LAMBTON COLLEGE

A Report on
[Lab 4,5,6 on AWS Academy Data Analytics]

121 Brunel Rd, Mississauga

ON L4Z 3E9

A Group assignment with screenshots of Lab 4, 5, and 6

on Aws academy

Big Data Analytics DSMM

Under the supervision
of
Professor Teresa Zhu

Submitted BY:
Aadarsha Chapagain (C0825975)
Roshan Acharya (C0831342)
Anjana Kuriakose (C0829580)
Onyinye Mbanefo (C0831578)

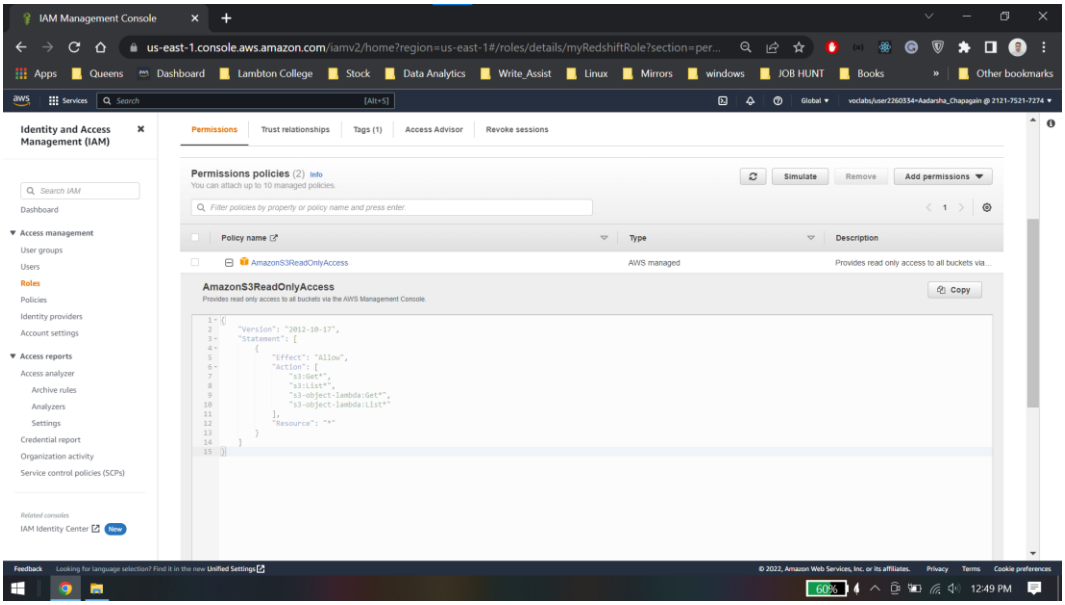
Submitted To:
Lambton College
Professor Teresa Zhu

Submission Date:
27th November 2022

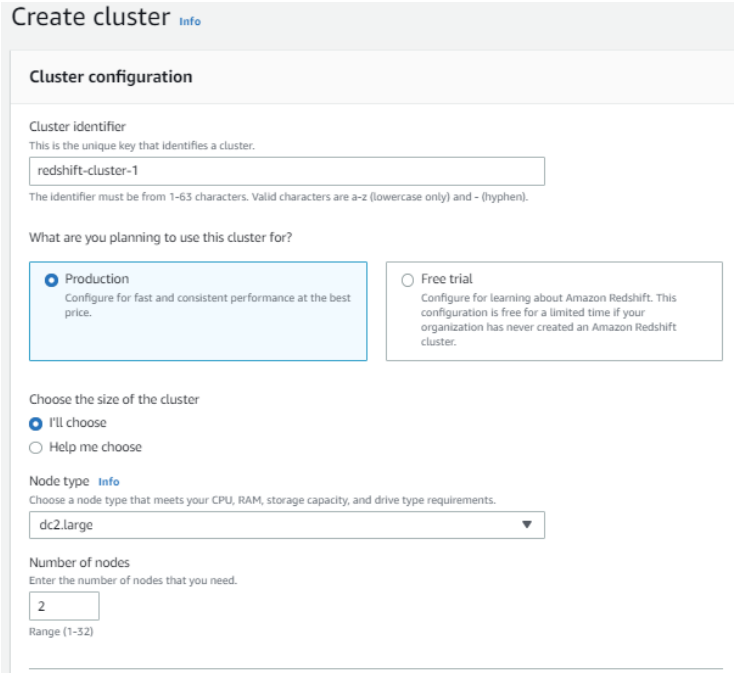
Lab4: Analyze Data with Amazon Redshift

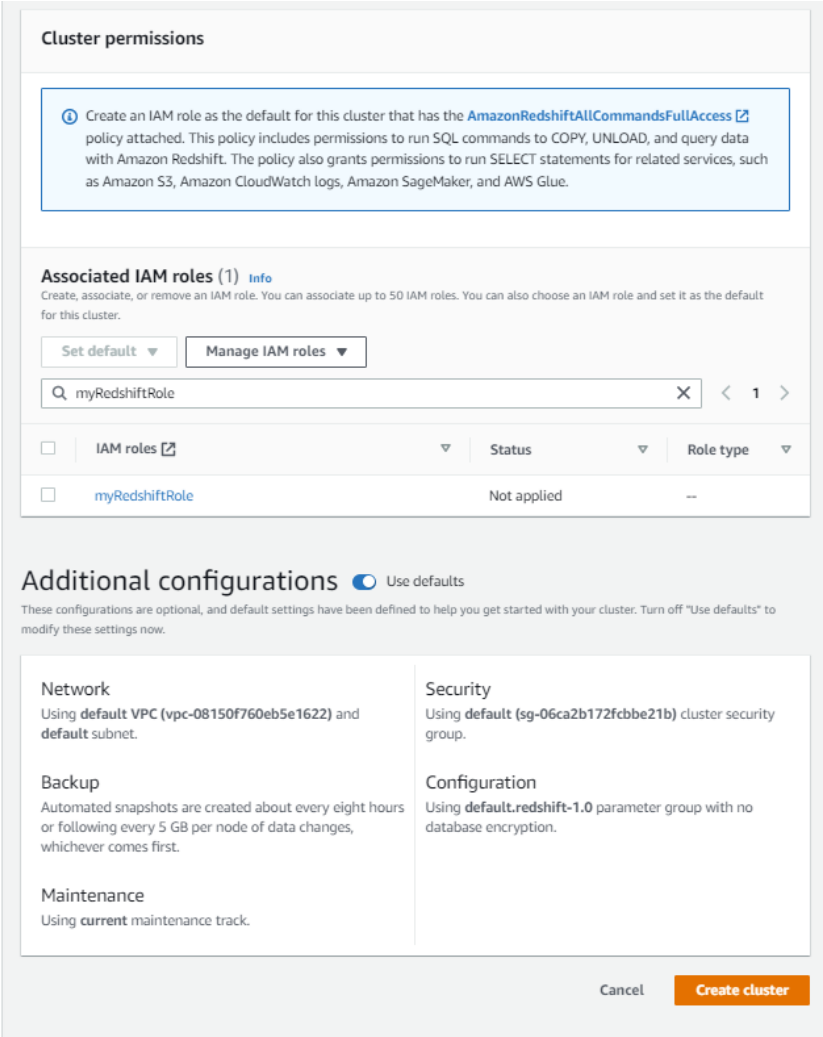
Task 1: Task 1: Review the security group for accessing the Amazon Redshift console

