

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

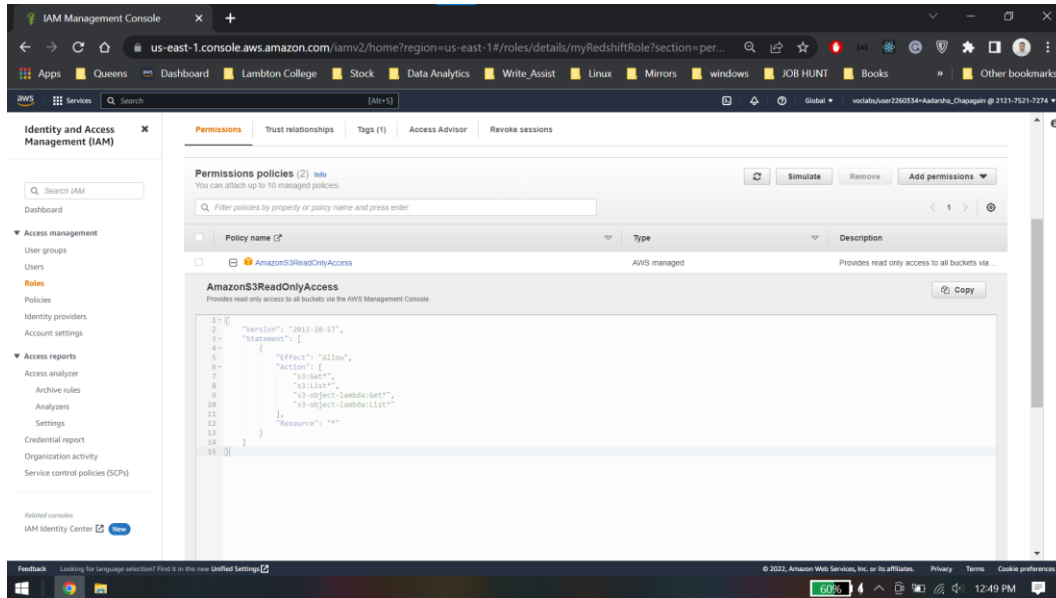
Submission Date:

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

Create cluster [Info](#)

Cluster configuration

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

☒ I'll choose

☐ Help me choose

Node type [Info](#)

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

dc2.large

Number of nodes

Enter the number of nodes that you need.

2

Range (1-32)

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 ▼

<input type="checkbox"/>	IAM roles <input type="button" value="X"/>	Status	Role type
<input type="checkbox"/>	myRedshiftRole	Not applied	--

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-08150f760eb5e1622)** and **default** subnet.

Backup

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

Maintenance

Using **current** maintenance track.

Security

Using **default (sg-06ca2b172fcbbe21b)** cluster security group.

Configuration

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

Task 2.1: Create a security group for your cluster

aws

Services

Search

[Alt+S]

N. Virginia

voclabs/user2260334-Audisha_Chappan @ 2121-7531-7274

EC2 > Security Groups > Create security group

Create security group [Info](#)

A security group acts as a virtual firewall for your instance to control inbound and outbound traffic. To create a new security group, complete the fields below.

Basic details

Security group name [Info](#)

Name cannot be edited after creation.

Description [Info](#)

VPC [Info](#)

Inbound rules

Type	Protocol	Port range	Source	Description - optional
Redshift	TCP	5439	Anywhere... <input type="text" value="0.0.0.0/0"/>	Redshift inbound rule


Feedback

Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

1:05 PM

Task 2.2: Configure your Amazon Redshift cluster

 Services [Alt+S]

≡ Edit network and security

▼ Network and security [Info](#)

Virtual private cloud (VPC)
This VPC defines the virtual networking environment for this cluster.

vpc-08150f760eb5e1622

VPC security groups
This VPC security group defines which subnets and IP ranges the cluster can use in the VPC.

Choose one or more security groups ▼

Redshift Security Group ✕
sg-07dfd920d9a90fb0d

Cluster subnet group
Choose the Amazon Redshift subnet group to launch the cluster in.

default

Availability Zone
Specify the Availability Zone to create the cluster in. Otherwise, Amazon Redshift chooses an Availability Zone for you.

No preference

Enhanced VPC routing
Enabling this option routes network traffic between your cluster and data repositories through a VPC, instead of through the internet. [Learn more](#)

☒ Turn off
☐ Turn on

Publicly accessible

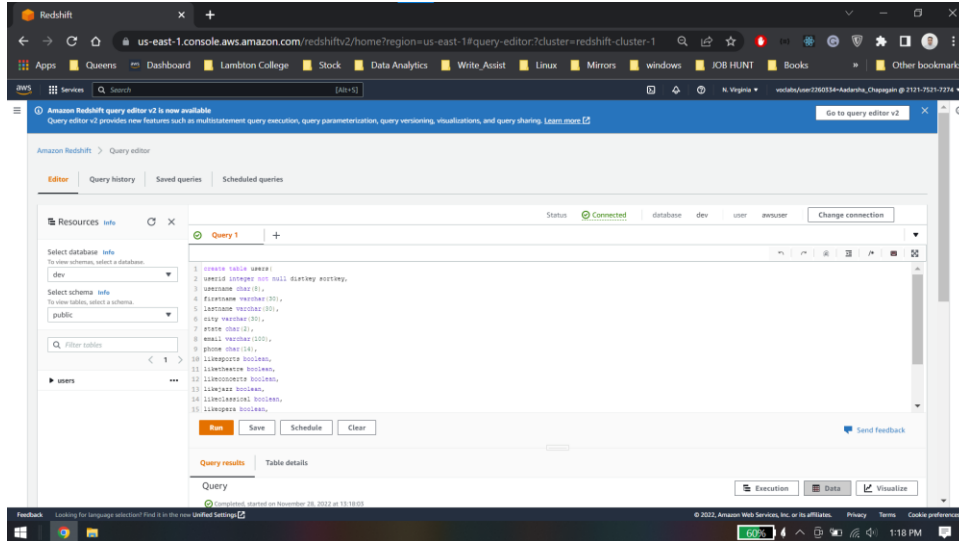
☐ Turn on Publicly accessible
Allow public connections to Amazon Redshift.

