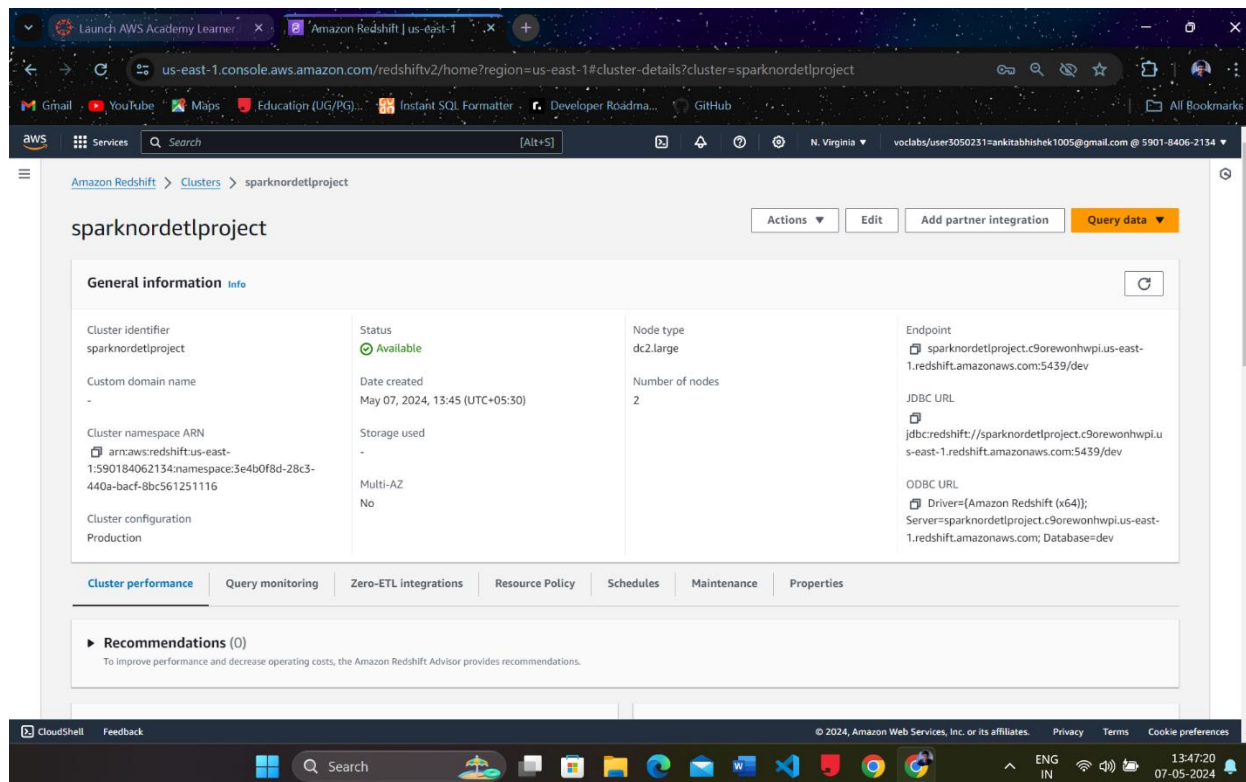


Creation of a Redshift Cluster

Screenshots of the configuration of the Redshift cluster created:



The screenshot displays the Amazon Redshift console interface for a cluster named 'sparknordetlproject'. The browser address bar shows the URL: `us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#cluster-details?cluster=sparknordetlproject`. The console header includes the AWS logo, a search bar, and the user's profile information (N. Virginia, voclabs/user3050231=ankitabhishek1005@gmail.com @ 5901-8406-2134).

The main content area shows the cluster details for 'sparknordetlproject'. The 'General information' tab is selected, displaying the following details:

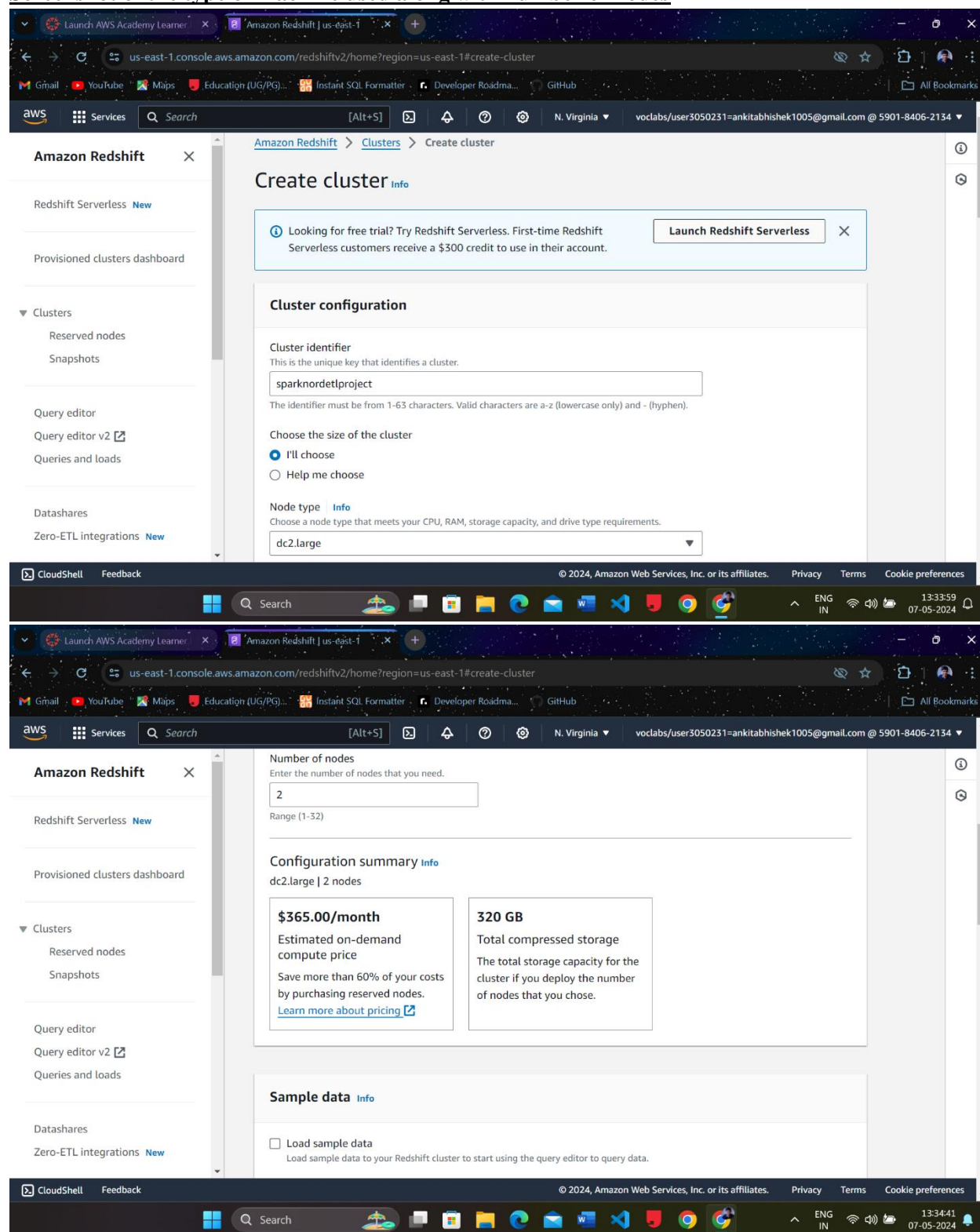
Cluster identifier	Status	Node type	Endpoint
sparknordetlproject	Available	dc2.large	sparknordetlproject.c9orewonhwpi.us-east-1.redshift.amazonaws.com:5439/dev
Custom domain name	Date created	Number of nodes	JDBC URL
-	May 07, 2024, 13:45 (UTC+05:30)	2	jdbc:redshift://sparknordetlproject.c9orewonhwpi.us-east-1.redshift.amazonaws.com:5439/dev
Cluster namespace ARN	Storage used		ODBC URL
arn:aws:redshift:us-east-1:590184062134:namespace:3e4b0f8d-28c3-440a-bacf-8bc561251116	-		Driver=[Amazon Redshift (x64)]; Server=sparknordetlproject.c9orewonhwpi.us-east-1.redshift.amazonaws.com; Database=dev
Cluster configuration	Multi-AZ		
Production	No		

Below the general information, there are tabs for 'Cluster performance', 'Query monitoring', 'Zero-ETL integrations', 'Resource Policy', 'Schedules', 'Maintenance', and 'Properties'. The 'Cluster performance' tab is currently selected.

At the bottom, there is a 'Recommendations' section with a heading 'Recommendations (0)' and a note: 'To improve performance and decrease operating costs, the Amazon Redshift Advisor provides recommendations.'

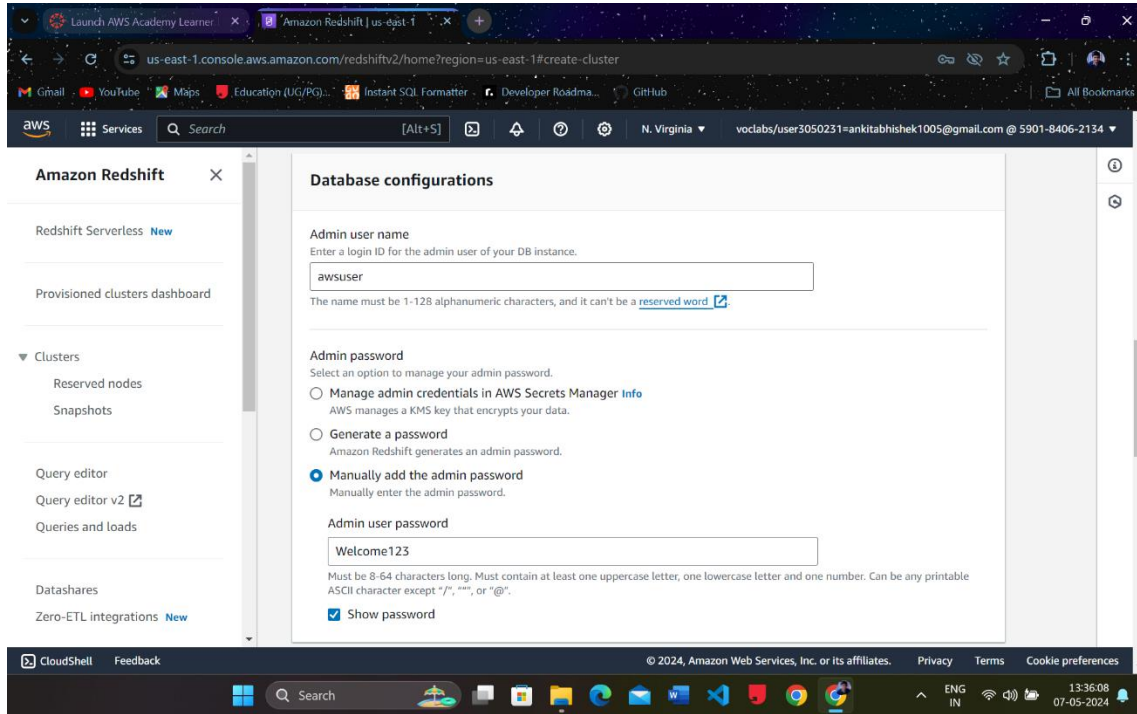
The bottom of the screenshot shows the Windows taskbar with the Start button, search bar, and various application icons. The system clock indicates the time is 13:47:20 on 07-05-2024.