Myredshiftrole

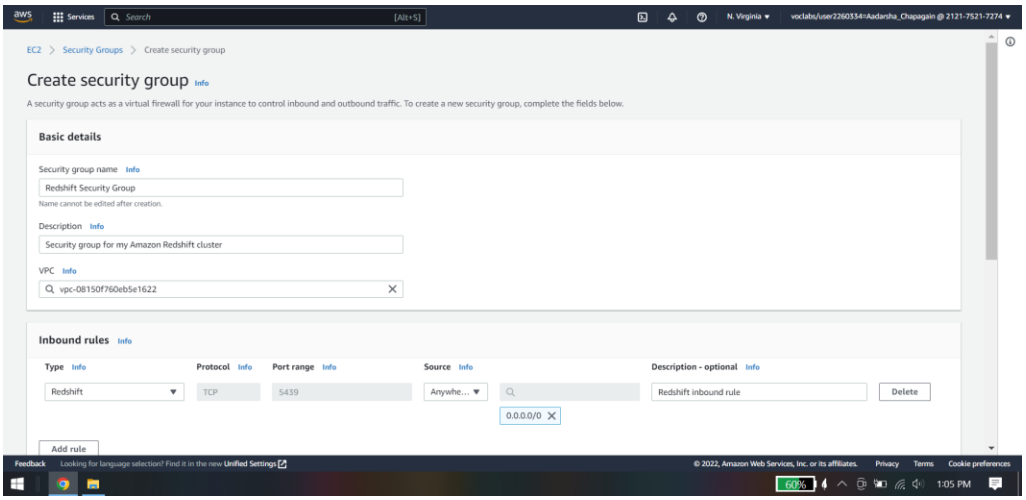


Task 2: Create and configure an Amazon Redshift cluster

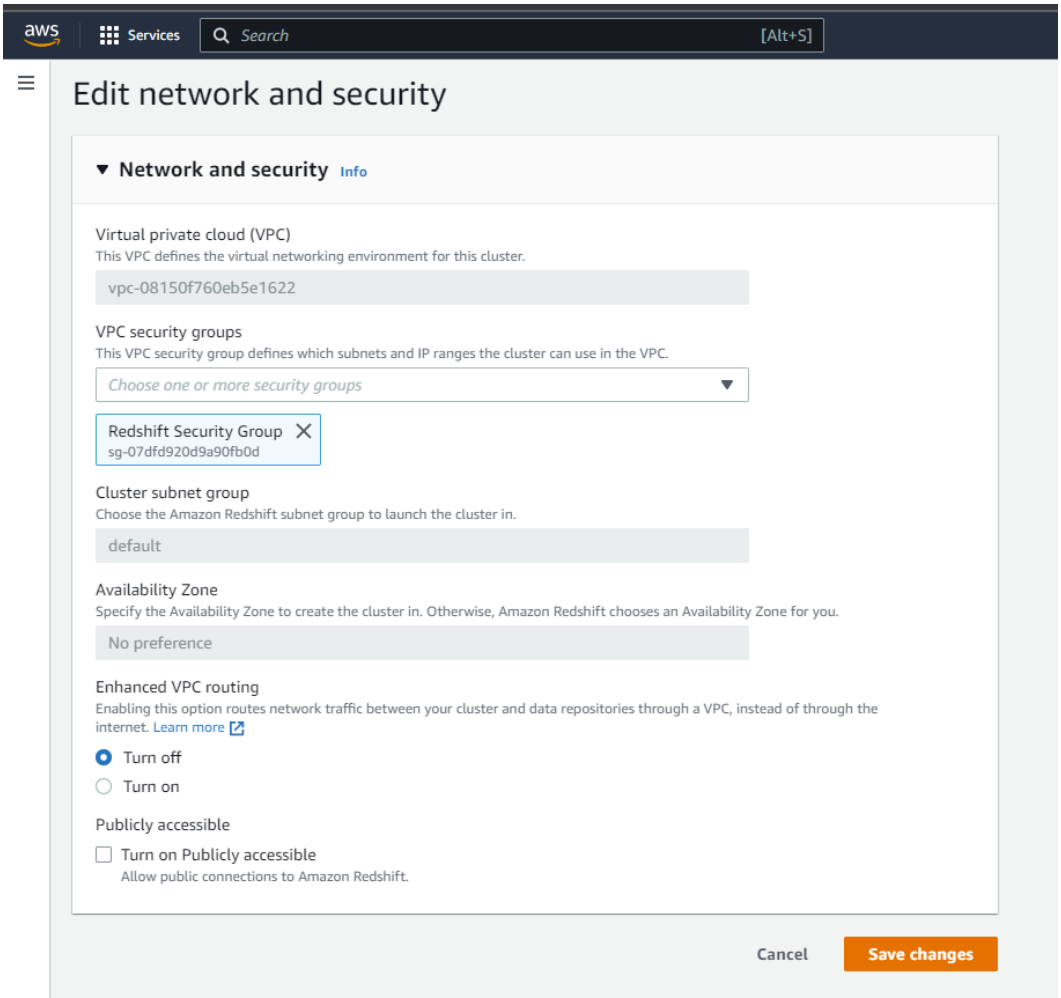




Task 2.1: Create a security group for your cluster



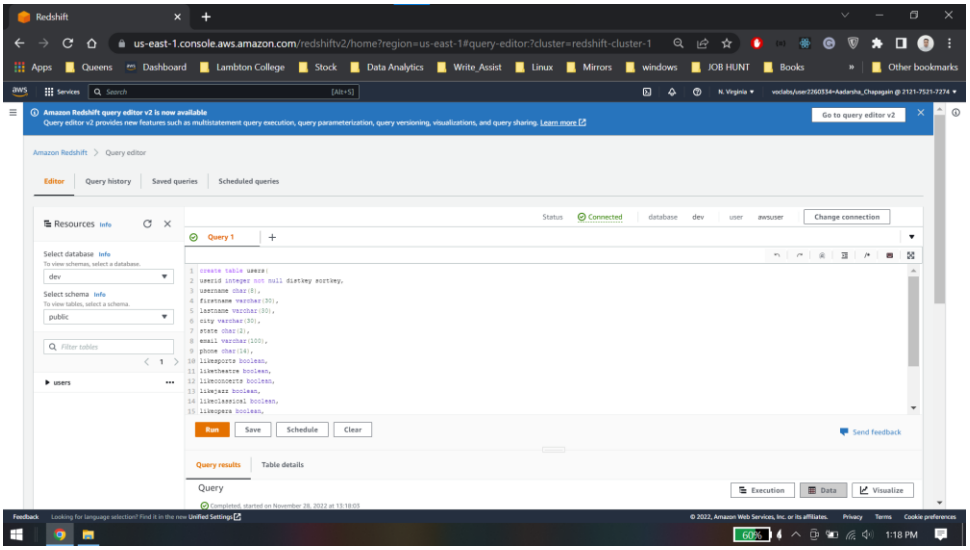
Task 2.2: Configure your Amazon Redshift cluster



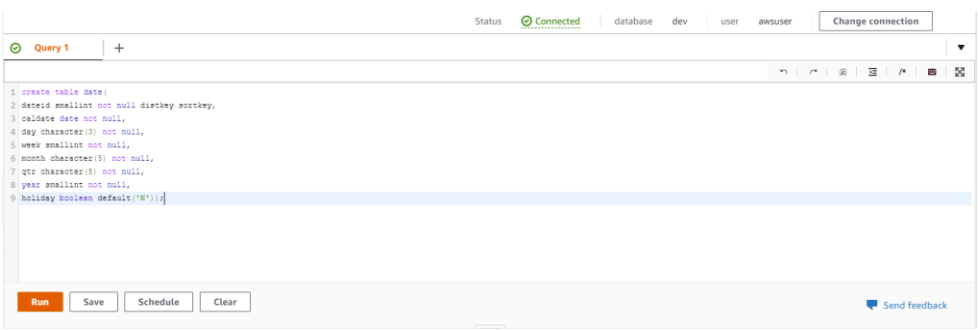
Task 3: Load data into your Amazon Redshift cluster

Task 3.1: Create the tables in the dev database

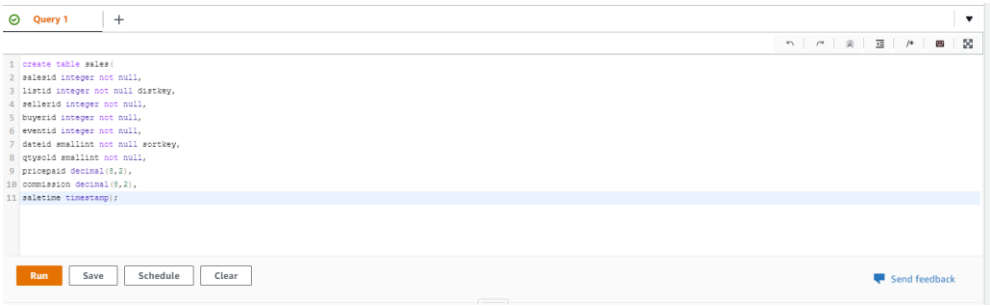
User table:



Date table



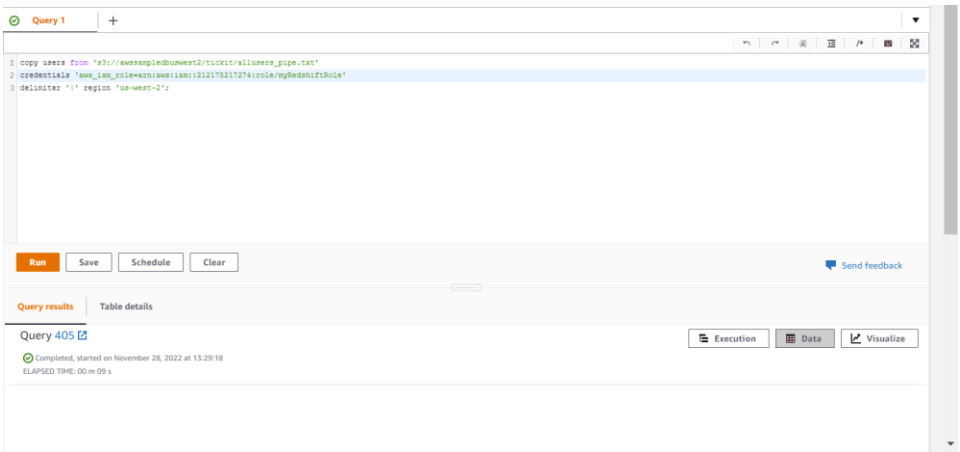
Sales table



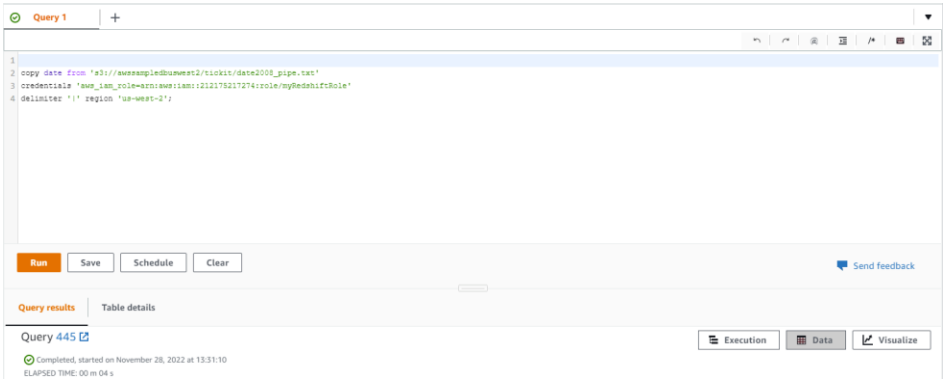
Task 3.2: Load data from Amazon S3

arn:aws:iam::212175217274:role/myRedshiftRole

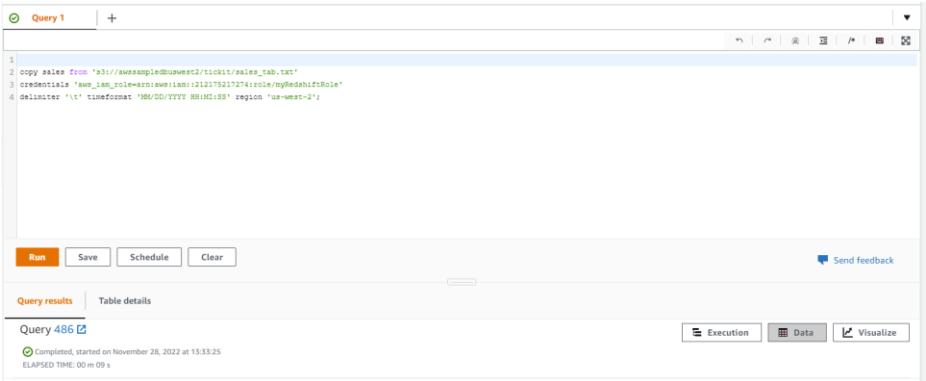
Copy users table



Copy date table

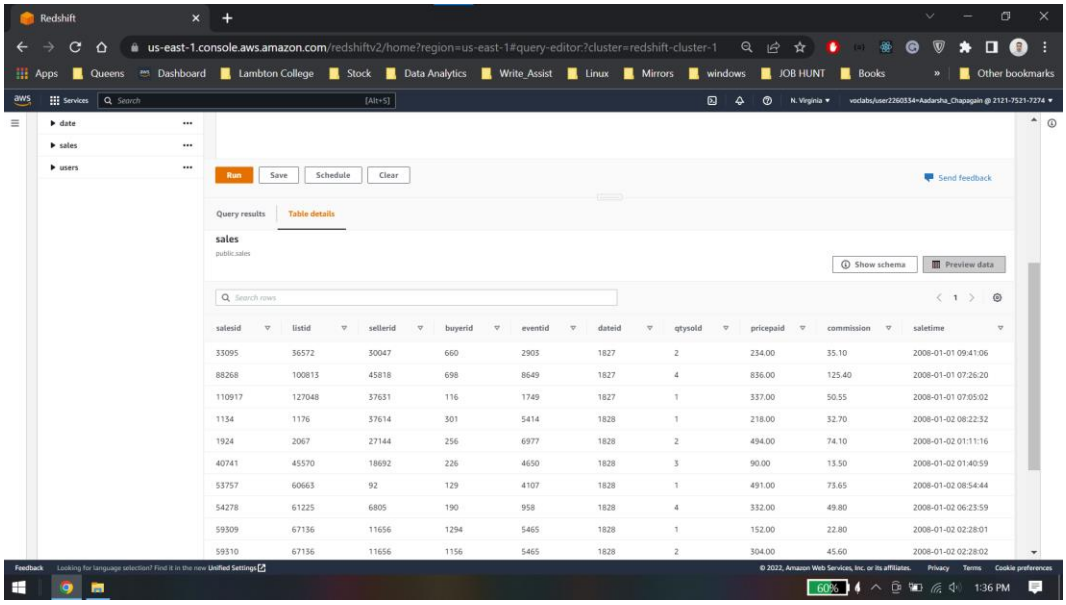


Copy sales table

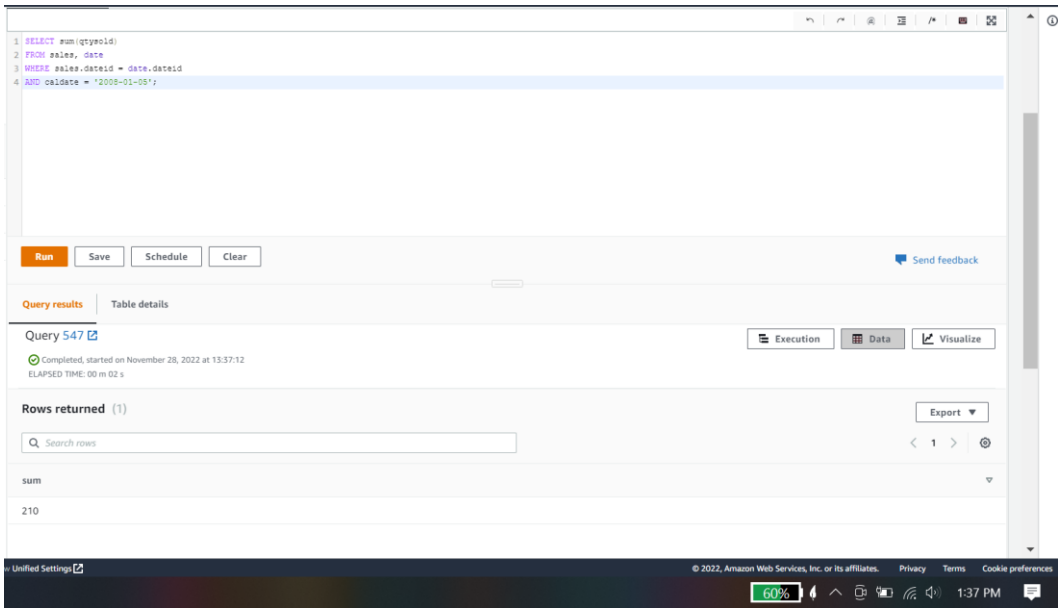


Task 4: Query the data

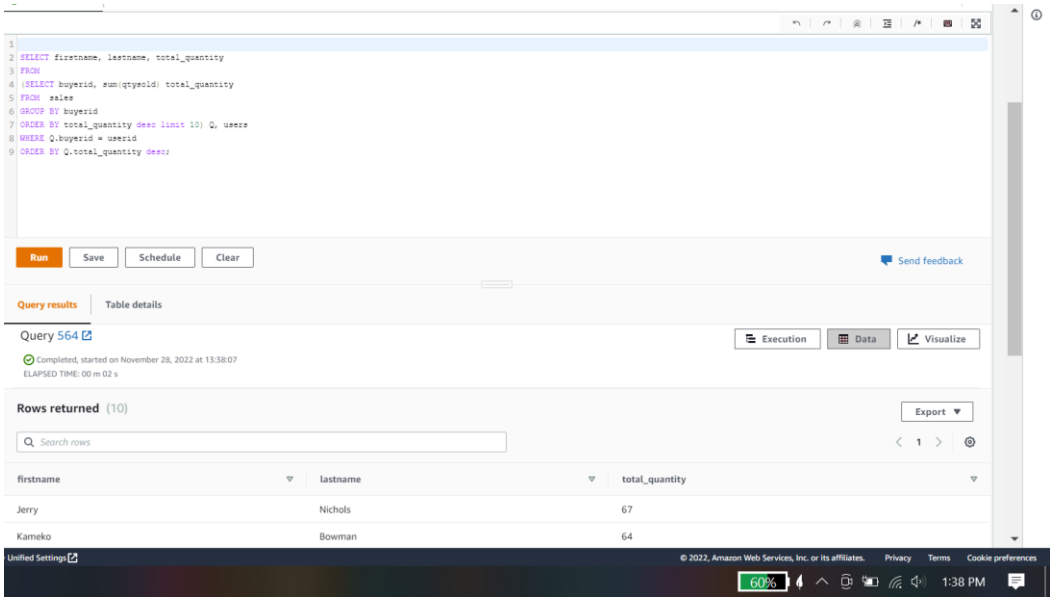
Preview sales table