Cancel **Save changes**

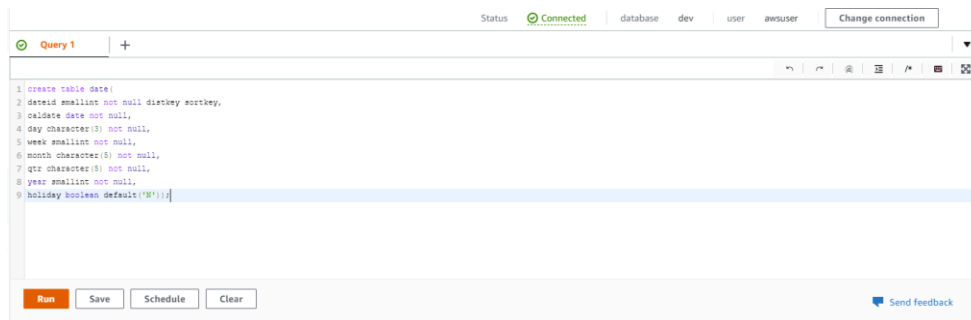
Task 3: Load data into your Amazon Redshift cluster

Task 3.1: Create the tables in the dev database

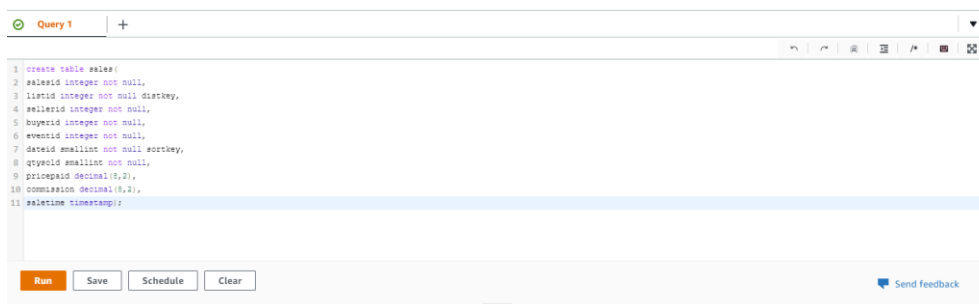
User table:



Date table



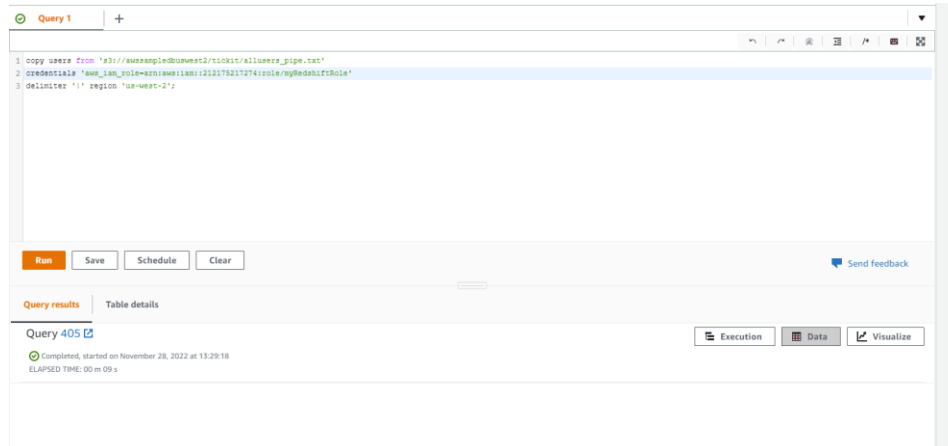
Sales table



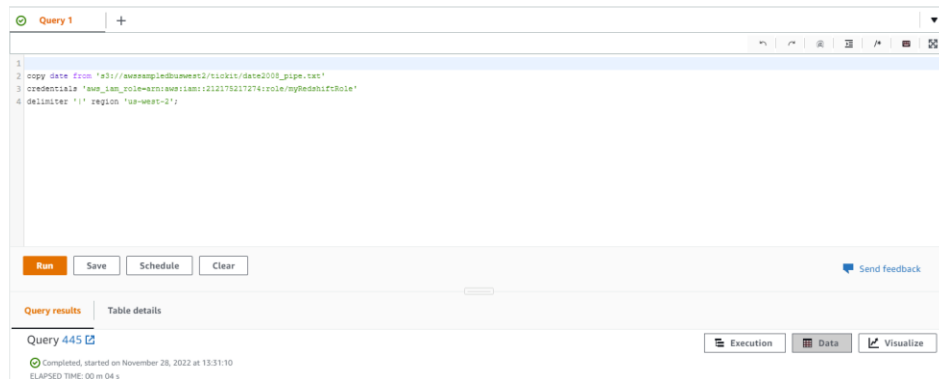
Task 3.2: Load data from Amazon S3

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

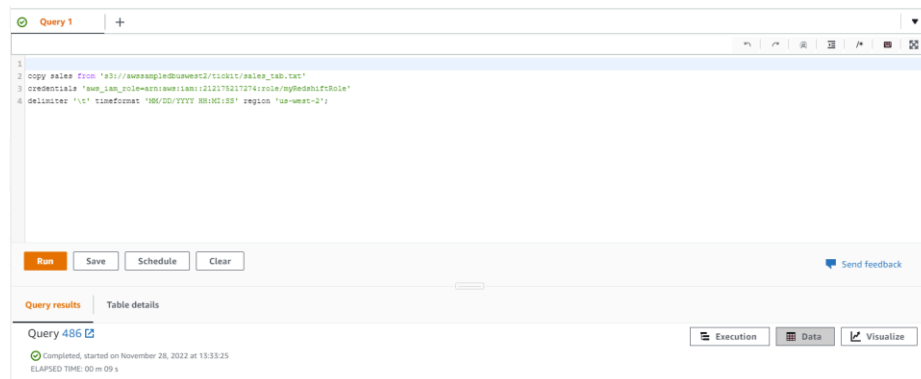
Copy users table



Copy date table

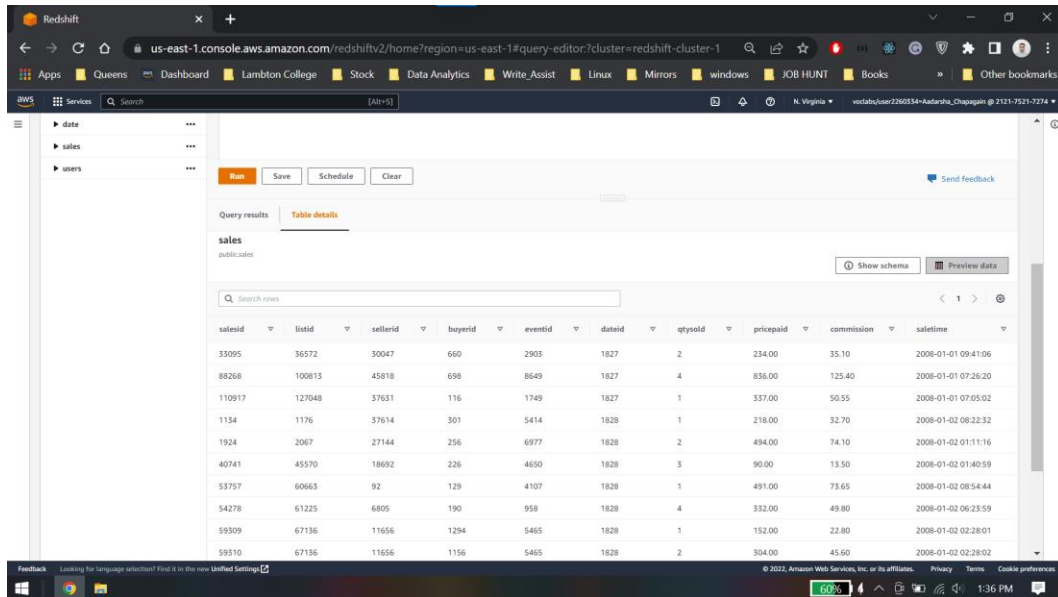


Copy sales table



Task 4: Query the data

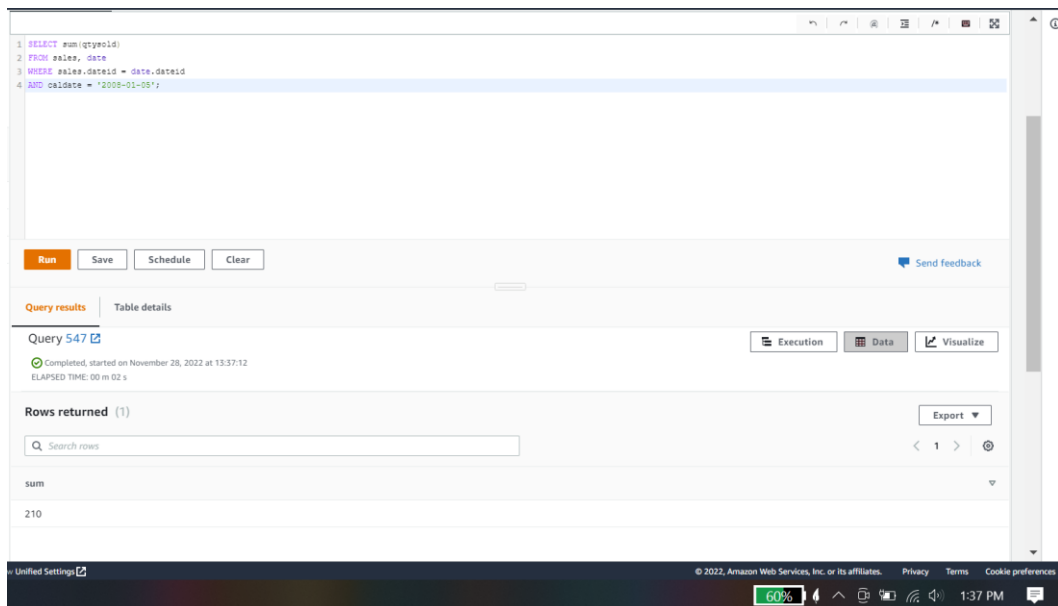
Preview sales table



The screenshot shows the AWS Redshift console interface. On the left, a sidebar lists databases: 'date', 'sales', and 'users'. The 'sales' database is selected, and the 'public.sales' table is chosen. The 'Table details' tab is active, displaying a preview of the table data. The table has 11 columns: salesid, listid, sellerid, buyerid, eventid, dateid, qtysold, pricepaid, commission, and saletime. The data is presented in a table with 10 rows visible. The bottom of the screen shows the Windows taskbar with the time 1:36 PM.

salesid	listid	sellerid	buyerid	eventid	dateid	qtysold	pricepaid	commission	saletime
33095	36572	30047	660	2903	1827	2	234.00	35.10	2008-01-01 09:41:06
88268	100813	45818	698	8649	1827	4	836.00	125.40	2008-01-01 07:29:20
110917	127048	37631	116	1749	1827	1	337.00	50.55	2008-01-01 07:05:02
1134	1176	37614	301	5414	1828	1	218.00	32.70	2008-01-02 08:22:32
1924	2067	27144	256	6977	1828	2	494.00	74.10	2008-01-02 01:11:16
40741	45570	18692	226	4650	1828	3	90.00	13.50	2008-01-02 01:40:59
53757	60663	92	129	4107	1828	1	491.00	73.65	2008-01-02 08:54:44
54278	61225	6805	190	958	1828	4	332.00	49.80	2008-01-02 06:23:59
59309	67136	11656	1294	5465	1828	1	152.00	22.80	2008-01-02 02:28:01
59310	67136	11656	1156	5465	1828	2	304.00	45.60	2008-01-02 02:28:02

Number of item sold on particular date



The screenshot shows the AWS Redshift console with a SQL query executed. The query is: `SELECT sum(qtysold) FROM sales, date WHERE sales.dateid = date.dateid AND caldate = '2008-01-05';`. The query results show a single row with the sum of items sold, which is 210. The bottom of the screen shows the Windows taskbar with the time 1:37 PM.

```
1 SELECT sum(qtysold)
2 FROM sales, date
3 WHERE sales.dateid = date.dateid
4 AND caldate = '2008-01-05';
```

Query 547

Completed, started on November 28, 2022 at 13:37:12
ELAPSED TIME: 00 m 02 s

Rows returned (1)

sum
210

Top 10 buyers by quantity

The screenshot displays the Amazon Redshift console interface. At the top, a SQL query is entered in a text area:

```
1
2 SELECT firstname, lastname, total_quantity
3 FROM
4 (SELECT buyerid, sum(qtysold) total_quantity
5 FROM sales
6 GROUP BY buyerid
7 ORDER BY total_quantity desc limit 10) Q, users
8 WHERE Q.buyerid = usersid
9 ORDER BY Q.total_quantity desc;
```

Below the query editor are buttons for "Run", "Save", "Schedule", and "Clear", along with a "Send feedback" link. The "Query results" tab is active, showing "Query 564" with a status of "Completed, started on November 28, 2022 at 13:38:07" and an elapsed time of "00 m 02 s".