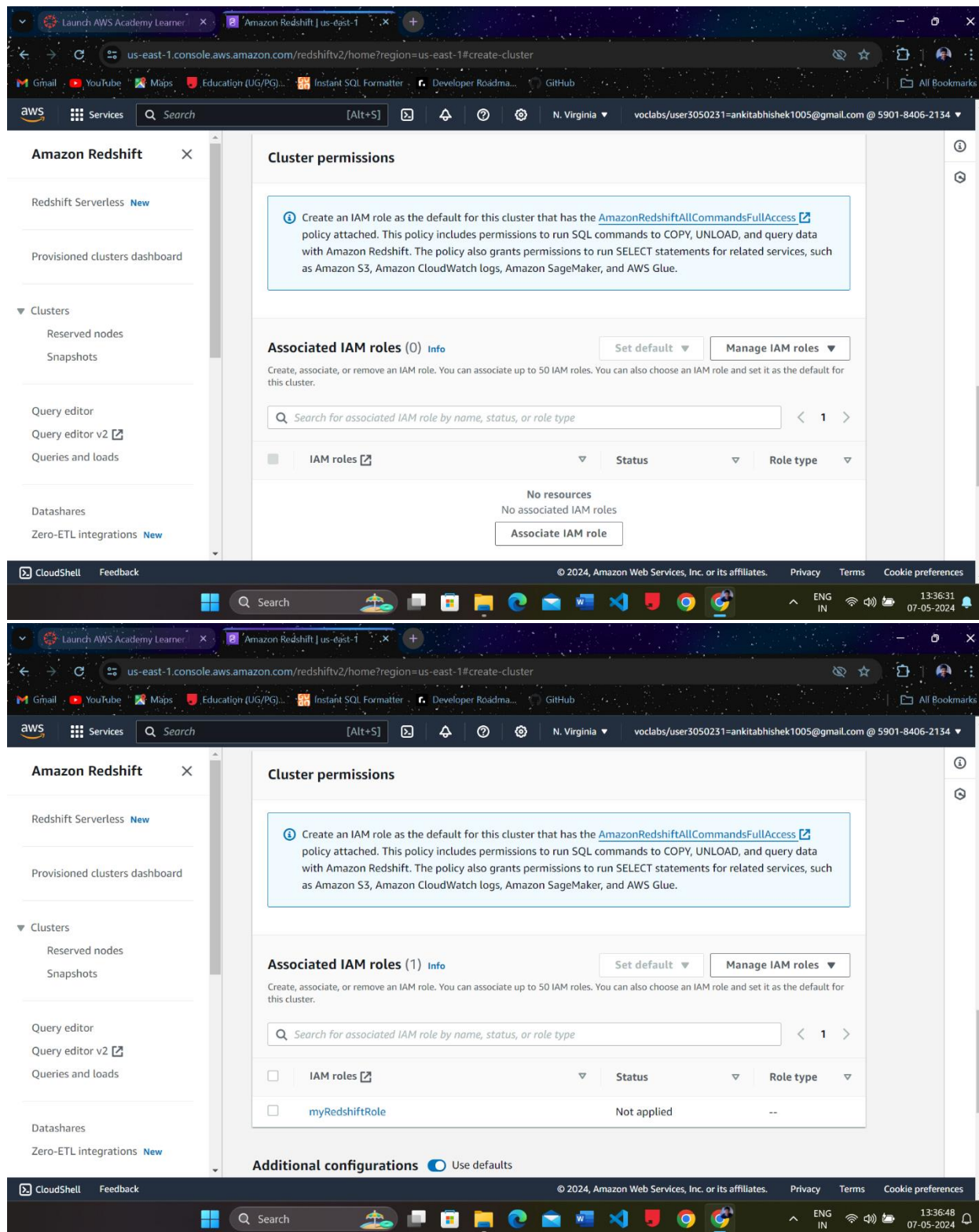
Screenshot of the type of machine used along with number of nodes



The screenshot shows the Amazon Redshift console's 'Create cluster' page. The left sidebar contains navigation options like 'Redshift Serverless', 'Provisioned clusters dashboard', 'Clusters', 'Query editor', and 'Datashares'. The main content area is titled 'Create cluster' and includes a 'Cluster configuration' section. In this section, the 'Cluster identifier' is 'sparknordetlproject', the 'Node type' is 'dc2.large', and the 'Number of nodes' is set to 2. A 'Configuration summary' box displays the estimated on-demand compute price as \$365.00/month and the total compressed storage as 320 GB. Below this, there is a 'Sample data' section with a checkbox for 'Load sample data'.

Screenshot of the redshift database creation associated with IAM roles:





Amazon Redshift

Redshift Serverless **New**

Provisioned clusters dashboard

Clusters

- Reserved nodes
- Snapshots

Query editor

Query editor v2

Queries and loads

Datashares

Zero-ETL integrations **New**

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 (0) [Info](#) Set default Manage IAM roles

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.

Search for associated IAM role by name, status, or role type

<input type="checkbox"/>	IAM roles	Status	Role type
No resources			
No associated IAM roles			
Associate IAM role			

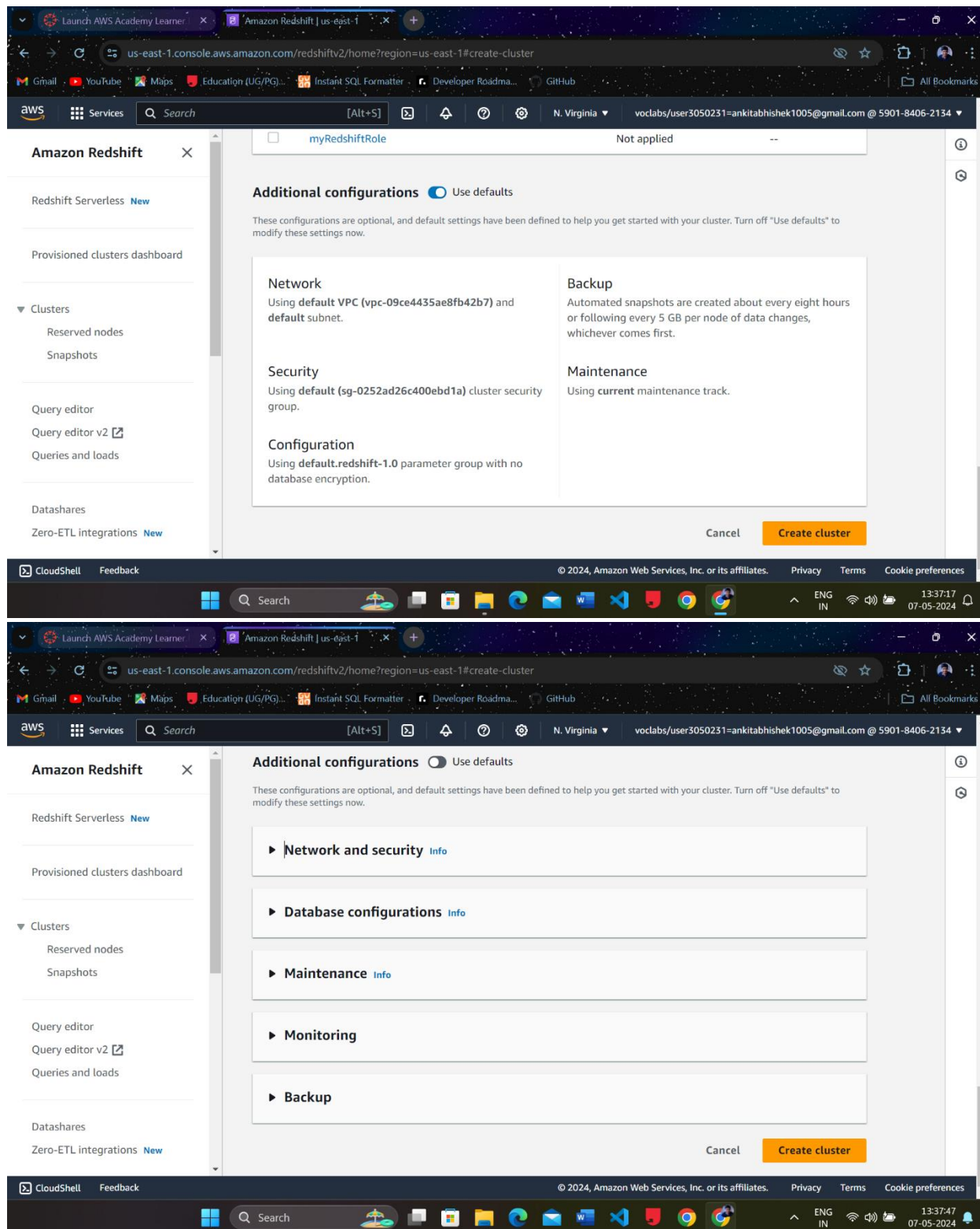
Associated IAM roles (1) [Info](#) Set default Manage IAM roles

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.

