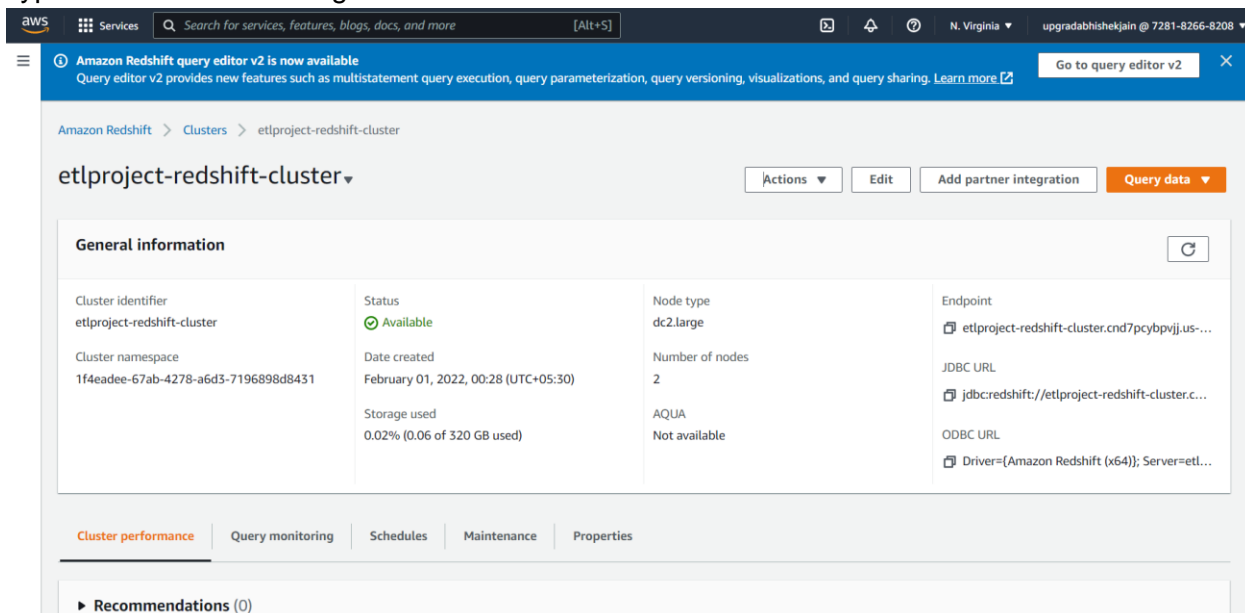


## Creation of a Redshift Cluster

### Screenshots of the configuration of the Redshift cluster that you have created:

Type of machine : dc2.large



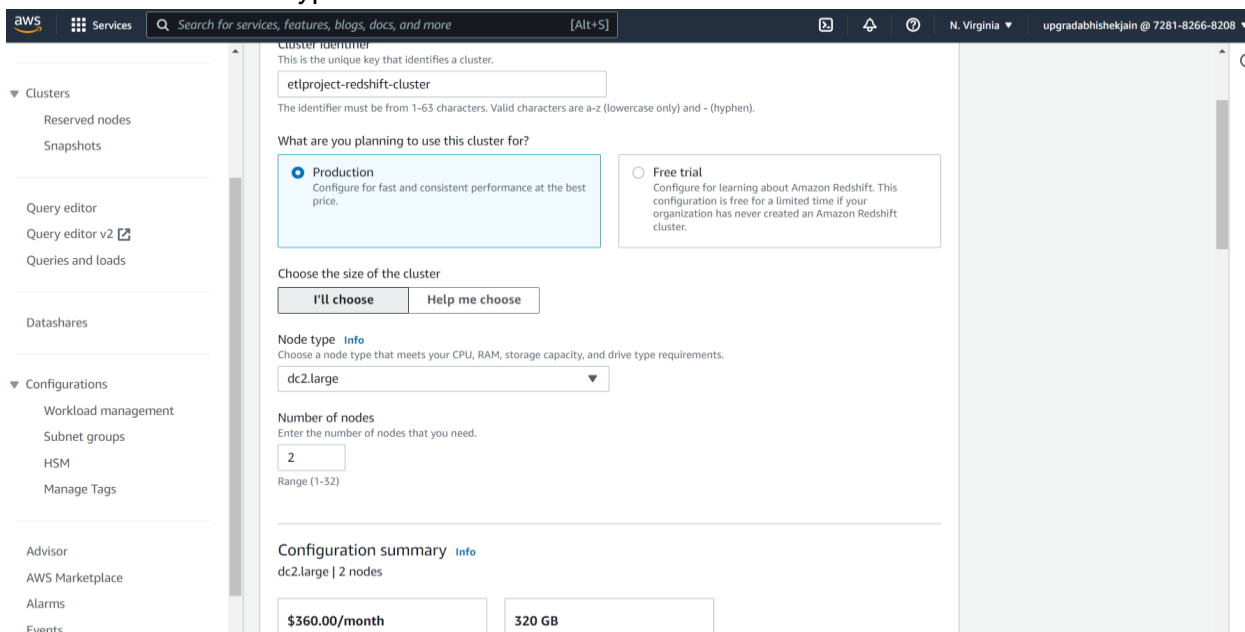
The screenshot shows the Amazon Redshift console interface. At the top, there's a navigation bar with the AWS logo, 'Services' link, a search bar, and user information. Below the navigation bar, a blue banner announces 'Amazon Redshift query editor v2 is now available'. The main content area shows the cluster 'etlproject-redshift-cluster' with a dropdown menu for 'Actions', 'Edit', 'Add partner integration', and 'Query data'. The 'General information' tab is selected, displaying a table with the following details:

General information			
Cluster identifier etlproject-redshift-cluster	Status Available	Node type dc2.large	Endpoint etlproject-redshift-cluster.cnd7pcybpvj.us-...
Cluster namespace 1f4eadee-67ab-4278-a6d3-7196898d8431	Date created February 01, 2022, 00:28 (UTC+05:30)	Number of nodes 2	JDBC URL jdbc:redshift://etlproject-redshift-cluster.c...
	Storage used 0.02% (0.06 of 320 GB used)	AQUA Not available	ODBC URL Driver={Amazon Redshift (x64)}; Server=etl...

Below the table, there are tabs for 'Cluster performance', 'Query monitoring', 'Schedules', 'Maintenance', and 'Properties'. At the bottom, there's a 'Recommendations (0)' section.

### Screenshot of various configurations while setting up Red shift cluster

#### Cluster use and node type



The screenshot shows the 'Create new cluster' wizard in the Amazon Redshift console. The left sidebar contains navigation links for 'Clusters', 'Reserved nodes', 'Snapshots', 'Query editor', 'Query editor v2', 'Queries and loads', 'Datashares', 'Configurations', 'Workload management', 'Subnet groups', 'HSM', 'Manage Tags', 'Advisor', 'AWS Marketplace', 'Alarms', and 'Events'. The main content area is titled 'Cluster identifier' and includes a text input field with 'etlproject-redshift-cluster'. Below this, there's a section 'What are you planning to use this cluster for?' with two radio buttons: 'Production' (selected) and 'Free trial'. The 'Production' option is highlighted with a blue border. Below this, there's a section 'Choose the size of the cluster' with a button 'I'll choose' and a button 'Help me choose'. The 'Node type' section shows a dropdown menu with 'dc2.large' selected. The 'Number of nodes' section has a text input field with '2' and a range '(1-32)'. At the bottom, there's a 'Configuration summary' section showing 'dc2.large | 2 nodes' with a cost of '\$360.00/month' and a storage capacity of '320 GB'.

## Database Configurations

▼ Clusters

Reserved nodes

Snapshots

Query editor

Query editor v2

Queries and loads

Datashares

▼ Configurations

### 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**

☐ **Show 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 `@`.

## IAM Roles

aws

Services

Search for services, features, blogs, docs, and more

[Alt+S]

N. Virginia

upgradabhishekjain @ 7281-8266-8208

▼ Clusters

Reserved nodes

Snapshots

Query editor

Query editor v2

Queries and loads

Datashares

▼ Configurations

Workload management

Subnet groups

HSM

Manage Tags

### 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.

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.