The results section shows "Rows returned (10)" and a search bar. Below this is a table with the following data:

firstname	lastname	total_quantity
Jerry	Nichols	67
Kameko	Bowman	64

The bottom of the screen shows the AWS Management Console footer with copyright information and a system tray at the bottom of the browser window.

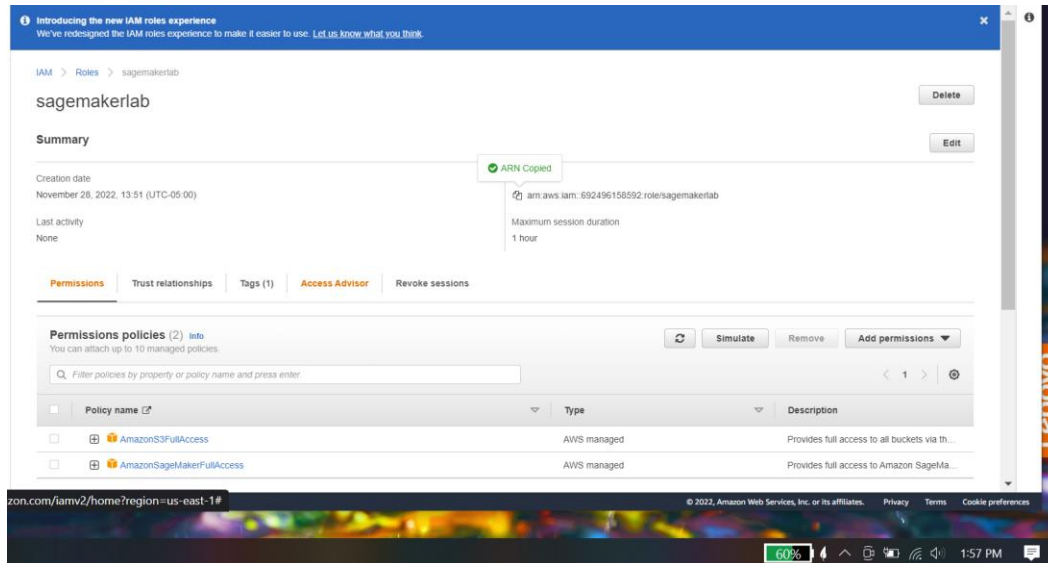
Lab 4 Conclusion.

- 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

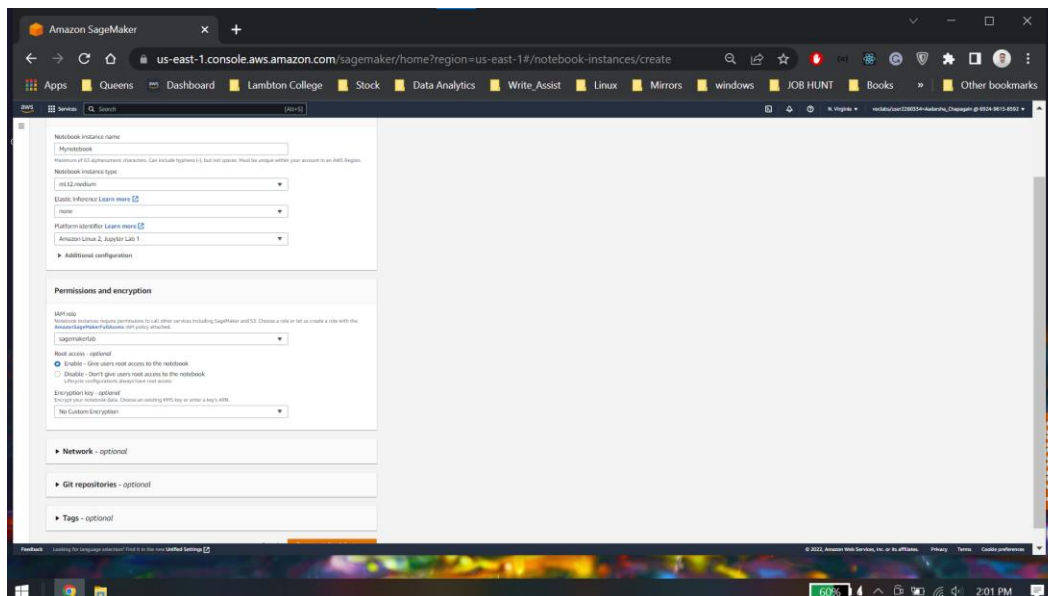
Lab5: Analyze Data with Amazon Sagemaker, Jupyter Notebooks and Bokeh

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

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



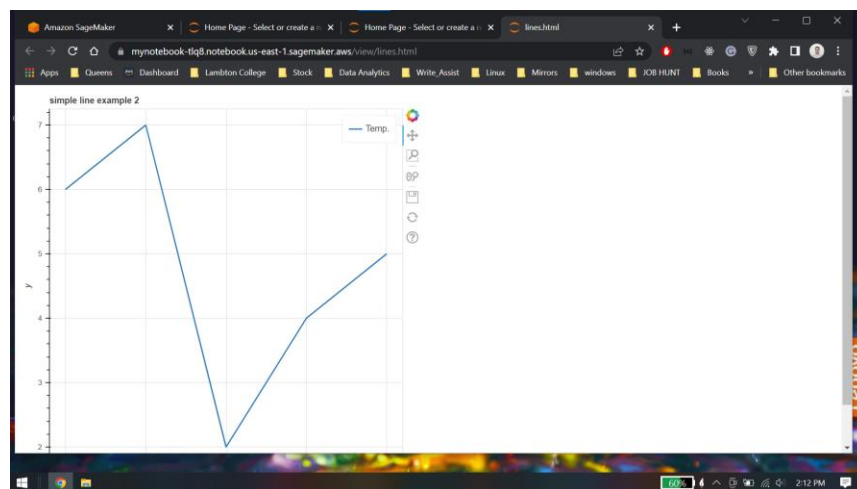
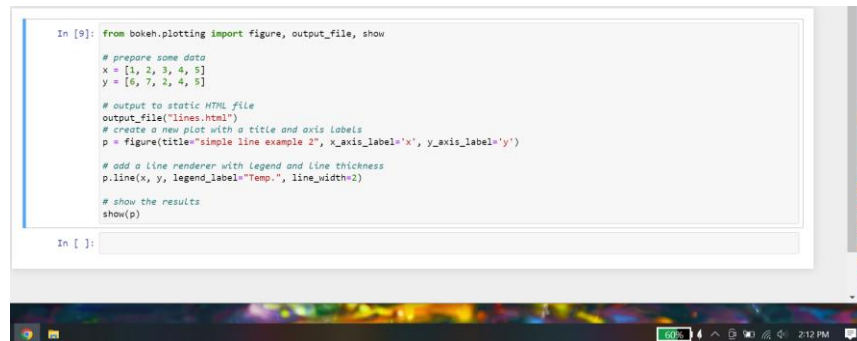
Task 2: Create a Jupyter notebook with Amazon SageMaker