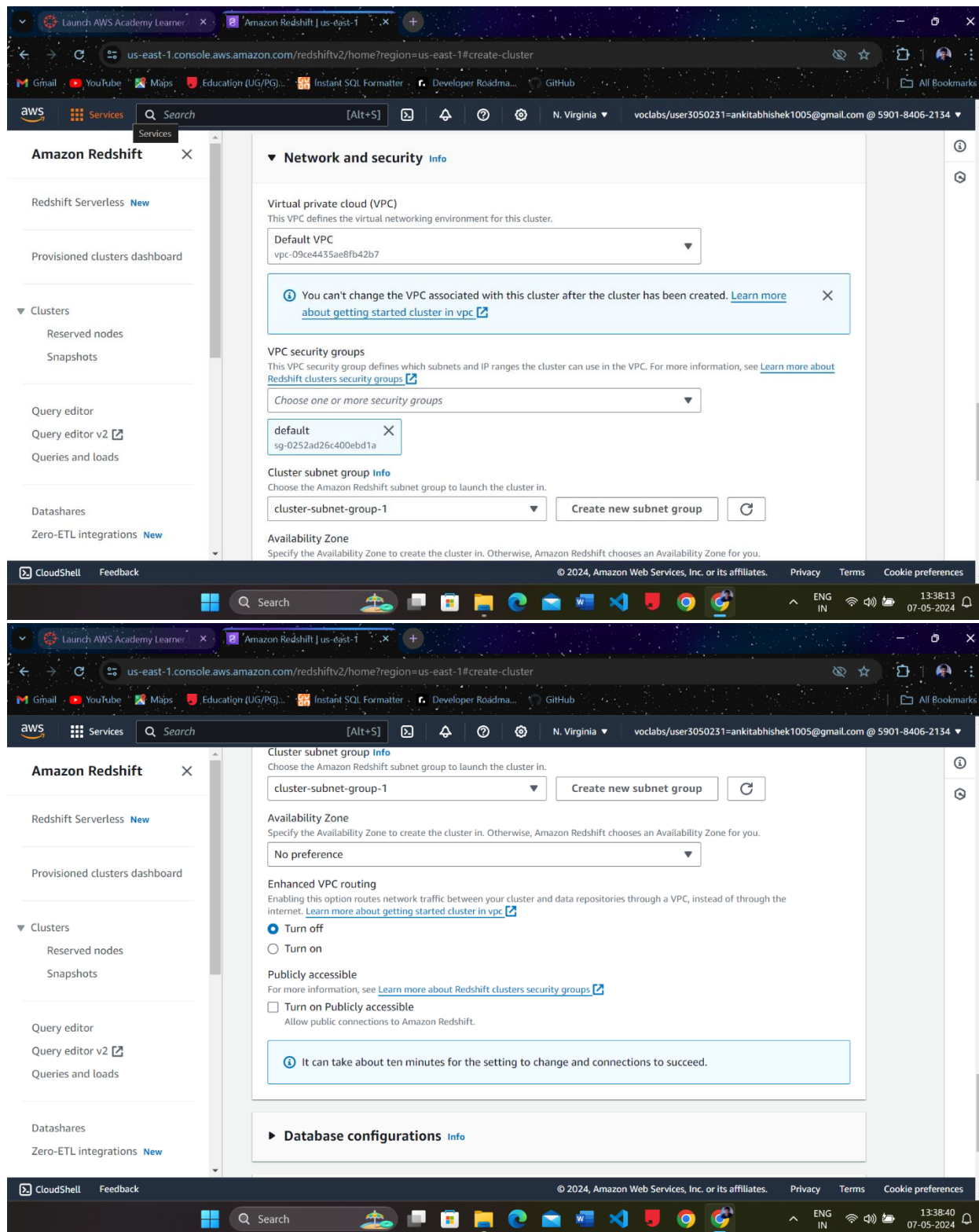
Search for associated IAM role by name, status, or role type

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

Additional configurations ☒ Use defaults



The screenshot displays the AWS Redshift console interface for creating a new cluster. The left sidebar shows the navigation menu with options like 'Redshift Serverless', 'Provisioned clusters dashboard', 'Clusters', 'Query editor', and 'Datashares'. The main content area is titled 'Additional configurations' and shows the 'Use defaults' option selected. Below this, there are five configuration sections: Network, Security, Configuration, Backup, and Maintenance. Each section provides a brief description of the default settings. At the bottom right, there are 'Cancel' and 'Create cluster' buttons. The browser's address bar shows the URL 'us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#create-cluster'.



Amazon Redshift

Redshift Serverless **New**

Provisioned clusters dashboard

Clusters

- Reserved nodes
- Snapshots

Query editor

Query editor v2

Queries and loads

Datashares

Zero-ETL integrations **New**

Network and security Info

Virtual private cloud (VPC)

This VPC defines the virtual networking environment for this cluster.

Default VPC
vpc-09ce4435ae8fb42b7

You can't change the VPC associated with this cluster after the cluster has been created. [Learn more](#) about getting started cluster in vpc

VPC security groups

This VPC security group defines which subnets and IP ranges the cluster can use in the VPC. For more information, see [Learn more about Redshift clusters security groups](#)

Choose one or more security groups

default
sg-0252ad26c400ebd1a

Cluster subnet group Info

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

cluster-subnet-group-1

Create new subnet group

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 about getting started cluster in vpc](#)

Turn off

Publicly accessible

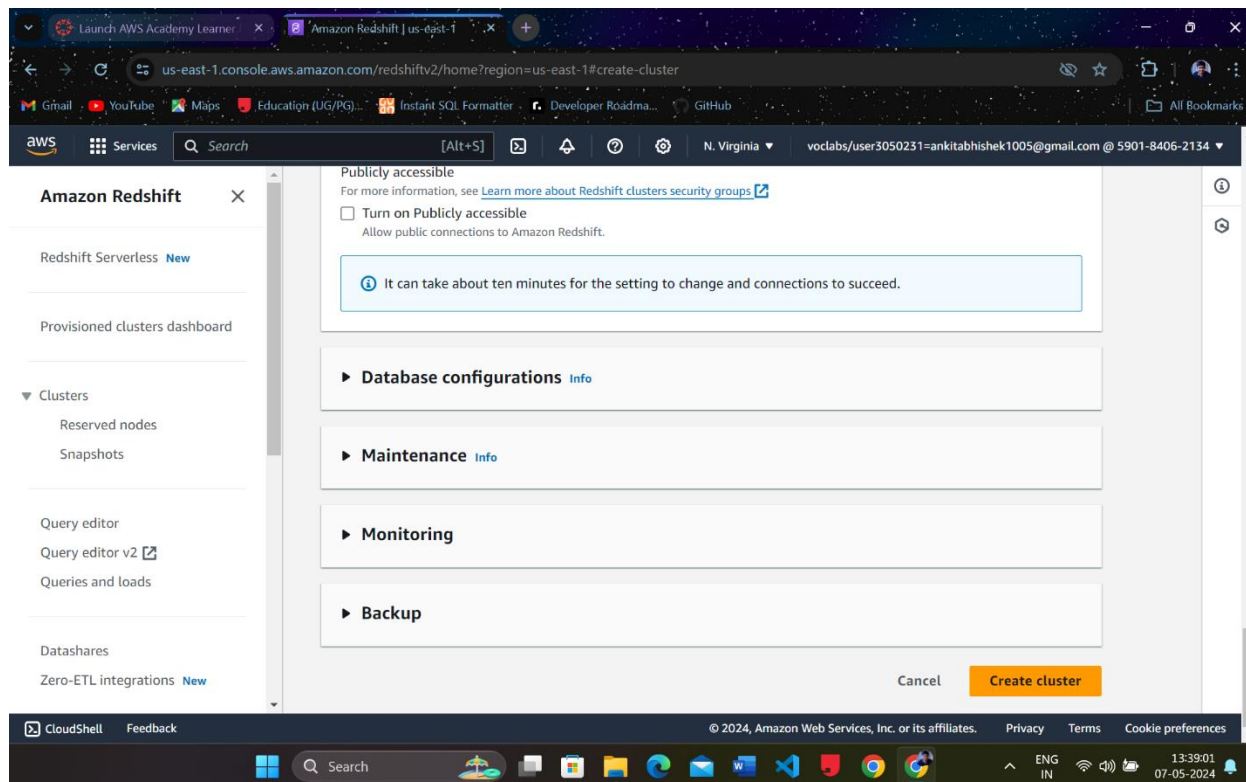
For more information, see [Learn more about Redshift clusters security groups](#)

Turn on Publicly accessible

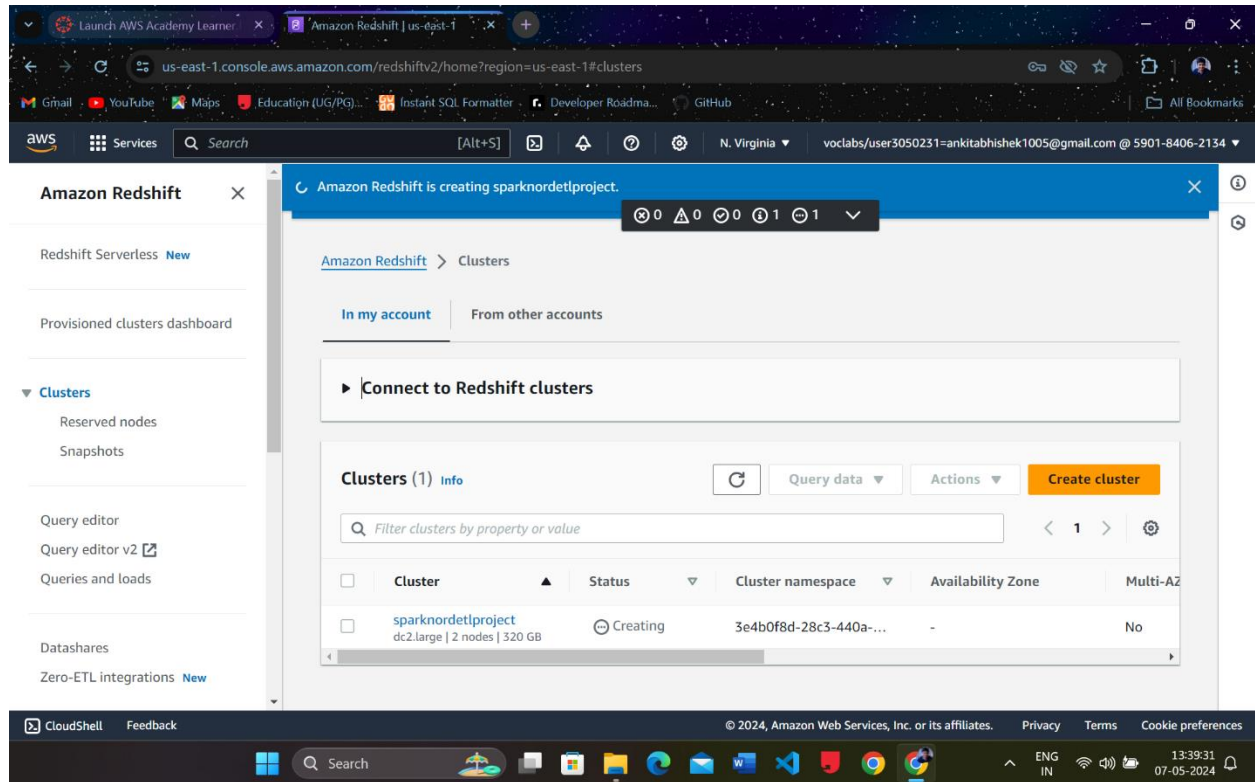
Allow public connections to Amazon Redshift.

