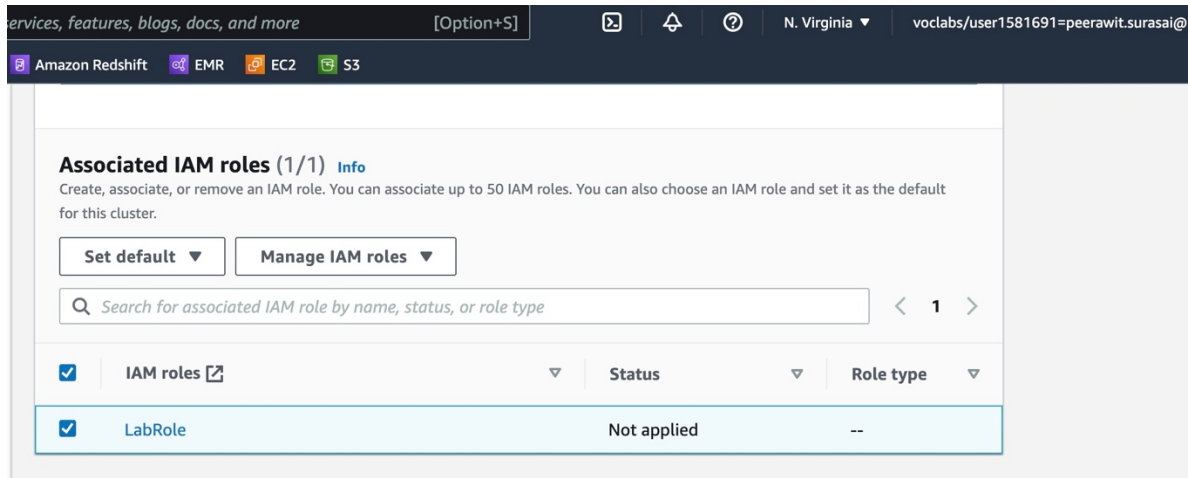
The screenshot shows the 'Cluster configuration' page in the AWS Management Console. The page is titled 'Cluster configuration' and includes a search bar at the top. The main content area is divided into several sections:

- Cluster identifier**: A text input field containing 'redshift-cluster-1'. Below it, a note states: '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?**: Two radio button options are present. The 'Production' option is selected, with a description: 'Configure for fast and consistent performance at the best price.' The 'Free trial' option is unselected, with a description: '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**: Two buttons are shown: 'I'll choose' (selected) and 'Help me choose'.
- Node type**: A dropdown menu is set to 'ra3.xlplus'. A note below says: 'Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.'
- Number of nodes**: A text input field contains the value '1'. A note below says: 'Enter the number of nodes that you need.'

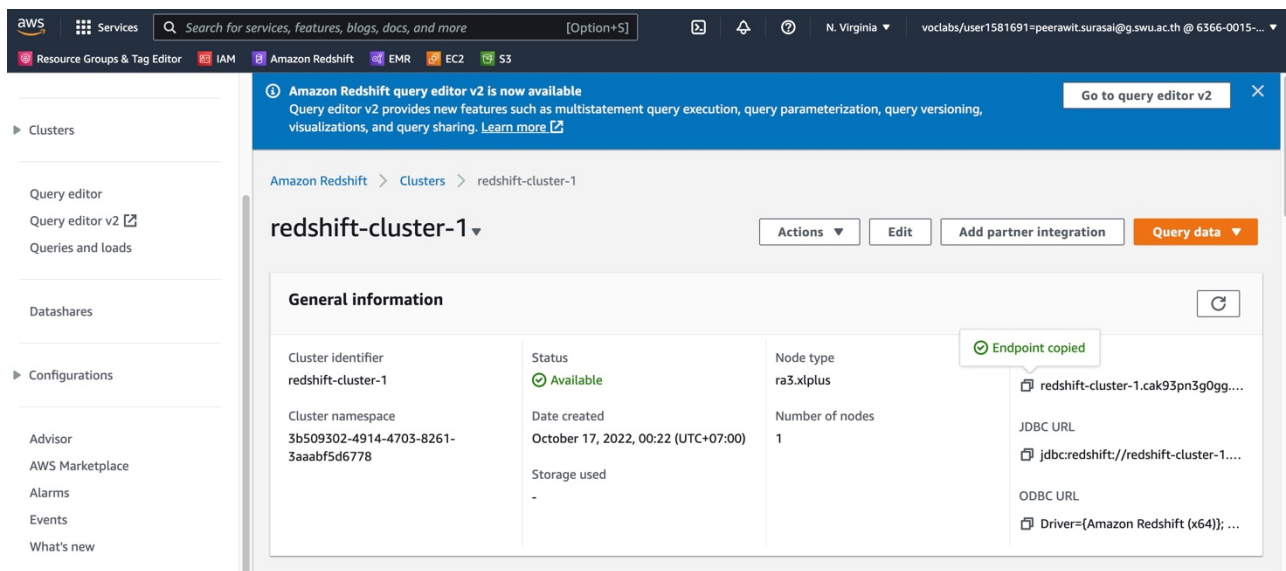
The screenshot shows the 'Database configurations' page in the AWS Management Console. The page is titled 'Database configurations' and includes a search bar at the top. The main content area is divided into several sections:

- Admin user name**: A text input field contains 'awsuser'. Below it, a note states: 'Enter a login ID for the admin user of your DB instance.' and 'The name must be 1-128 alphanumeric characters, and it can't be a reserved word'.
- Auto generate password**: An unchecked checkbox. Below it, a note states: 'Amazon Redshift can generate a password for you, or you can specify your own password.'
- Admin user password**: A text input field contains a masked password (represented by dots). Below it, an unchecked checkbox labeled 'Show password' is present. A note below states: '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 "@'.

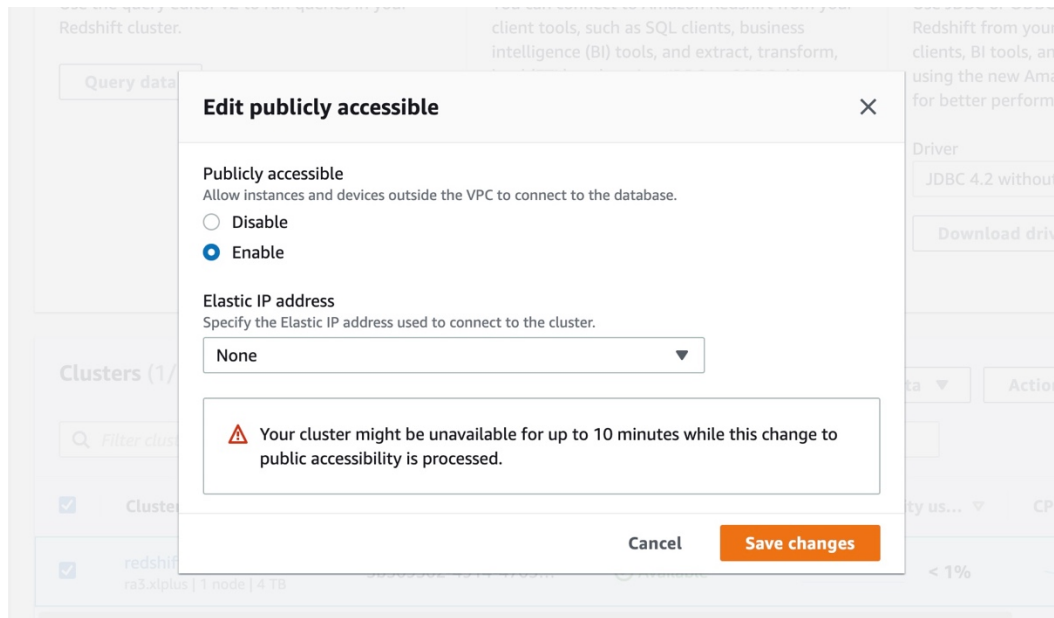
Set LabRole for Associated IAM roles



Redshift Cluster after creating

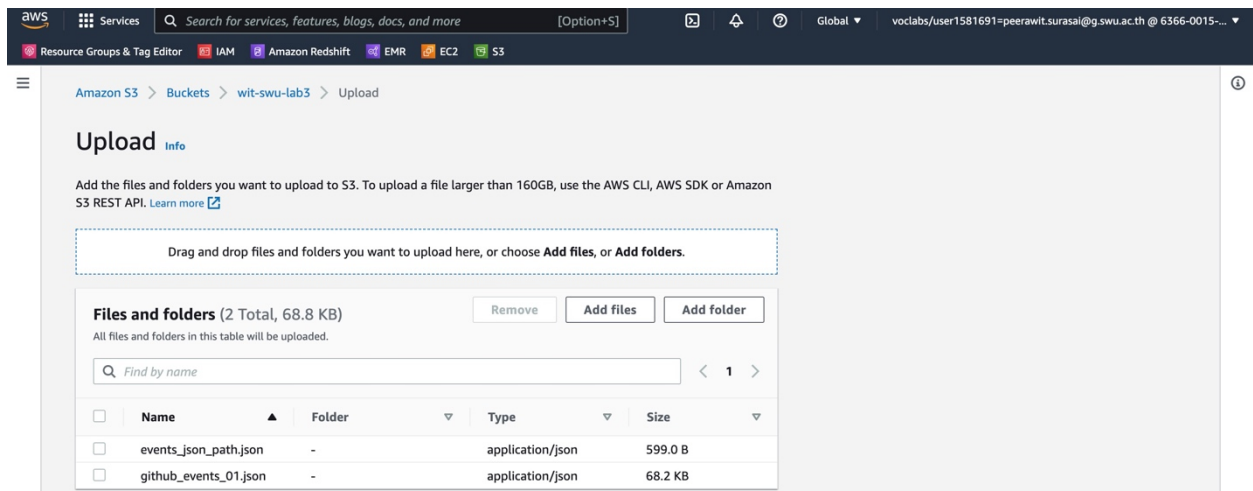


Set Enable publicly access for Redshift



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:

- Host : copy from AWS Redshift endpoint
- Port : 5439
- Dbname : dev
- User/Password : as define when create the cluster

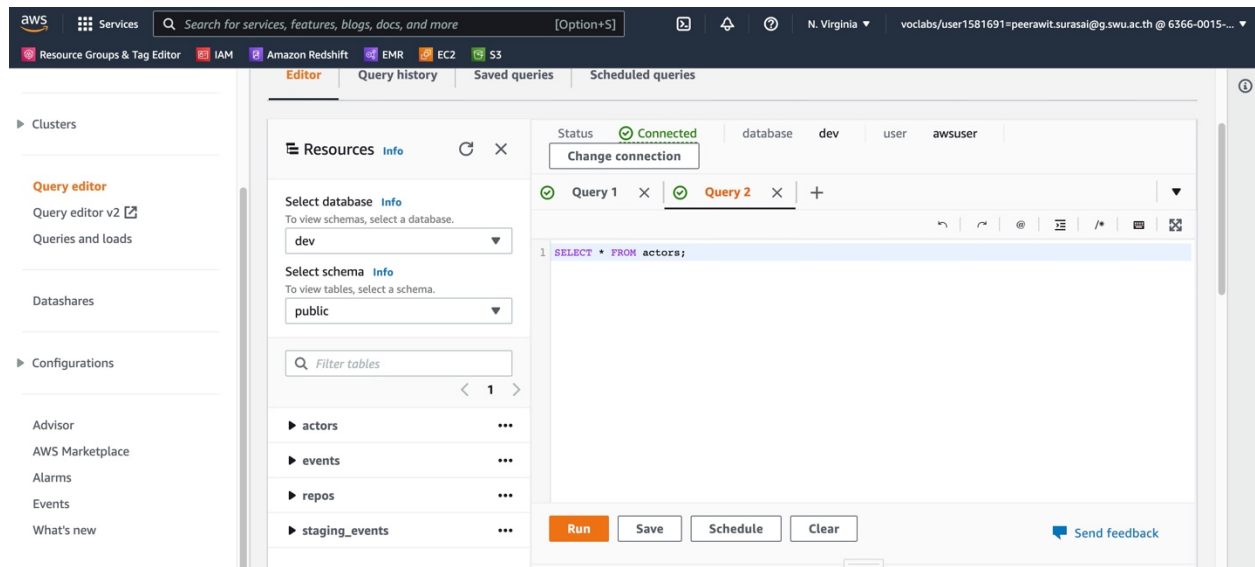
```
def main():  
    host = "redshift-cluster-1.cak93pn3g0gg.us-east-1.redshift.amazonaws.com:5439/dev"  
    port = "5439"  
    dbname = "dev"  
    user = "awsuser"  
    password = "awsPassword1"  
    conn_str = f"host={host} dbname={dbname} user={user} password={password} port={port}"  
    conn = psycopg2.connect(conn_str)  
    cur = conn.cursor()
```

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

- From : the URI to data file
- Credentials : the ARN of LabRole
- 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'  
    """,  
]
```

The screenshot shows the AWS IAM console interface. The left sidebar contains the 'Identity and Access Management (IAM)' menu with options like Dashboard, User groups, Users, Roles, Policies, and Identity providers. The main content area displays the details for the 'LabRole'. A 'Summary' section shows the creation date as 'September 03, 2022, 17:59 (UTC+07:00)' and the last activity as '51 minutes ago'. A 'Summary' table lists the 'Instance profile ARN' as 'arn:aws:iam::636600157353:instance-profile/LabRole' and the 'Maximum session duration' as '1 hour'. A green checkmark and 'ARN Copied' message are visible above the ARN field. The top navigation bar includes the AWS logo, a search bar, and the user's account information.



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