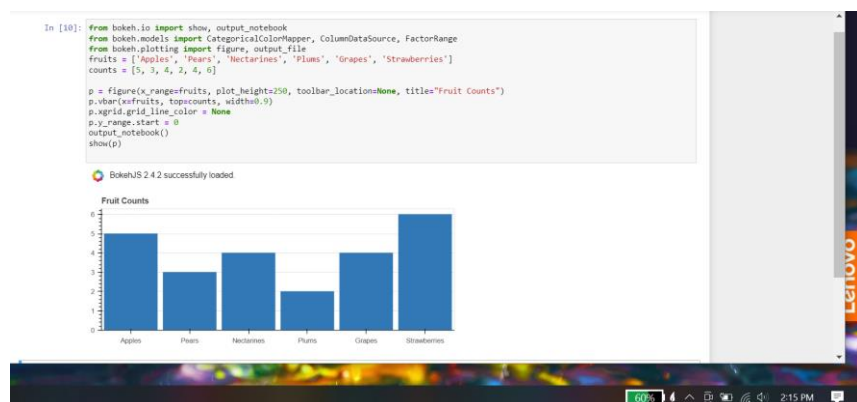
Task 3: Open your Jupyter notebook instance

Task 4: Create visualizations with Bokeh

Task 4.1: Create a line graph



Task 4.2: Create a bar chart



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]}

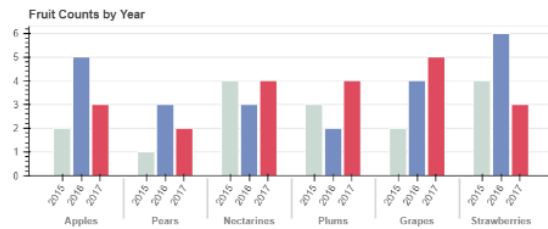
# this creates [ ("Apples", "2015"), ("Apples", "2016"), ("Apples", "2017"), ("Pears", "2015"), ... ]
x = [ (fruit, year) for fruit in fruits for year in years ]
counts = sum(zip(data['2015'], data['2016'], data['2017']), ()) # Like an hstack

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

```
sh-4.2$ aws s3 cp s3://aws-tc-largeobjects/CUR-TF-200-ACBDFO-1/Lab5/yellow_tripdata_2017-01.csv yellow_tripdata_2017-01.csv
download: s3://aws-tc-largeobjects/CUR-TF-200-ACBDFO-1/Lab5/yellow_tripdata_2017-01.csv to ./yellow_tripdata_2017-01.csv
sh-4.2$
```

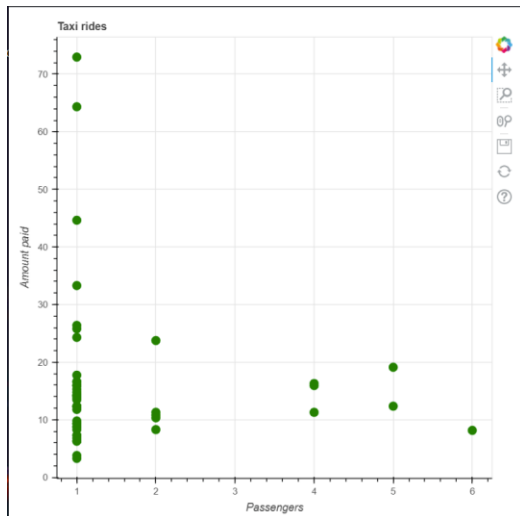
Task 5.2: Create the notebook

```
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 [ ]:
```