It can take about ten minutes for the setting to change and connections to succeed.

Database configurations Info

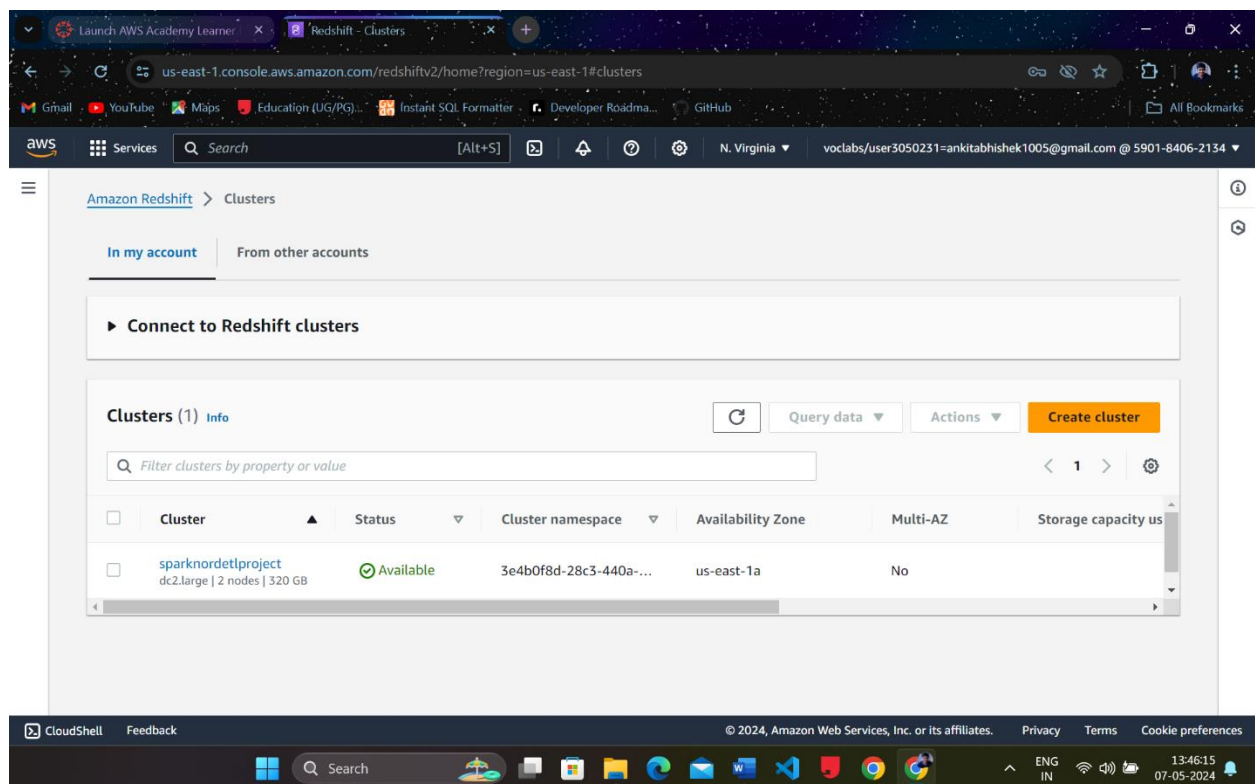


Screenshot of the redshift being created and its status getting changed from creating to available:



The screenshot shows the Amazon Redshift console interface. A blue banner at the top states "Amazon Redshift is creating sparknordetlproject." The left sidebar contains navigation links for "Redshift Serverless", "Provisioned clusters dashboard", "Clusters", "Reserved nodes", "Snapshots", "Query editor", "Query editor v2", "Queries and loads", "Datashares", and "Zero-ETL integrations". The main content area displays the "Clusters" page with a table showing one cluster in the "Creating" state.

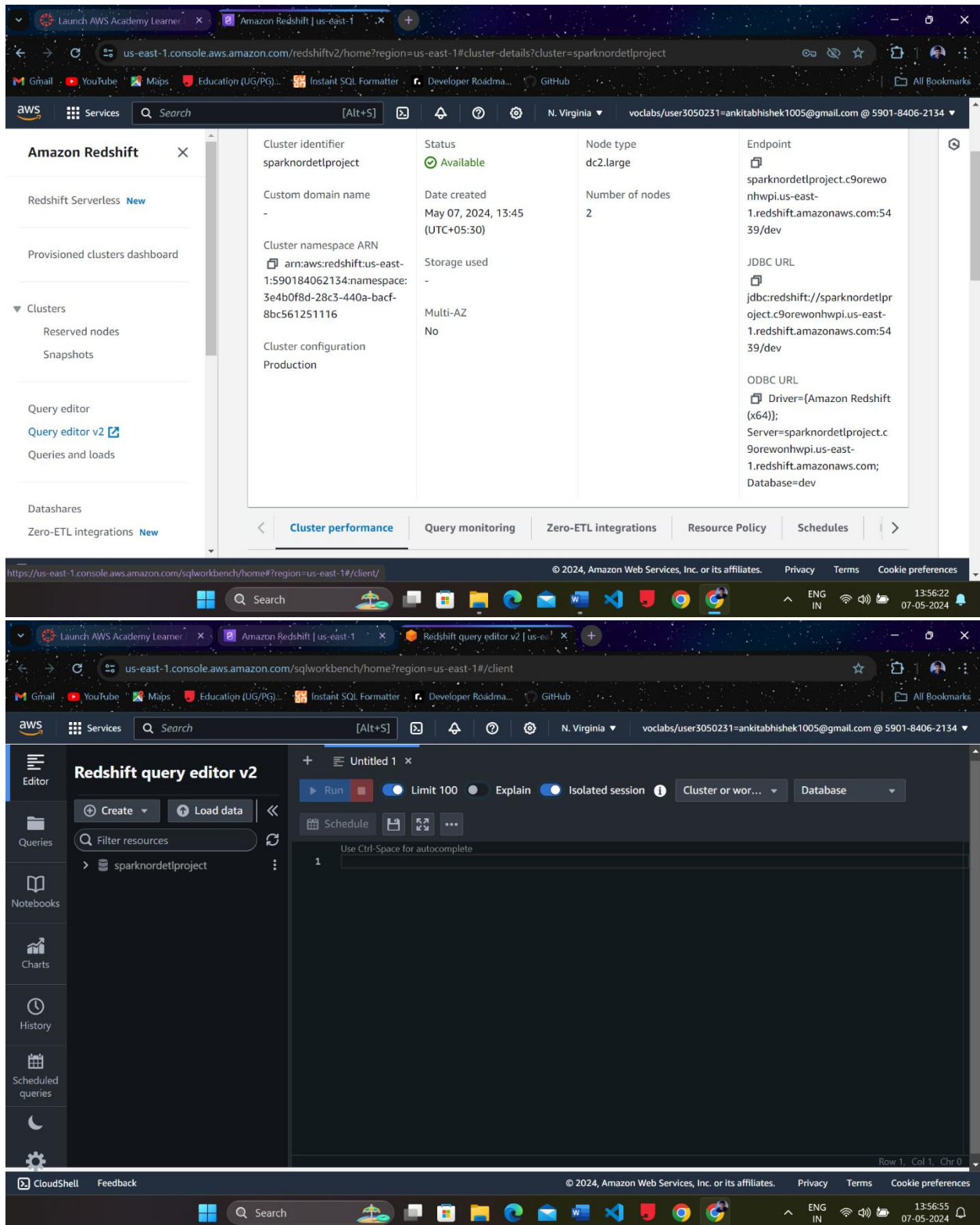
Cluster	Status	Cluster namespace	Availability Zone	Multi-AZ
sparknordetlproject dc2.large 2 nodes 320 GB	Creating	3e4b0f8d-28c3-440a-...	-	No



The screenshot shows the Amazon Redshift console interface after the cluster has been created. The status of the cluster "sparknordetlproject" has changed to "Available". The table below shows the updated cluster details.

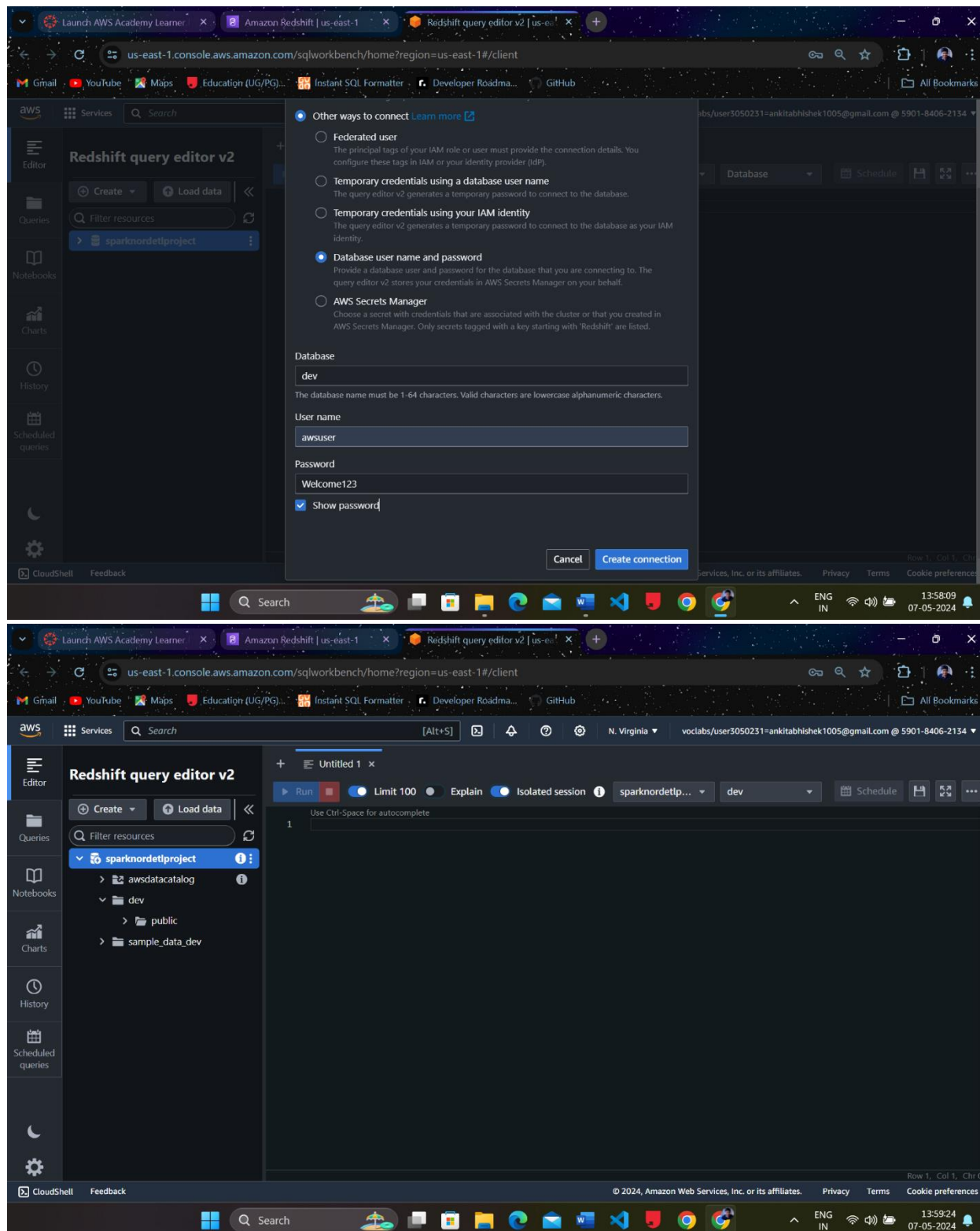
Cluster	Status	Cluster namespace	Availability Zone	Multi-AZ	Storage capacity us
sparknordetlproject dc2.large 2 nodes 320 GB	Available	3e4b0f8d-28c3-440a-...	us-east-1a	No	