Number of item sold on particular date



Top 10 buyers by quantity



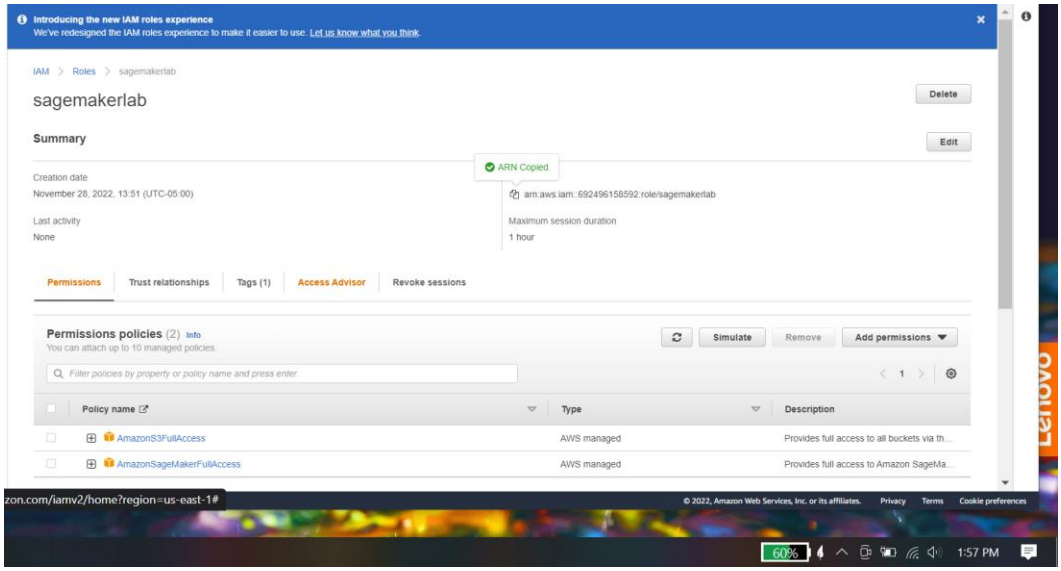
Lab 4 Conclusion.

- Accessed Amazon Redshift in the AWS Management Console
- Created an Amazon Redshift cluster.
- Load data from Amazon Simple Storage Service (Amazon S3) into an Amazon Redshift table
- Queried data in Amazon Redshift

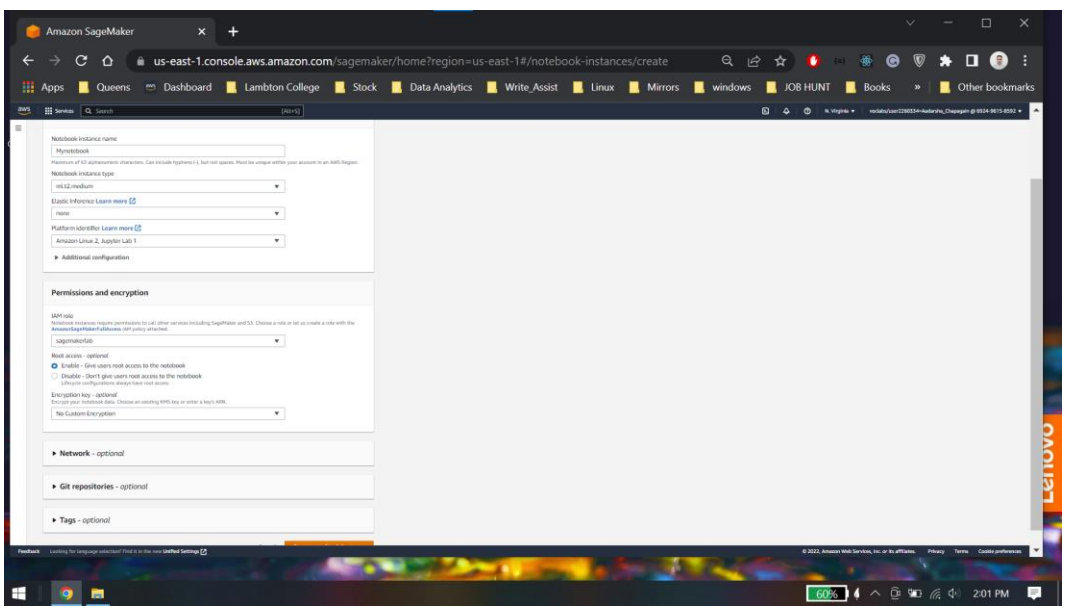
Lab5: Analyze Data with Amazon Sagemaker, Jupyter Notebooks and Bokeh