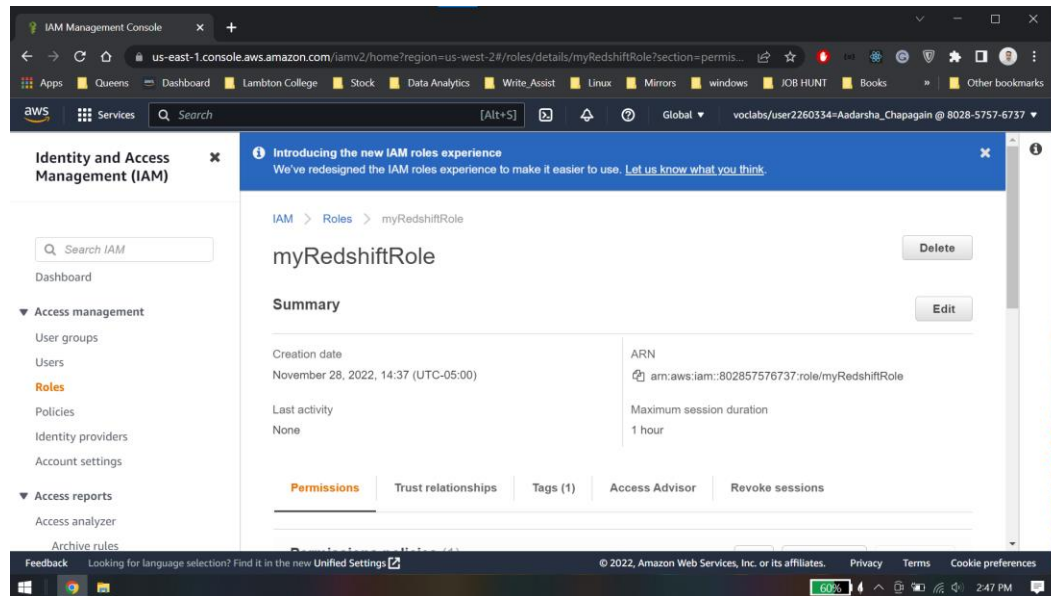
Lab 5 Conclusion

- 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

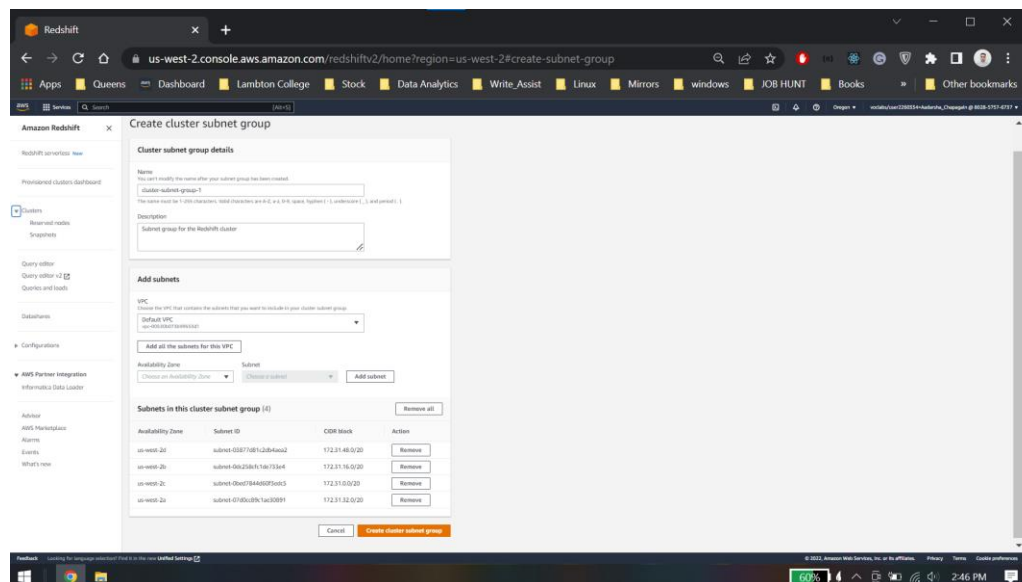
Task 1: Review the security group for accessing the Amazon Redshift console

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



Task 2: Create and configure an Amazon Redshift cluster

Create subnet cluster group



Create cluster

Create cluster Info

Cluster configuration

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

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

☒ **I'll choose**

☐ Help me choose

Find it in the new [Unified Settings](#)

© 2022, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

60% 2:51 PM

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

Cluster permissions

Find it in the new [Unified Settings](#)

© 2022, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

60% 2:52 PM

Redshift

us-west-2.console.aws.amazon.com/redshiftv2/home?region=us-west-2#create-cluster

Services

Search

[Alt+S]

Regions

Oregon

us-west-2

Amazon Redshift

Redshift serverless New

Provisioned clusters dashboard

Clusters

- Reserved nodes
- Snapshots

Query editor

- Query editor v2
- Queries and loads

Datashares

Configurations

- Workload management

Network and security Info

Virtual private cloud (VPC)
This VPC defines the virtual networking environment for this cluster.

Default VPC

vpc-0053060780955361

VPC security groups
This VPC security group defines which subnets and IP ranges the cluster can use in the VPC.

Choose one or more security groups

default

sg-0f6d4f9e06050620c

Cluster subnet group Info
Choose the Amazon Redshift subnet group to launch the cluster in.

cluster-subnet-group-1

Availability Zone
Specify the Availability Zone to create the cluster in. Otherwise, Amazon Redshift chooses an Availability Zone for you.

No preference

Enhanced VPC routing
Enabling this option routes network traffic between your cluster and data repositories through a VPC, instead of through the internet. [Learn more](#)

Turn off

Turn on

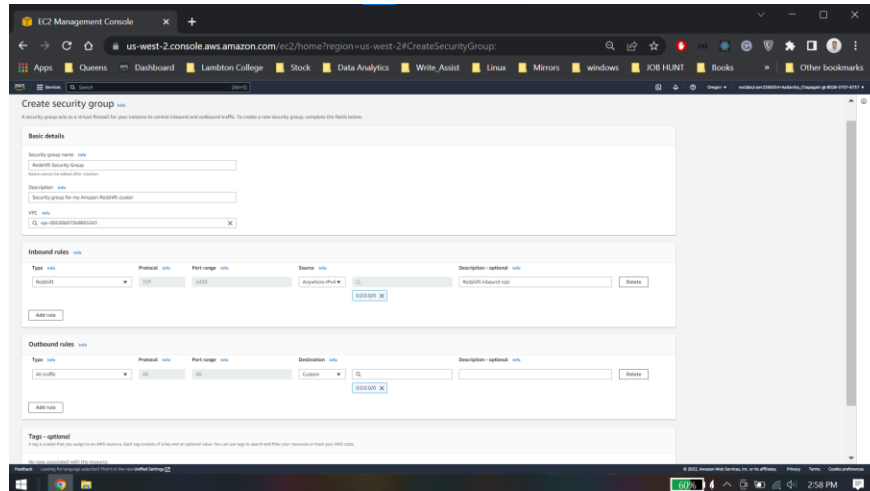
Feedback

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

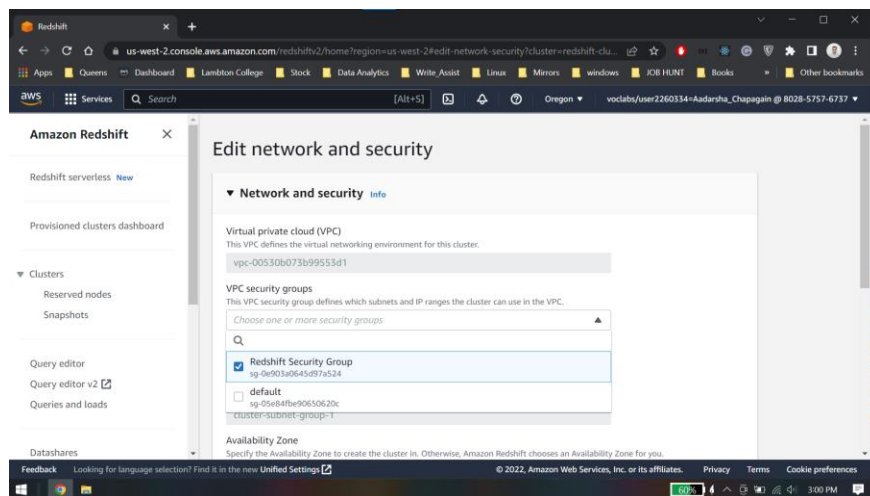
© 2022, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