Setting up a database in the Redshift cluster and running queries to create the dimension and fact tables



The screenshot displays the AWS Redshift console interface. The top section shows the 'Cluster identifier' as 'sparknordetlproject', which is in an 'Available' status. Other details include 'Node type' as 'dc2.large', 'Number of nodes' as '2', and the 'Endpoint' as 'sparknordetlproject.c9orewonhwp1.us-east-1.redshift.amazonaws.com:5439/dev'. The 'JDBC URL' and 'ODBC URL' are also provided.

The bottom section shows the 'Redshift query editor v2' interface. The left sidebar contains navigation options like 'Editor', 'Queries', 'Notebooks', 'Charts', 'History', and 'Scheduled queries'. The main area is titled 'Untitled 1' and includes a 'Run' button, 'Limit 100', 'Explain', 'Isolated session', and 'Cluster or wor...' dropdown. The query editor is currently empty, with a prompt to 'Use Ctrl-Space for autocomplete'.



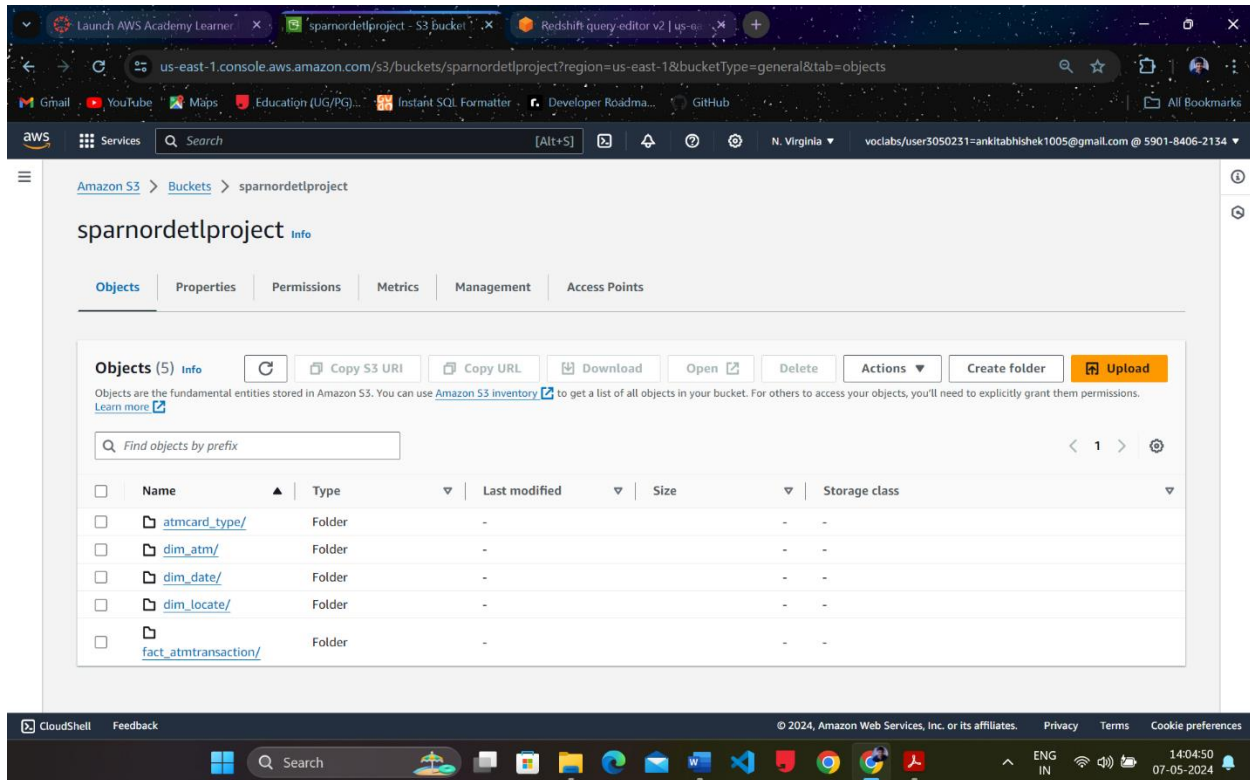
The screenshot displays the AWS Redshift query editor v2 interface. The top navigation bar shows the AWS logo, a search bar, and the current region (us-east-1). The left sidebar contains navigation options: Editor, Queries, Notebooks, Charts, History, and Scheduled queries. The main content area is titled 'Redshift query editor v2' and features a 'Create' button and a 'Load data' button. A modal dialog box is open, titled 'Other ways to connect', which lists several connection methods:

- ☐ Federated user: The principal tags of your IAM role or user must provide the connection details. You configure these tags in IAM or your identity provider (IdP).
- ☐ Temporary credentials using a database user name: The query editor v2 generates a temporary password to connect to the database.
- ☐ Temporary credentials using your IAM identity: The query editor v2 generates a temporary password to connect to the database as your IAM identity.
- ☒ Database user name and password: Provide a database user and password for the database that you are connecting to. The query editor v2 stores your credentials in AWS Secrets Manager on your behalf.
- ☐ AWS Secrets Manager: Choose a secret with credentials that are associated with the cluster or that you created in AWS Secrets Manager. Only secrets tagged with a key starting with 'Redshift' are listed.

Below the list, the 'Database' field is set to 'dev'. The 'User name' field is set to 'awsuser'. The 'Password' field is set to 'Welcome123', and the 'Show password' checkbox is checked. The 'Create connection' button is visible at the bottom right of the dialog box.

The bottom screenshot shows the same interface after the connection has been created. The 'Create' button is now disabled, and the 'Load data' button is active. The 'Database' dropdown is set to 'dev', and the 'User name' dropdown is set to 'awsuser'. The 'Password' field is now empty. The 'Show password' checkbox is unchecked. The 'Create connection' button is now disabled, and the 'Load data' button is active.

Viewing all the data in Amazon S3 bucket:



Amazon S3 > Buckets > sparnordetlproject

sparnordetlproject [Info](#)

Objects (5) [Info](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

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](#)

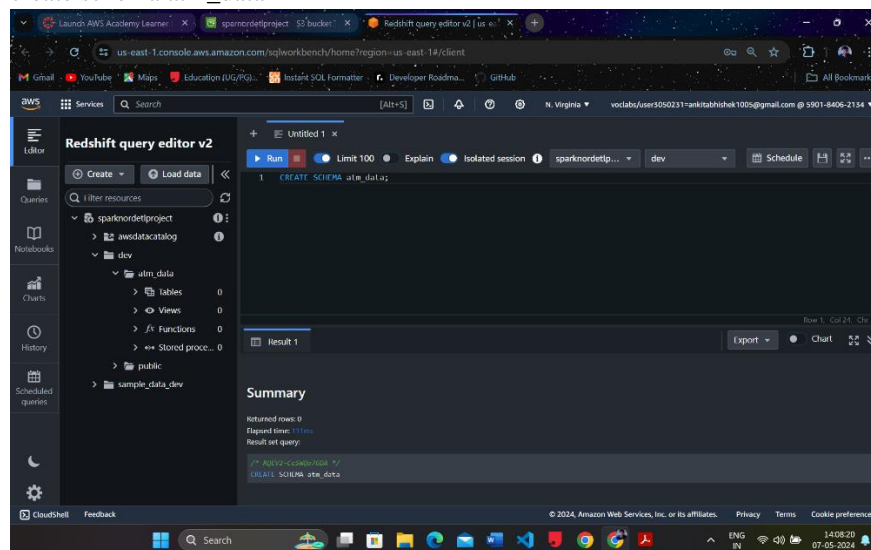
Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	atmcard_type/	Folder	-	-	-
<input type="checkbox"/>	dim_atm/	Folder	-	-	-
<input type="checkbox"/>	dim_date/	Folder	-	-	-
<input type="checkbox"/>	dim_locate/	Folder	-	-	-
<input type="checkbox"/>	fact_atmtransaction/	Folder	-	-	-

Queries to create the various dimension and fact tables with appropriate primary and foreign keys:

1. **SCHEMA CREATION** within awsuser database of the sparknordetlproject redshift cluster :

create schema atm_data



Redshift query editor v2

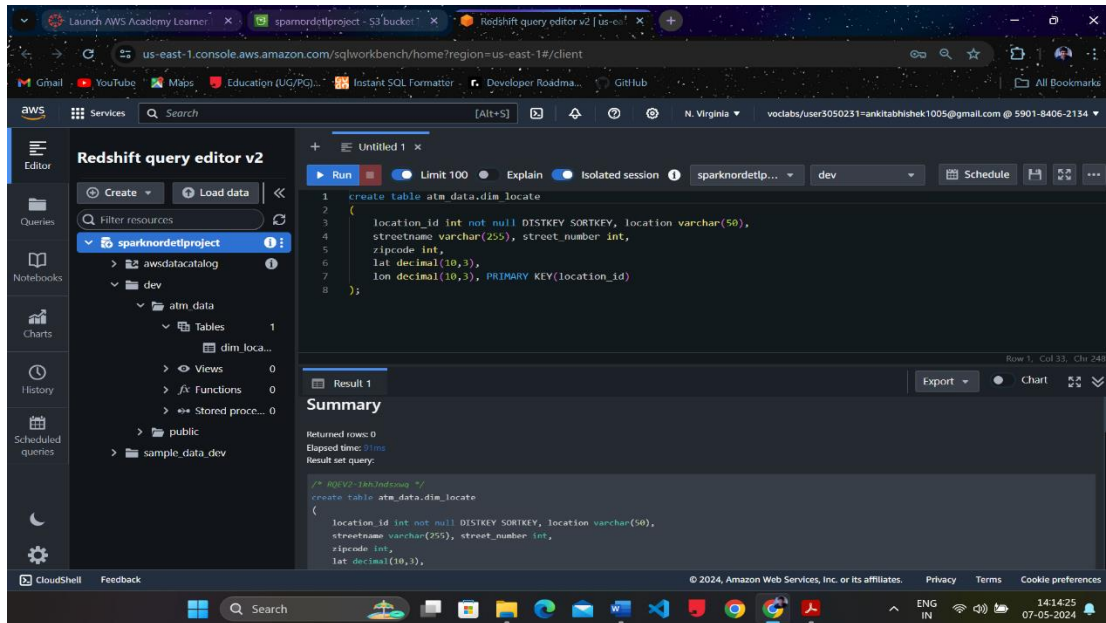
CREATE SCHEMA atm_data;