Task 1: Obtain the AWS Identity and Access Management (IAM) role

Sagemakerrolearn: arn:aws:iam::692496158592:role/sagemakerlab



Task 2: Create a Jupyter notebook with Amazon SageMaker

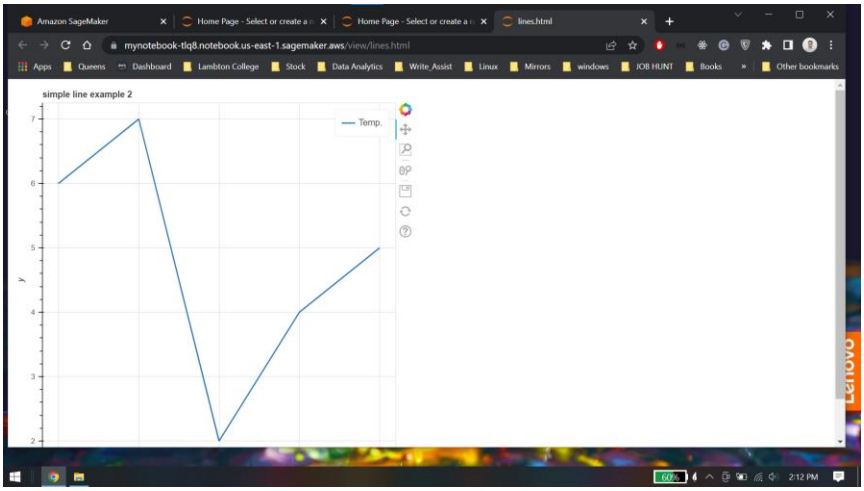


Task 3: Open your Jupyter notebook instance

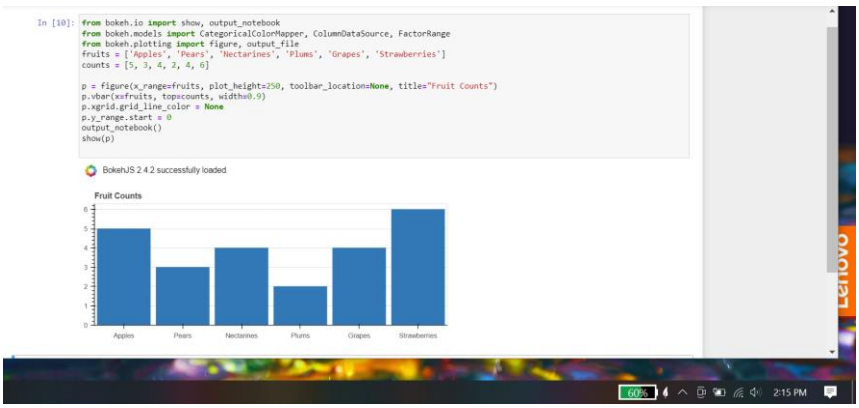
Task 4: Create visualizations with Bokeh

Task 4.1: Create a line graph

```
In [9]: from bokeh.plotting import figure, output_file, show

        # prepare some data
        x = [1, 2, 3, 4, 5]
        y = [6, 7, 2, 4, 5]

        # output to static HTML file
        output_file("lines.html")
        # create a new plot with a title and axis labels
        p = figure(title="simple line example 2", x_axis_label='x', y_axis_label='y')

        # add a line renderer with legend and line thickness
        p.line(x, y, legend_label="Temp.", line_width=2)

        # show the results
        show(p)

In [ ]:
```



Task 4.2: Create a bar chart

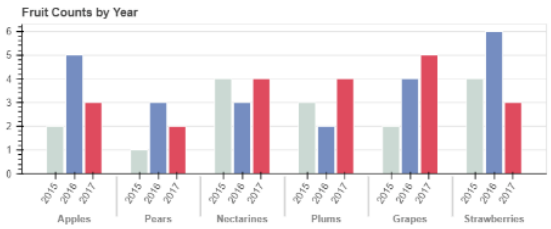


```
In [10]: from bokeh.io import show, output_notebook
         from bokeh.models import CategoricalColorMapper, ColumnDataSource, FactorRange
         from bokeh.plotting import figure, output_file
         fruits = ['Apples', 'Pears', 'Nectarines', 'Plums', 'Grapes', 'Strawberries']
         counts = [5, 3, 4, 2, 4, 6]

         p = figure(x_range=fruits, plot_height=250, toolbar_location=None, title="Fruit Counts")
         p.vbar(x=fruits, top=counts, width=0.9)
         p.xgrid.grid_line_color = None
         p.y_range.start = 0
         output_notebook()
         show(p)
```

Task 4.3: Create a grouped bar chart

In [11]:
```
from bokeh.transform import factor_cmap

fruits = ['Apples', 'Pears', 'Nectarines', 'Plums', 'Grapes', 'Strawberries']
years = ['2015', '2016', '2017']

data = {'fruits' : fruits,
          '2015'   : [2, 1, 4, 3, 2, 4],
          '2016'   : [5, 3, 3, 2, 4, 6],
          '2017'   : [3, 2, 4, 4, 5, 3]}

# this creates [ ("Apples", "2015"), ("Apples", "2016"), ("Apples", "2017"), ("Pears", "2015), ... ]
x = [ (fruit, year) for fruit in fruits for year in years ]
counts = sum(zip(data['2015'], data['2016'], data['2017']), ()) # like an hstack

source = ColumnDataSource(data=dict(x=x, counts=counts))

p = figure(x_range=FactorRange(*x), plot_height=250, toolbar_location=None, title="Fruit Counts by Year")
p.vbar(x='x', top='counts', width=0.9, source=source, line_color="white",
         fill_color=factor_cmap('x', palette=["#c9d9d3", "#718dbf", "#e84d60"], factors=years, start=1, end=2))

p.x_range.range_padding = 0.1
p.xgrid.grid_line_color = None
p.y_range.start = 0
p.xaxis.major_label_orientation = 1
output_notebook()
show(p)
```

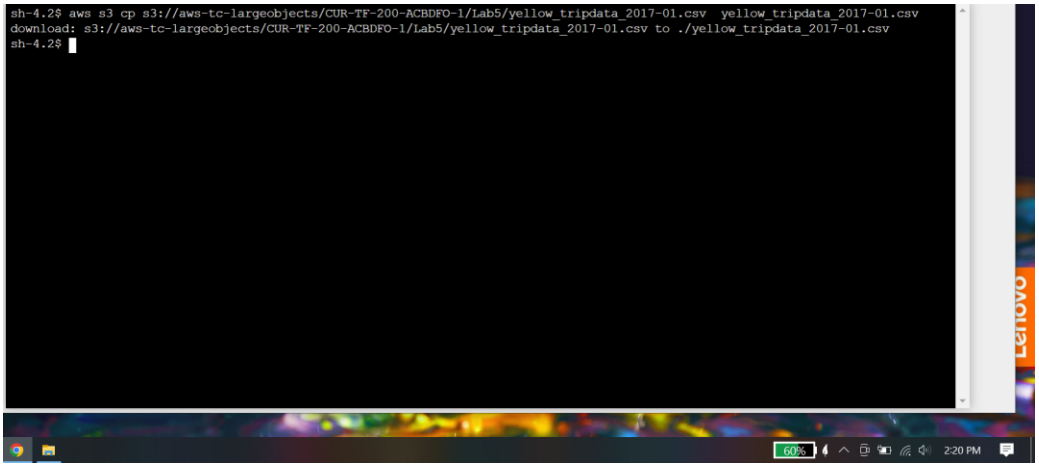
BokehJS 2.4.2 successfully loaded.



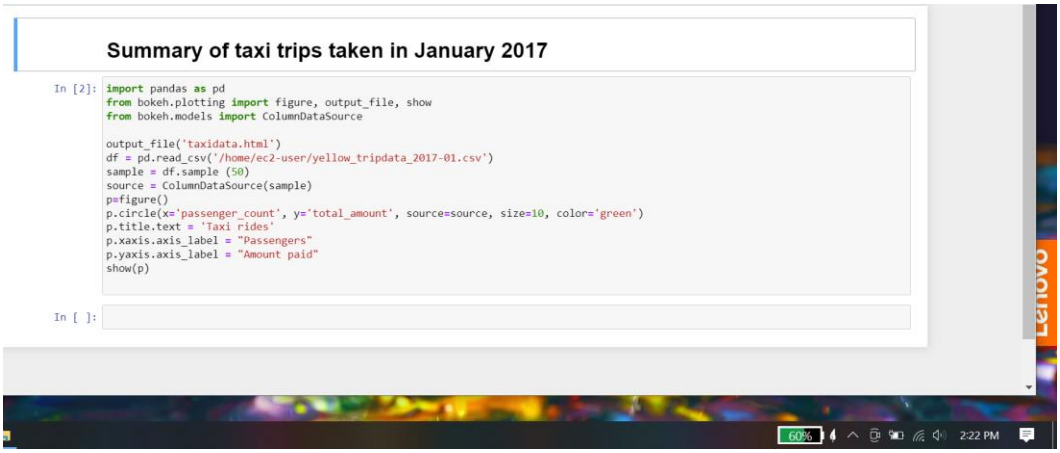
In [ ]:

Task 5: Create a visualization from a dataset

Task 5.1: Download the data file



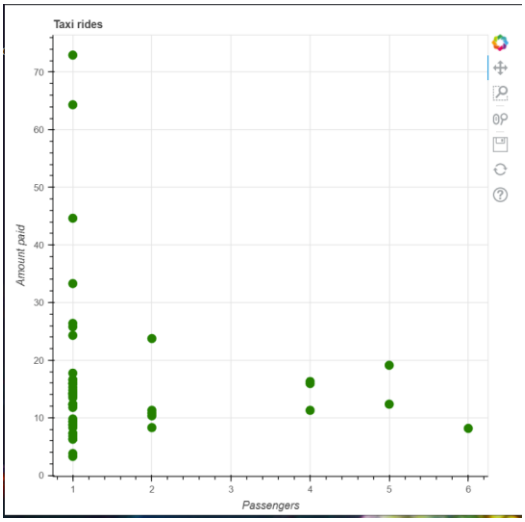
Task 5.2: Create the notebook



Summary of taxi trips taken in January 2017

In [2]:
```
import pandas as pd
from bokeh.plotting import figure, output_file, show
from bokeh.models import ColumnDataSource

output_file('taxidata.html')
df = pd.read_csv('/home/ec2-user/yellow_tripdata_2017-01.csv')
sample = df.sample (50)
source = ColumnDataSource(sample)
p=figure()
p.circle(x='passenger_count', y='total_amount', source=source, size=10, color='green')
p.title.text = 'Taxi rides'
p.xaxis.axis_label = "Passengers"
p.yaxis.axis_label = "Amount paid"
show(p)
```

In [ ]:



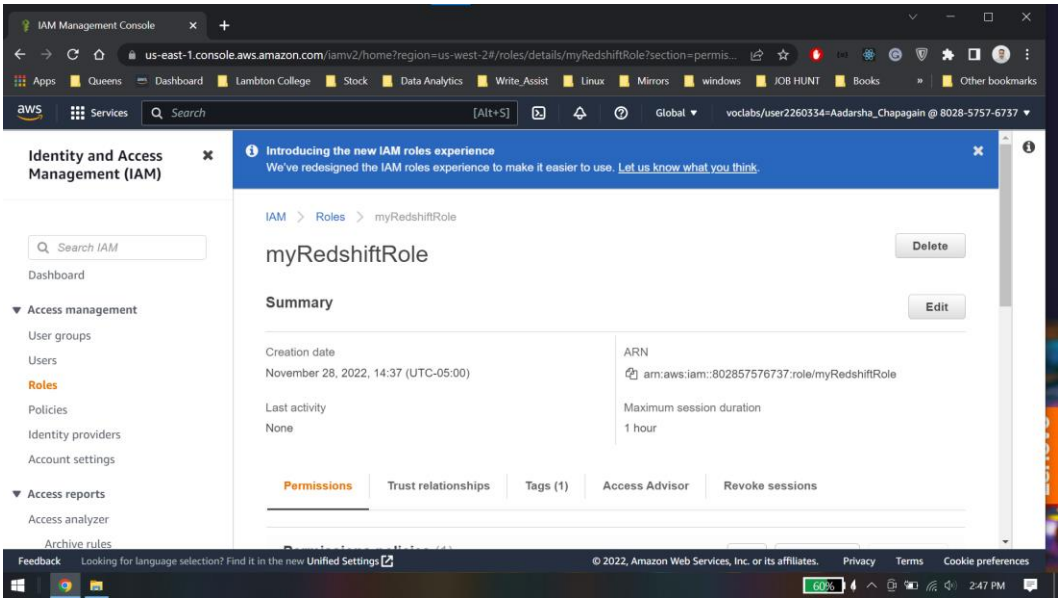
Lab 5 Conclusion

- Described Jupyter notebooks and the Bokeh visualization package.
- Created a Jupyter notebook with Amazon SageMaker.
- Imported data into a Jupyter notebook.
- Created a presentation with a Jupyter notebook.
- Visualized data with the open-source Bokeh Python package.

Lab 6: Automate Loading Data with AWS Data Pipeline

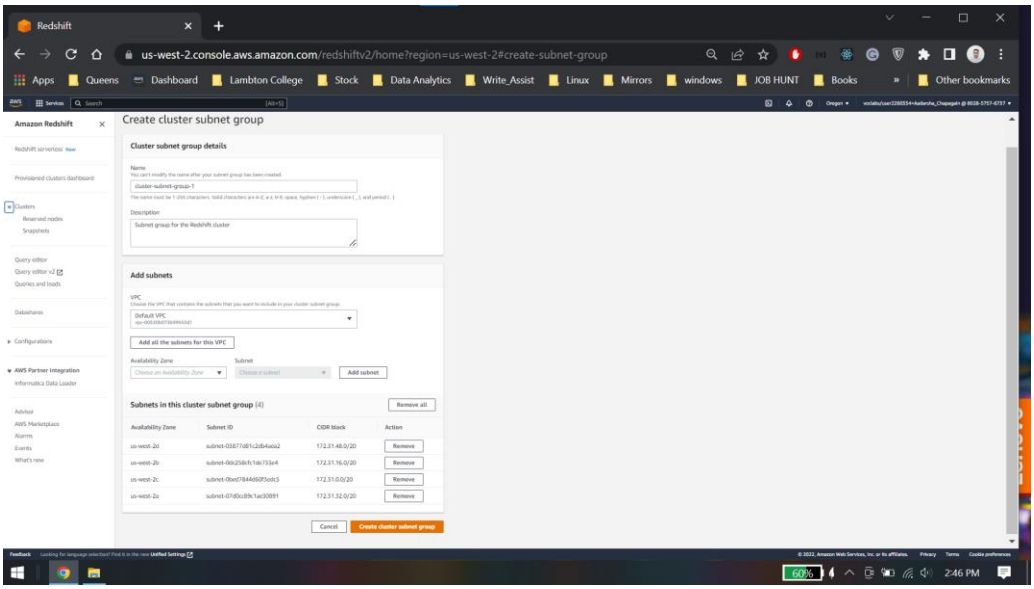
Task 1: Review the security group for accessing the Amazon Redshift console

Arn: arn:aws:iam::802857576737:role/myRedshiftRole

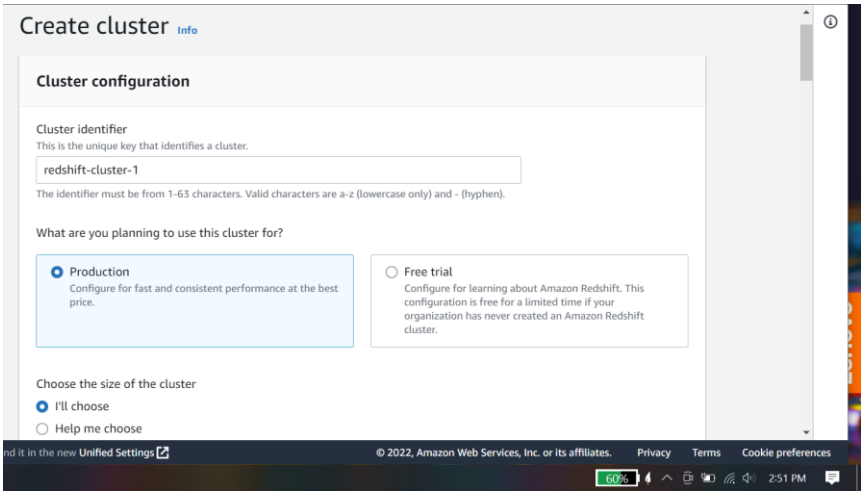


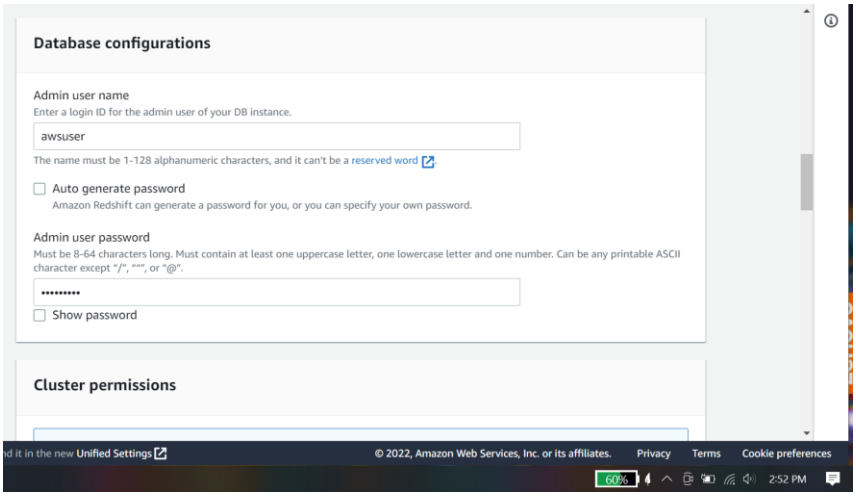
Task 2: Create and configure an Amazon Redshift cluster