60% 2:52 PM

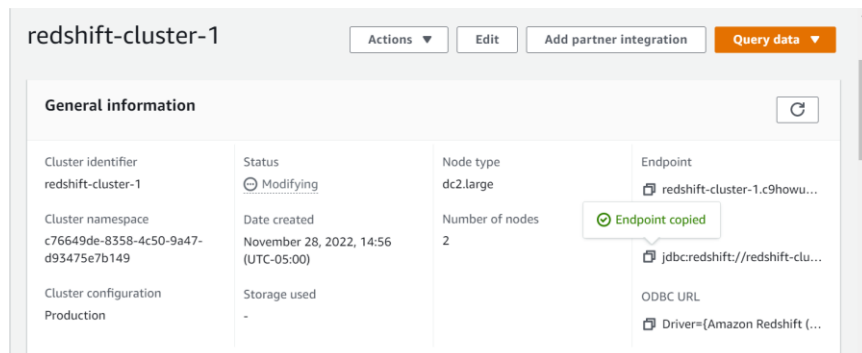
Task 2.1: Create a security group for your cluster



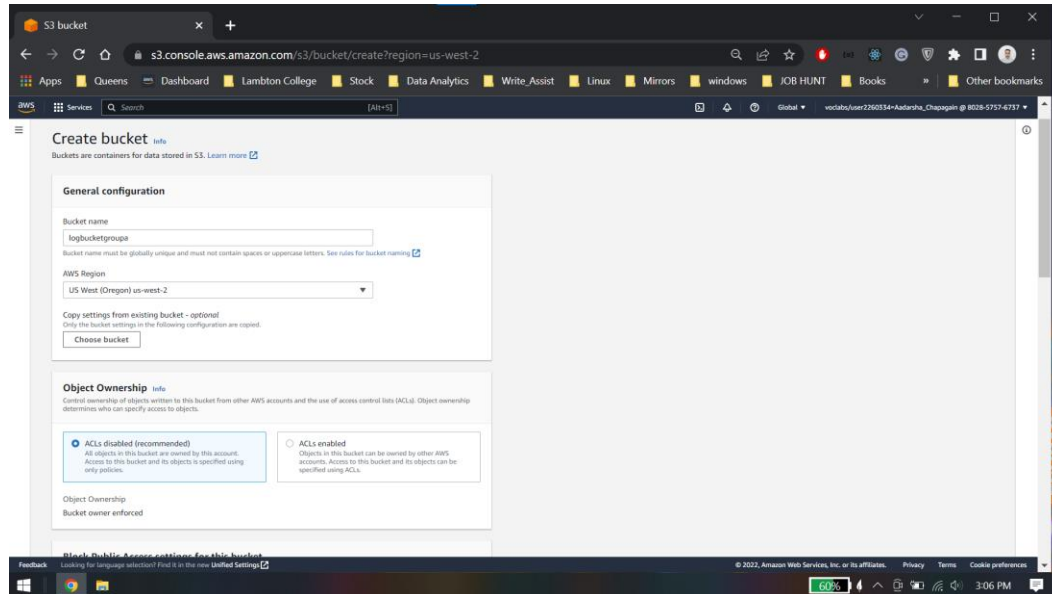
Task 2.2: Configure your Amazon Redshift cluster



Task 2.3: Capture JDBC connection information



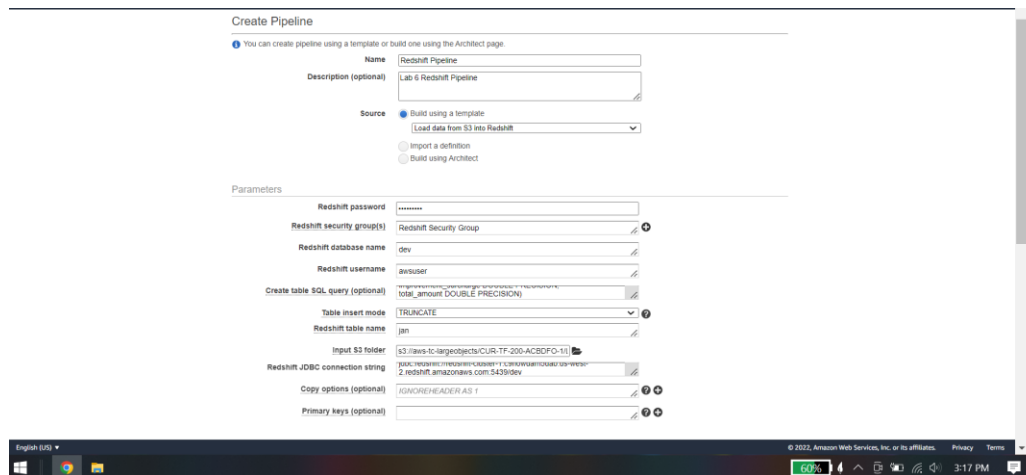
Task 2.4: Create an S3 bucket to store log files



Bucketname:logbucketgroupa

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

Task 3: Create the pipeline



Schedule

You can run your pipeline once or specify a schedule. [More](#)

Run ☒ on pipeline activation ☐ on a schedule

Pipeline Configuration

Logging ☒ Enabled ☐ Disabled [Copy execution logs to S3](#) [More](#)

S3 location for logs

Security/Access

IAM roles IAM Roles let you control permissions for AWS Data Pipeline and your EC2 applications. [More](#)

Pipeline role [More](#) Control what AWS Data Pipeline can do with resources in your account. [More](#)

EC2 instance role [More](#) Control what EC2 applications can do with resources in your account. [More](#)

Tags

Add up to 10 tags to your pipeline. These tags will be applied to the pipeline as well as any resources created by the pipeline. A tag consists of a case-sensitive key-value pair. [Learn more](#)

Key	Value (Optional)
<input type="text" value="Add key to create"/>	<input type="text"/>

This pipeline launches an Amazon EC2 instance (t1.micro) in your account on every scheduled execution of the pipeline. [Normal service charges](#) for this resource will apply in addition to charges for other AWS services used by this pipeline.