Summary

Returned rows: 0
Elapsed time: 111ms
Result set query:

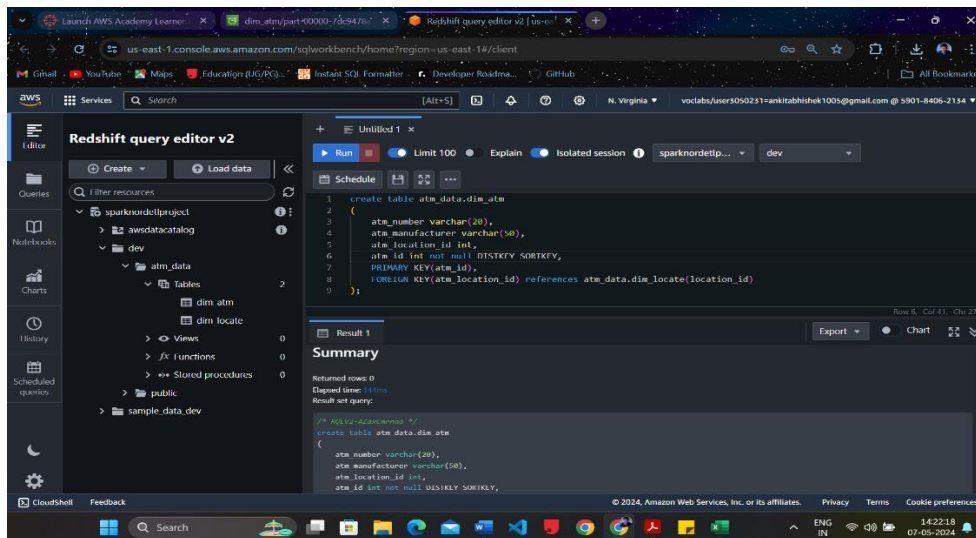
2. Creating location dimension table:

```
create table atm_data.dim_locate
(
    location_id int not null DISTKEY SORTKEY, location varchar(50),
    streetname varchar(255), street_number int,
    zipcode int,
    lat decimal(10,3),
    lon decimal(10,3), PRIMARY KEY(location_id)
);
```



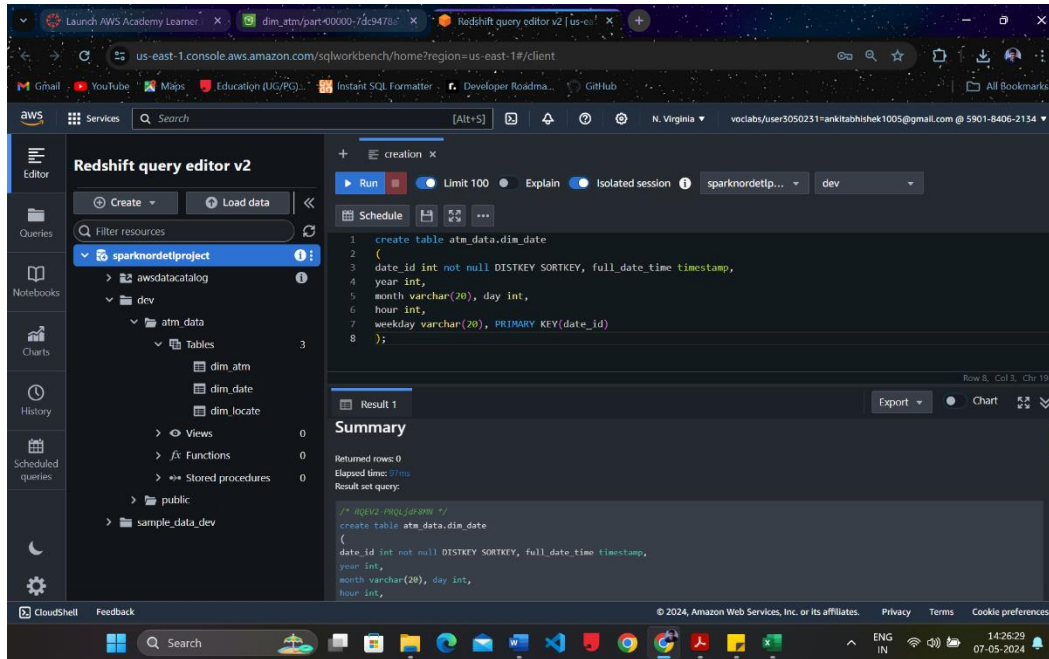
3. Creating atm dimension table:

```
create table atm_data.dim_atm
(
    atm_number varchar(20), atm_manufacturer varchar(50), atm_location_id int,
    atm_id int not null DISTKEY SORTKEY, PRIMARY KEY(atm_id),
    FOREIGN KEY(atm_location_id) references atm_data.dim_locate(location_id)
);
```



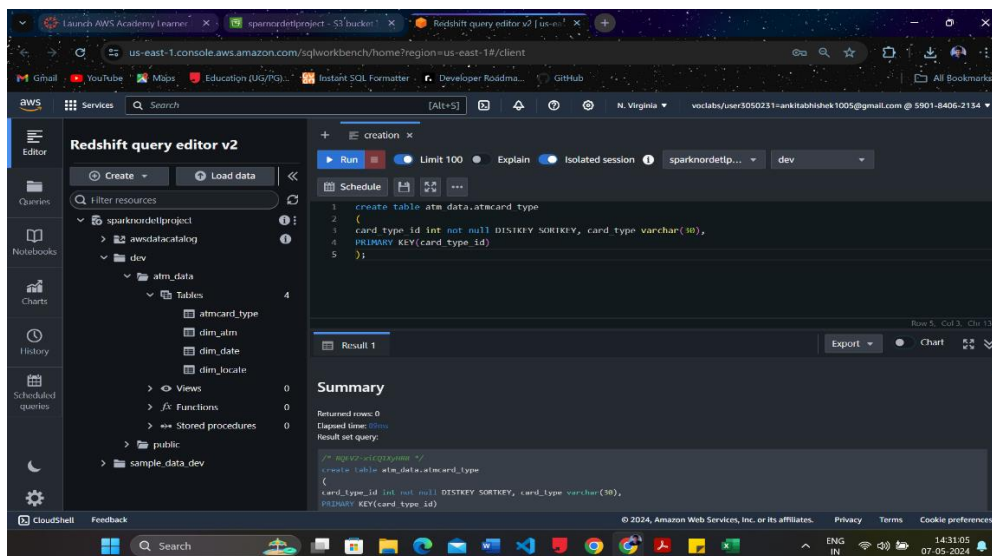
4. Creating date dimension table:

```
create table atm_data.dim_date
(
    date_id int not null DISTKEY SORTKEY, full_date_time timestamp,
    year int, month varchar(20), day int, hour int,
    weekday varchar(20), PRIMARY KEY(date_id)
);
```



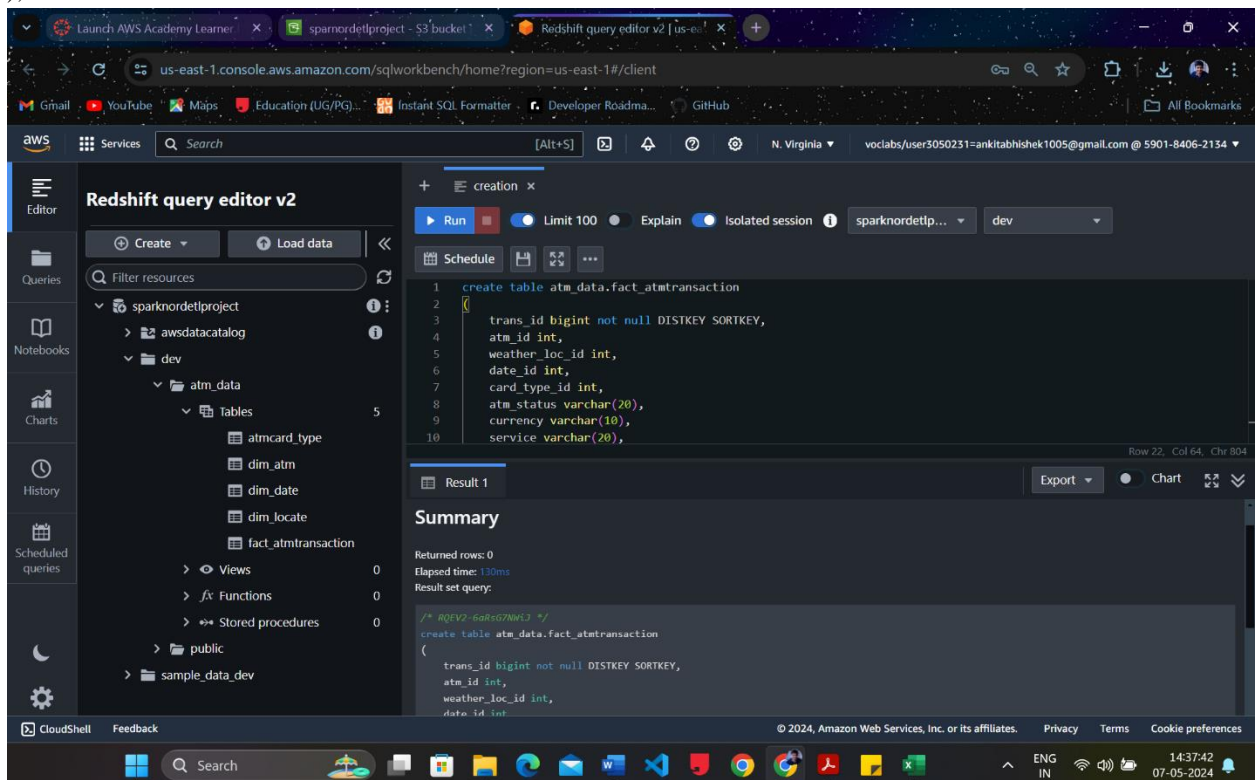
5. Creating atmcard_type dimension table:

```
create table atm_data.atmcard_type
(
  card_type_id int not null DISTKEY SORTKEY, card_type varchar(30),
  PRIMARY KEY(card_type_id)
);
```