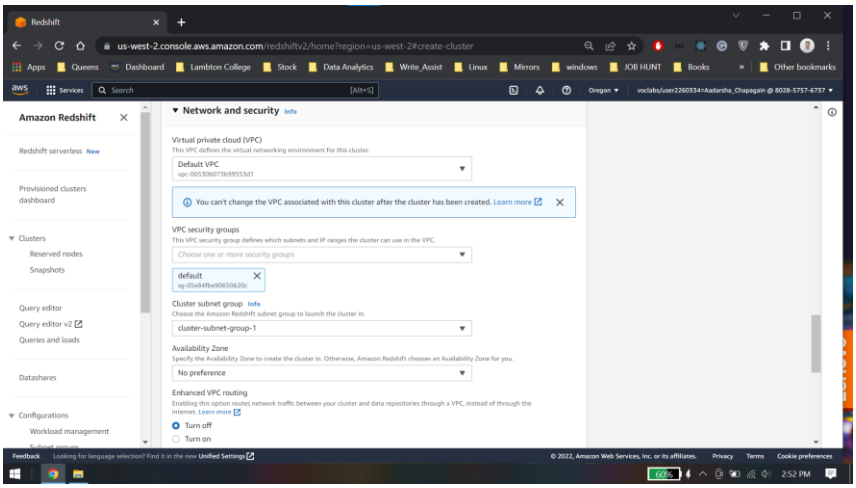
Create subnet cluster group



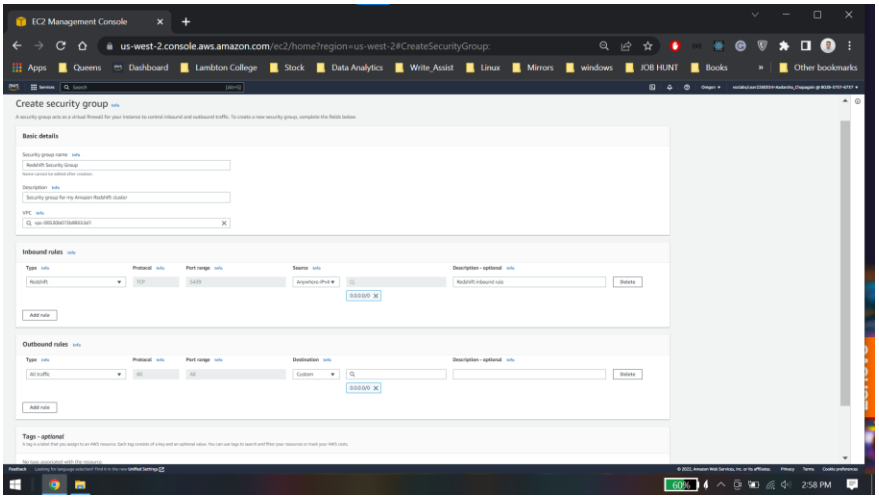
Create cluster



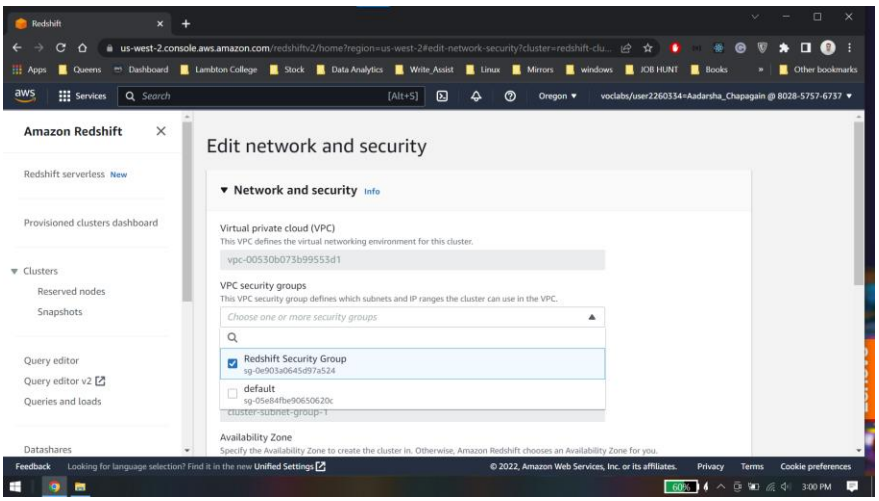




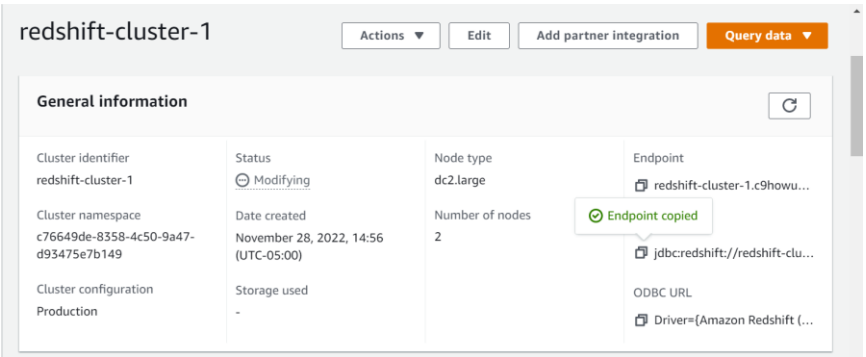
Task 2.1: Create a security group for your cluster



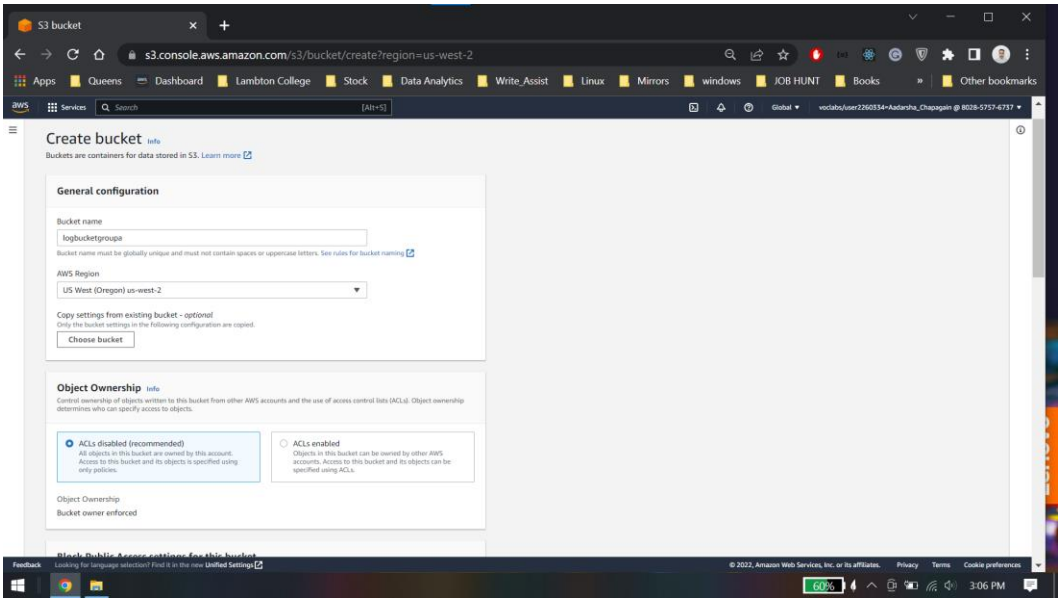
Task 2.2: Configure your Amazon Redshift cluster