[Cancel](#) [Edit in Architect](#) [Activate](#)

Task 4: Monitor your pipeline

Show components in state with between UTC and UTC [Apply](#)

Filter: [Filter instances ...](#) 1 instances (all loaded)

Component Name	Schedule Interval (UTC)	Type	Status	Execution Start (UTC)	Execution End (UTC)	Attempt
RedshiftLoadActivity	2022-11-28 20:18:26 - 2022-11-28 20:18:26	RedshiftCopyActivity	WAITING_FOR_RUNNER	2022-11-28 20:18:26	-	1 of 3

Dependencies **Attempts**

Input data node S3InputDataNode

Runs on resource **Ec2Instance**

RedshiftLoadActivity waiting on **Ec2Instance**

Name Ec2Instance
Type Ec2Resource
Schedule start time 2022-11-28 20:18:26
Execution start time 2022-11-28 20:18:29
Status CREATING
Attempt 1 of 3 [Latest Attempt \(1\) Details](#)
[Logs](#) [Logs not available](#)

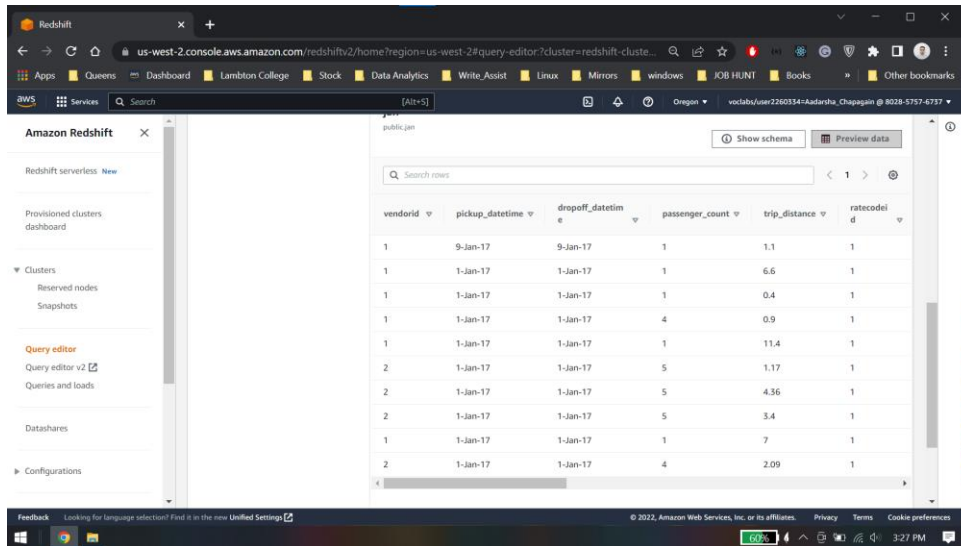
Task 5: View log files

```

AWS Data Pipeline Management x https://logbucketgroupa.s3.us-west-2.amazonaws.com/df-08751552FNME4E3E3Z955/RedshiftLoadActivity/640RedshiftLoa...
logbucketgroupa.s3.us-west-2.amazonaws.com/df-08751552FNME4E3E3Z955/RedshiftLoadActivity/640RedshiftLoa...
28 Nov 2022 20:21:54,329 [INFO] (TaskRunnerService-resource:df-08751552FNME4E3E3Z955_Ec2Instance_2022-11-28T20:18:26-0) df-08751552FNME4E3E3Z955
amazonaws.datapipeline.taskrunner.TaskPoller: Executing: amazonaws.datapipeline.activity.RedshiftCopyActivity@3b092ed9
28 Nov 2022 20:21:54,333 [INFO] (TaskRunnerService-resource:df-08751552FNME4E3E3Z955_Ec2Instance_2022-11-28T20:18:26-0) df-08751552FNME4E3E3Z955
private.com.amazonaws.services.datapipeline.factory.S3ClientFactory: Returning cached AmazonS3Client for the region [us-west-2]
28 Nov 2022 20:21:54,658 [INFO] (TaskRunnerService-resource:df-08751552FNME4E3E3Z955_Ec2Instance_2022-11-28T20:18:26-0) df-08751552FNME4E3E3Z955
amazonaws.datapipeline.database.ConnectionFactory: Created connection jdbc:postgresql://redshift-cluster-1.c9howuam5uab.us-west-2.redshift.amazonaws.com:5439/dev
28 Nov 2022 20:22:03,063 [INFO] (TaskRunnerService-resource:df-08751552FNME4E3E3Z955_Ec2Instance_2022-11-28T20:18:26-0) df-08751552FNME4E3E3Z955
amazonaws.datapipeline.taskrunner.HeartBeatService: Finished waiting for heartbeat thread @RedshiftLoadActivity_2022-11-28T20:18:26 Attempt-1
28 Nov 2022 20:22:03,064 [INFO] (TaskRunnerService-resource:df-08751552FNME4E3E3Z955_Ec2Instance_2022-11-28T20:18:26-0) df-08751552FNME4E3E3Z955
amazonaws.datapipeline.taskrunner.TaskPoller: Work RedshiftCopyActivity took 0:8 to complete
  
```

Task 6: Query the Amazon Redshift database

Connect to redshift using username and password



The screenshot shows the Amazon Redshift console interface. On the left is a navigation sidebar with options like 'Redshift serverless', 'Provisioned clusters dashboard', 'Clusters', 'Query editor', 'Datashares', and 'Configurations'. The main area displays a table of data with columns: vendorid, pickup_datetime, dropoff_datetime, passenger_count, trip_distance, and ratecodeid. The table contains 10 rows of data. At the top of the table area, there are buttons for 'Show schema' and 'Preview data'. A search bar for rows is also present.

vendorid	pickup_datetime	dropoff_datetime	passenger_count	trip_distance	ratecodeid
1	9-Jan-17	9-Jan-17	1	1.1	1
1	1-Jan-17	1-Jan-17	1	6.6	1
1	1-Jan-17	1-Jan-17	1	0.4	1
1	1-Jan-17	1-Jan-17	4	0.9	1
1	1-Jan-17	1-Jan-17	1	11.4	1
2	1-Jan-17	1-Jan-17	5	1.17	1
2	1-Jan-17	1-Jan-17	5	4.36	1
2	1-Jan-17	1-Jan-17	5	3.4	1
1	1-Jan-17	1-Jan-17	1	7	1
2	1-Jan-17	1-Jan-17	4	2.09	1

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