6. Creating fact_transaction fact table:

```
create table atm_data.fact_atmtransaction
(
    trans_id bigint not null DISTKEY SORTKEY,
    atm_id int,
    weather_loc_id int,
    date_id int,
    card_type_id int,
    atm_status varchar(20),
    currency varchar(10),
    service varchar(20),
    transaction_amount int,
    message_code varchar(225),
    message_text varchar(225),
    rain_3h decimal(10,3),
    clouds_all int,
    weather_id int,
    weather_main varchar(50),
    weather_description varchar(255),
    PRIMARY KEY(trans_id),
    FOREIGN KEY(weather_loc_id) references atm_data.dim_locate(location_id),
    FOREIGN KEY(atm_id) references atm_data.dim_atm(atm_id),
    FOREIGN KEY(date_id) references atm_data.dim_date(date_id),
    FOREIGN KEY(card_type_id) references atm_data.atmcard_type(card_type_id)
);
```



Loading data into a Redshift cluster from Amazon S3 bucket

Queries to copy the data from S3 buckets to the Redshift cluster in the appropriate tables

1. Copying data to dim_locate table :

copy atm_data.dim_locate from

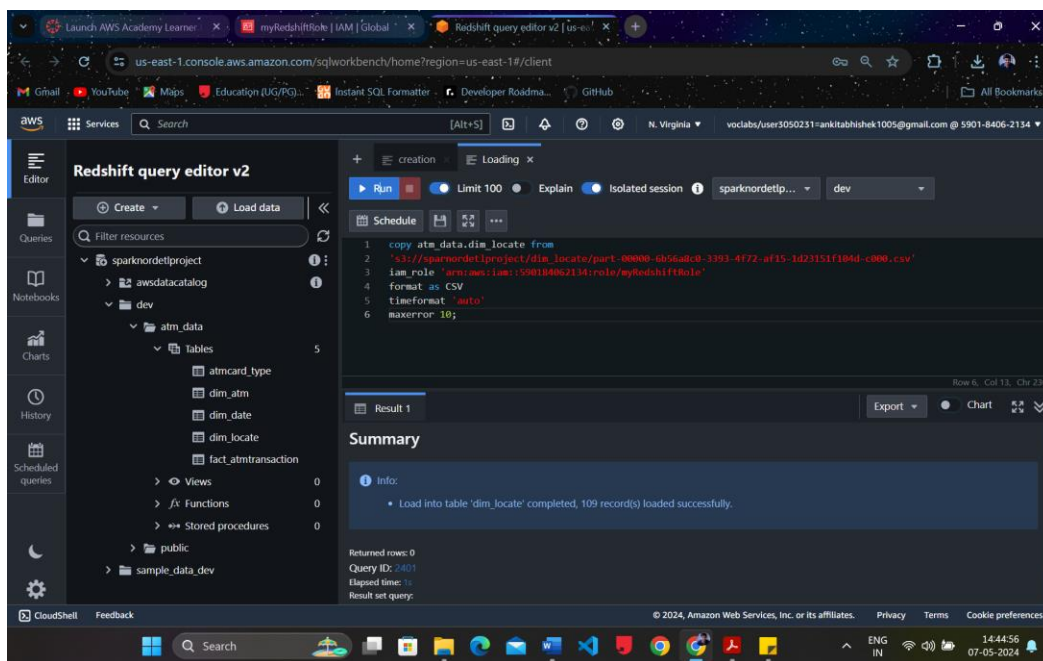
's3://sparnordetlproject/dim_locate/part-00000-6b56a8c0-3393-4f72-af15-1d23151f104d-c000.csv'

iam_role 'arn:aws:iam::590184062134:role/myRedshiftRole'

format as CSV

timeformat 'auto'

maxerror 10;



The screenshot displays the AWS Redshift query editor v2 interface. The left sidebar shows the 'Filter resources' section with a tree view of the 'sparnordetlproject' database, including 'awsdatacatalog', 'dev', 'atm_data', and 'Tables'. The main editor area contains a SQL query to copy data from an S3 bucket into the 'dim_locate' table. The query is as follows:

```
1 copy atm_data.dim_locate from
2 's3://sparnordetlproject/dim_locate/part-00000-6b56a8c0-3393-4f72-af15-1d23151f104d-c000.csv'
3 iam_role 'arn:aws:iam::590184062134:role/myRedshiftRole'
4 format as CSV
5 timeformat 'auto'
6 maxerror 10;
```

Below the query, the 'Summary' section shows the execution status: 'Load into table 'dim_locate' completed, 109 record(s) loaded successfully.' The bottom status bar indicates 'Returned rows: 0', 'Query ID: 2401', and 'Elapsed time: 1s'.

2. Copying data to dim_atm table :

copy atm_data.dim_atm from

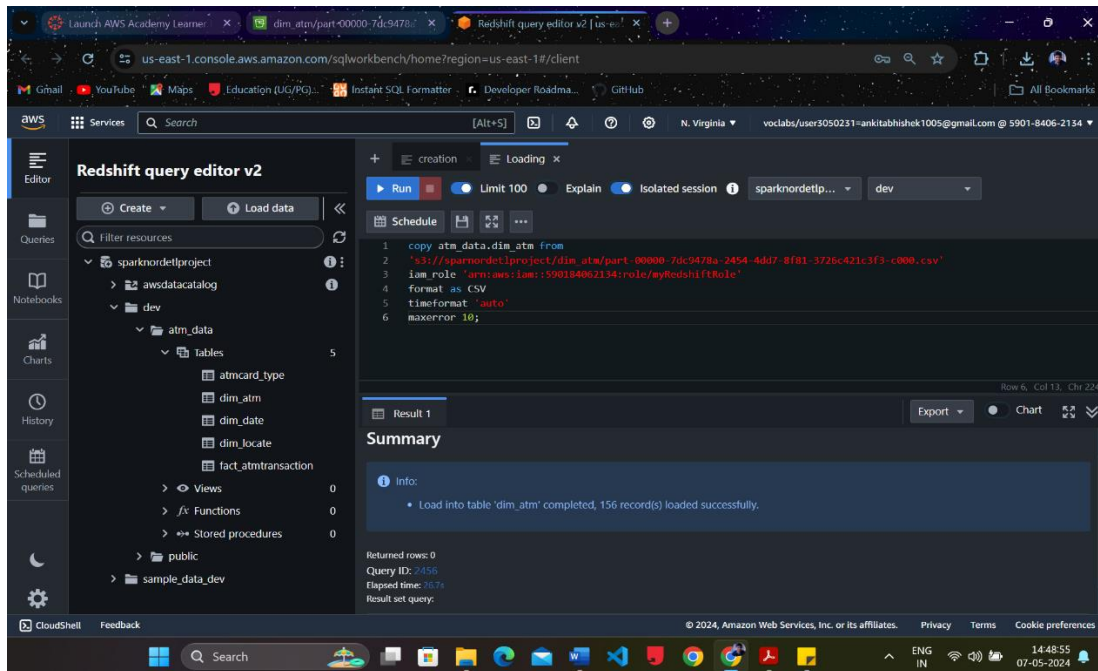
's3://sparknordetlproject/dim_atm/part-00000-7dc9478a-2454-4dd7-8f81-3726c421c3f3-c000.csv'

iam_role 'arn:aws:iam::590184062134:role/myRedshiftRole'

format as CSV

timeformat 'auto'

maxerror 10;



3. Copying data to dim_date table :

copy atm_data.dim_date from

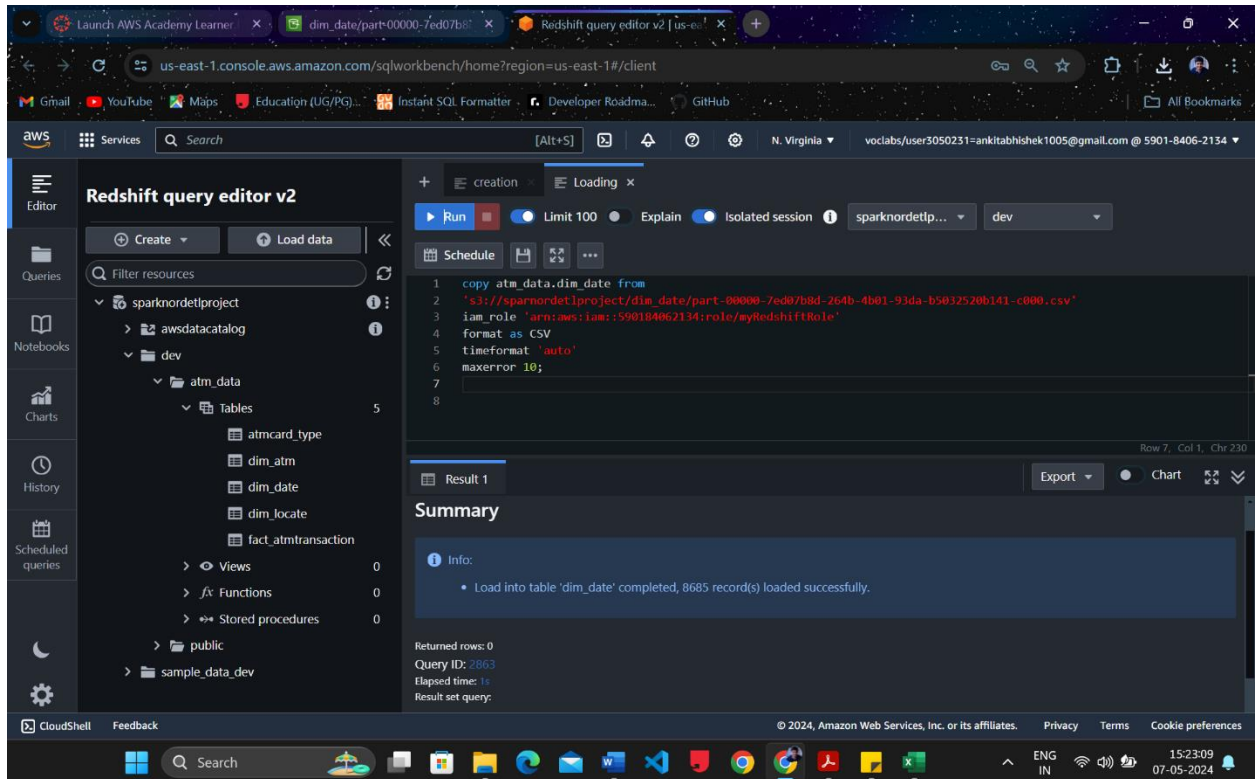
's3://sparknordetlproject/dim_date/part-00000-7ed07b8d-264b-4b01-93da-b5032520b141-c000.csv'

iam_role 'arn:aws:iam::590184062134:role/myRedshiftRole'

format as CSV

timeformat 'auto'

maxerror 10;



The screenshot displays the AWS Redshift Query Editor v2 interface. The left sidebar shows the 'Queries' section with a tree view of the 'sparknordetlproject' database, including 'dev', 'atm_data', and 'Tables'. The main editor area shows a SQL query being executed. The query is as follows:

```
1 copy atm_data.dim_date from
2 's3://sparknordetlproject/dim_date/part-00000-7ed07b8d-264b-4b01-93da-b5032520b141-c000.csv'
3 iam_role 'arn:aws:iam::590184062134:role/myRedshiftRole'
4 format as CSV
5 timeformat 'auto'
6 maxerror 10;
7
8
```

The 'Summary' section below the query shows the following information:

- Info: Load into table 'dim_date' completed, 8685 record(s) loaded successfully.
- Returned rows: 0
- Query ID: 2863
- Elapsed time: 1s
- Result set query:

The bottom of the interface shows the AWS CloudShell terminal and the Windows taskbar with various application icons.