Task 2.3: Capture JDBC connection information



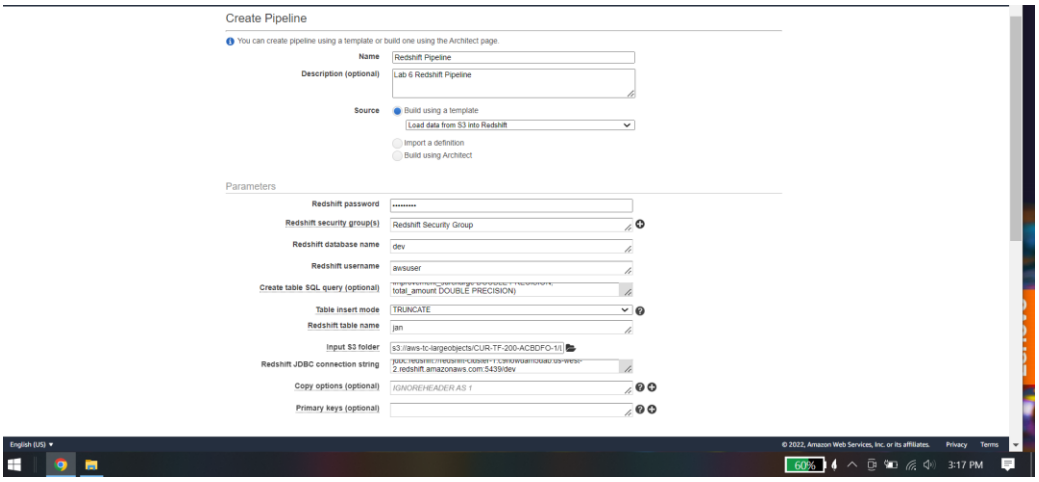
Task 2.4: Create an S3 bucket to store log files

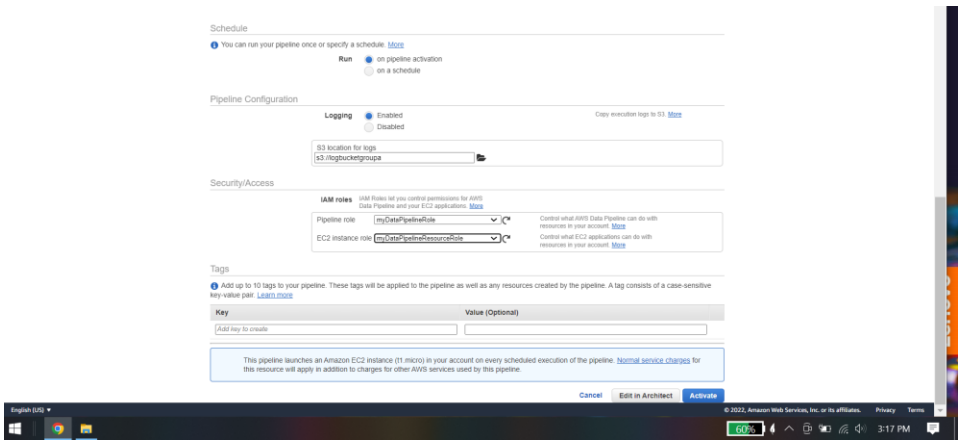


Bucketname:logbucketgroupa

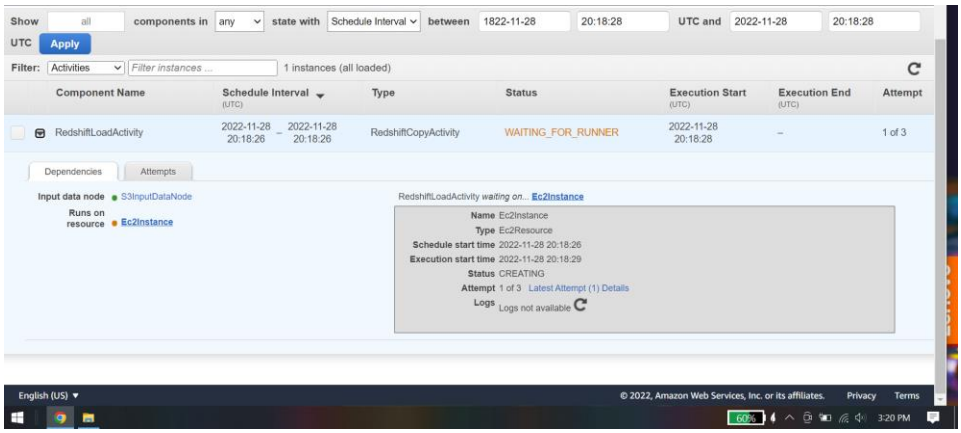
Jdbc Connection: jdbc:redshift://redshift-cluster-1.c9howuam5uab.us-west-2.redshift.amazonaws.com:5439/dev

Task 3: Create the pipeline

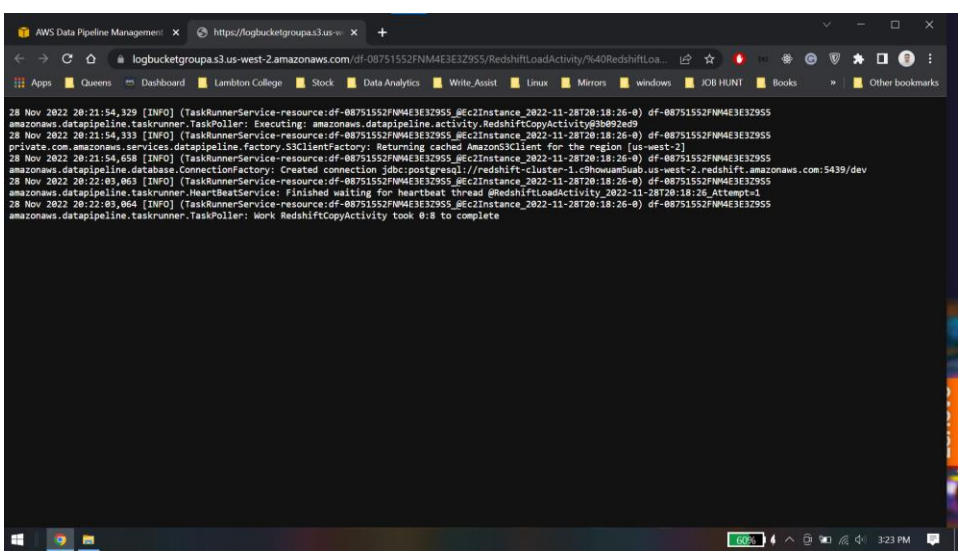




Task 4: Monitor your pipeline

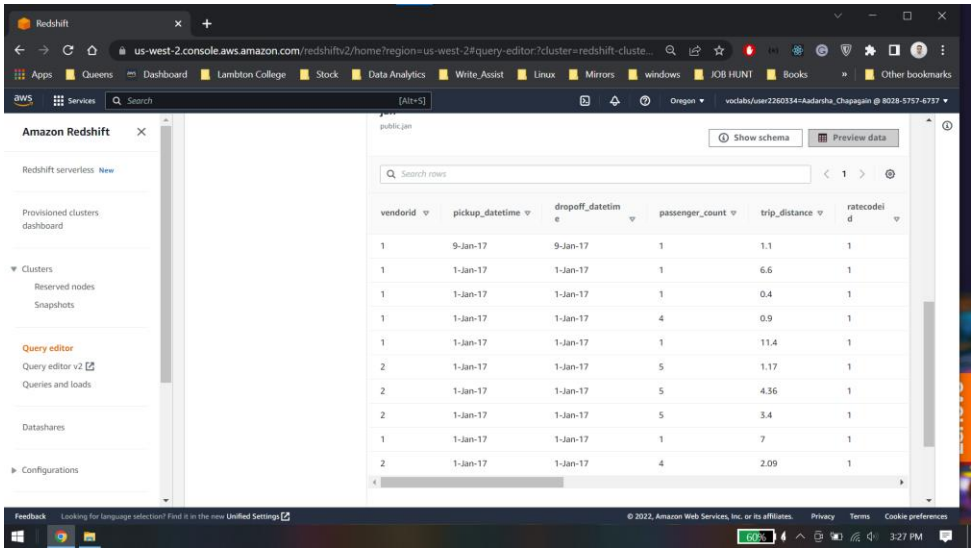


Task 5: View log files



Task 6: Query the Amazon Redshift database

Connect to redshift using username and password



Lab 6 Conclusion

- Accessed AWS Data Pipeline in the AWS Management Console.
- Created a data pipeline.
- Load data from Amazon S3 into Amazon Redshift with a data pipeline.
- Troubleshoot a data pipeline.
- Export data from Amazon Redshift to a Jupyter notebook