4. Copying data to atmcard_type table :

copy atm_data.atmcard_type from

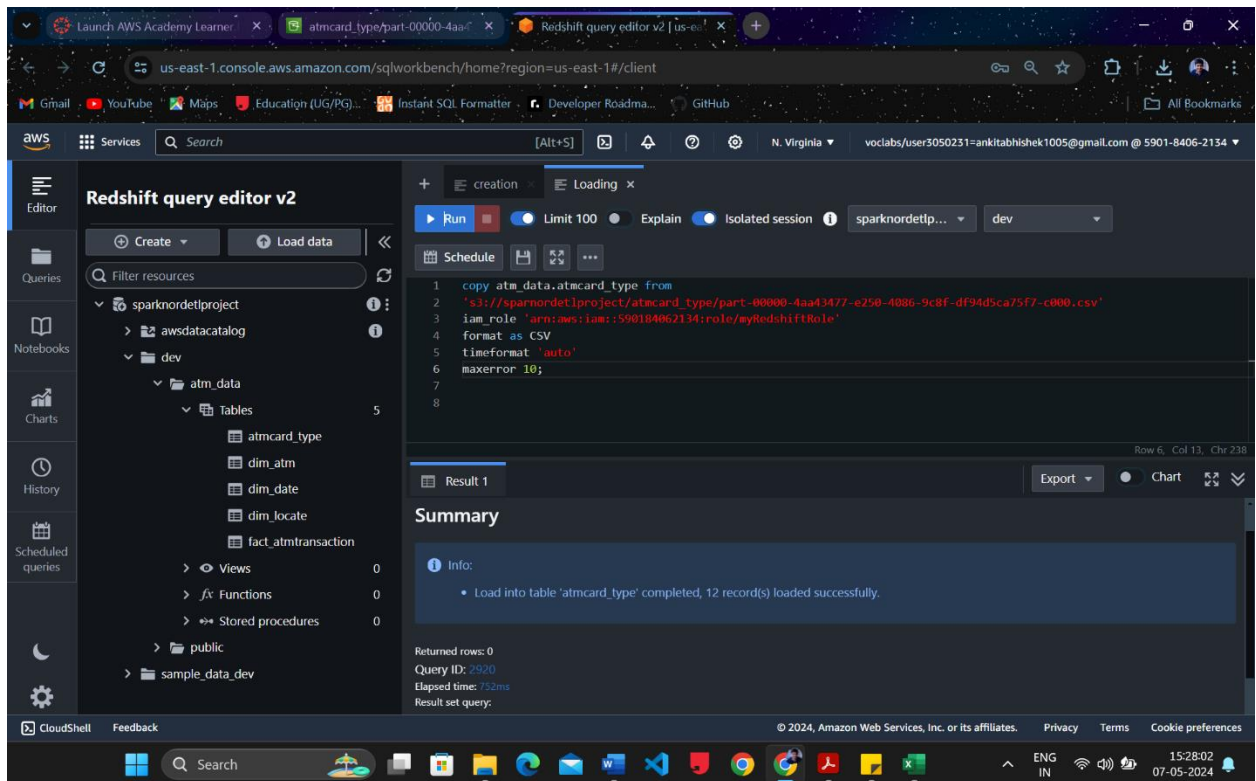
's3://sparknordetlproject/atmcard_type/part-00000-4aa43477-e250-4086-9c8f-df94d5ca75f7-c000.csv'

iam_role 'arn:aws:iam::590184062134:role/myRedshiftRole'

format as CSV

timeformat 'auto'

maxerror 10;



The screenshot displays the AWS Redshift Query Editor v2 interface. The left sidebar shows the 'Queries' section with a tree view of the 'sparknordetlproject' database, including 'dev', 'atm_data', and 'Tables'. The main editor area shows a SQL query being executed. The query is as follows:

```
1 copy atm_data.atmcard_type from
2 's3://sparknordetlproject/atmcard_type/part-00000-4aa43477-e250-4086-9c8f-df94d5ca75f7-c000.csv'
3 iam_role 'arn:aws:iam::590184062134:role/myRedshiftRole'
4 format as CSV
5 timeformat 'auto'
6 maxerror 10;
7
8
```

The 'Summary' section at the bottom indicates that the query was successful, with the message: 'Load into table 'atmcard_type' completed, 12 record(s) loaded successfully.' The interface also shows the 'Result 1' section, which is currently empty, and the 'Info' section, which provides details about the query execution, including the query ID (2920) and the elapsed time (752ms).

5. Copying data to fact_transaction fact table:

copy atm_data.fact_atmtransaction from

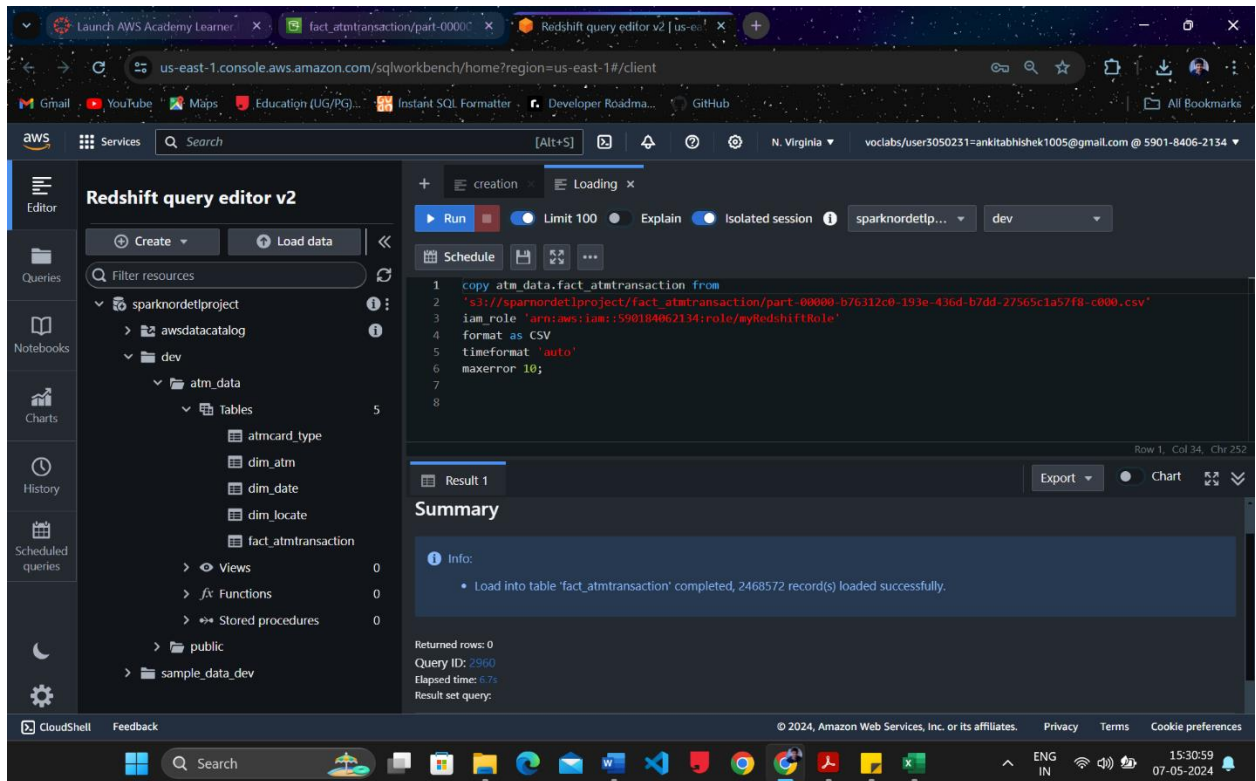
's3://sparknordetlproject/fact_atmtransaction/part-00000-b76312c0-193e-436d-b7dd-27565c1a57f8-c000.csv'

iam_role 'arn:aws:iam::590184062134:role/myRedshiftRole'

format as CSV

timeformat 'auto'

maxerror 10;



The screenshot displays the AWS Redshift query editor v2 interface. The left sidebar shows the 'Editor' tab with a file explorer for 'sparknordetlproject' containing 'awsdatacatalog', 'dev', and 'atm_data' folders. The 'atm_data' folder is expanded, showing a 'Tables' section with 5 items: 'atmcard_type', 'dim_atm', 'dim_date', 'dim_location', and 'fact_atmtransaction'. The 'fact_atmtransaction' table is selected. The main editor area shows a SQL query:

```
1 copy atm_data.fact_atmtransaction from
2 's3://sparknordetlproject/fact_atmtransaction/part-00000-b76312c0-193e-436d-b7dd-27565c1a57f8-c000.csv'
3 iam_role 'arn:aws:iam::590184062134:role/myRedshiftRole'
4 format as CSV
5 timeformat 'auto'
6 maxerror 10;
7
8
```

The query is executed, and the 'Summary' section shows the following information:

- Info: Load into table 'fact_atmtransaction' completed, 2468572 record(s) loaded successfully.
- Returned rows: 0
- Query ID: 2960
- Elapsed time: 6.7s
- Result set query:

The bottom status bar indicates the user is logged in as 'vociabs/user3050231=ankitabhishek1005@gmail.com @ 5901-8406-2134' in the 'N. Virginia' region. The footer shows the copyright notice: '© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences'.