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

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

<input type="checkbox"/>	<a href="#">IAM roles</a>	Status	Role type
<input type="checkbox"/>	<a href="#">upgrad_redshift_s3</a>	Not applied	--

### Network and security selection VPC and subnets

Clusters

Reserved nodes

Snapshots

Query editor

Query editor v2

Queries and loads

Datashares

Configurations

Workload management

Subnet groups

HSM

Manage Tags

Advisor

Virtual private cloud (VPC)

This VPC defines the virtual networking environment for this cluster.

myvpc

vpc-0470837f3df5b43b

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

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-035ab25b422f834e3

Cluster subnet group

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

cluster-subnet-group-etlproject

Availability Zone

Specify the Availability Zone that you want the cluster to be created in. Otherwise, Amazon Redshift chooses an Availability Zone for you.

No preference

## Configuring DB port

Clusters

Reserved nodes

Snapshots

Query editor

Query editor v2

Queries and loads

Datashares

Configurations

Workload management

Subnet groups

HSM

Manage Tags

Advisor

AWS Marketplace

Alarms

Allow instances and devices outside the VPC to connect to your database through the cluster endpoint.

☒ Disable
 ☐ Enable

Database configurations

Database name

Specify a database name to create an additional database.

dev

The name must be 1-64 alphanumeric characters (lowercase only), and it can't be a reserved word.

Database port

Port number where the database accepts inbound connections. You can't change the port after the cluster has been created.

5439

The port must be numeric (1150-65535).

Parameter groups

Defines database parameter and query queues for all the databases.

default.redshift-1.0

Default parameter group for redshift-1.0

Encryption

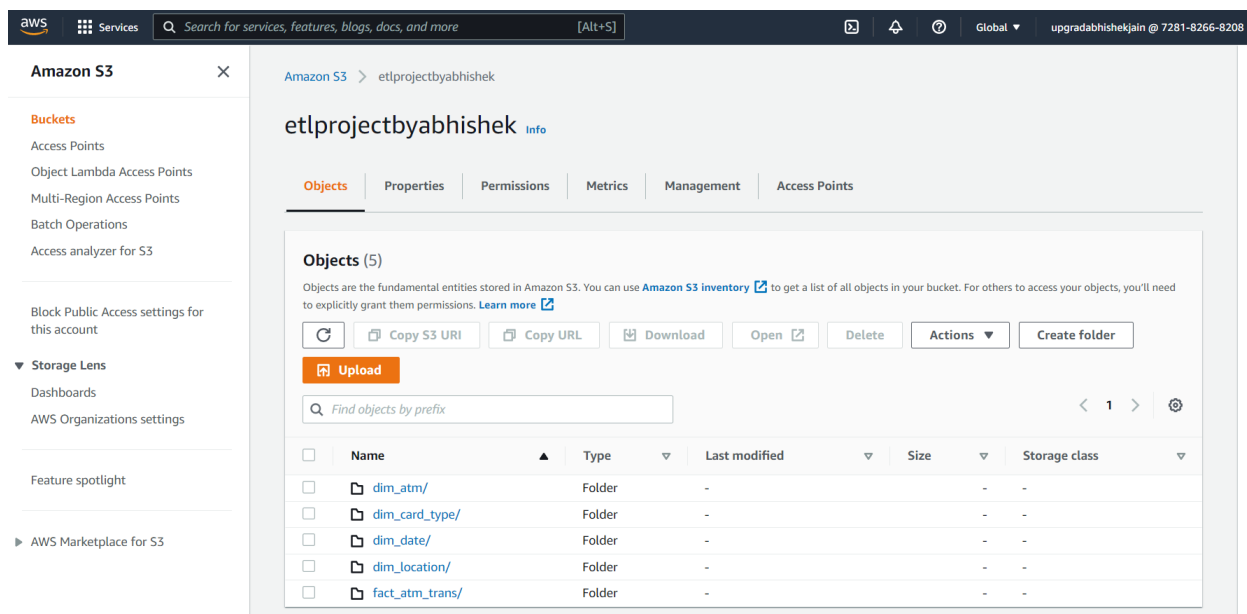
Encrypt all data on your cluster.

☒ Disabled
 ☐ Use AWS Key Management Service (AWS KMS)
 ☐ Use a hardware security module (HSM)

© Copyright. upGrad Education Pvt. Ltd. All rights reserved

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

### View all the data in Amazon S3 Bucket:



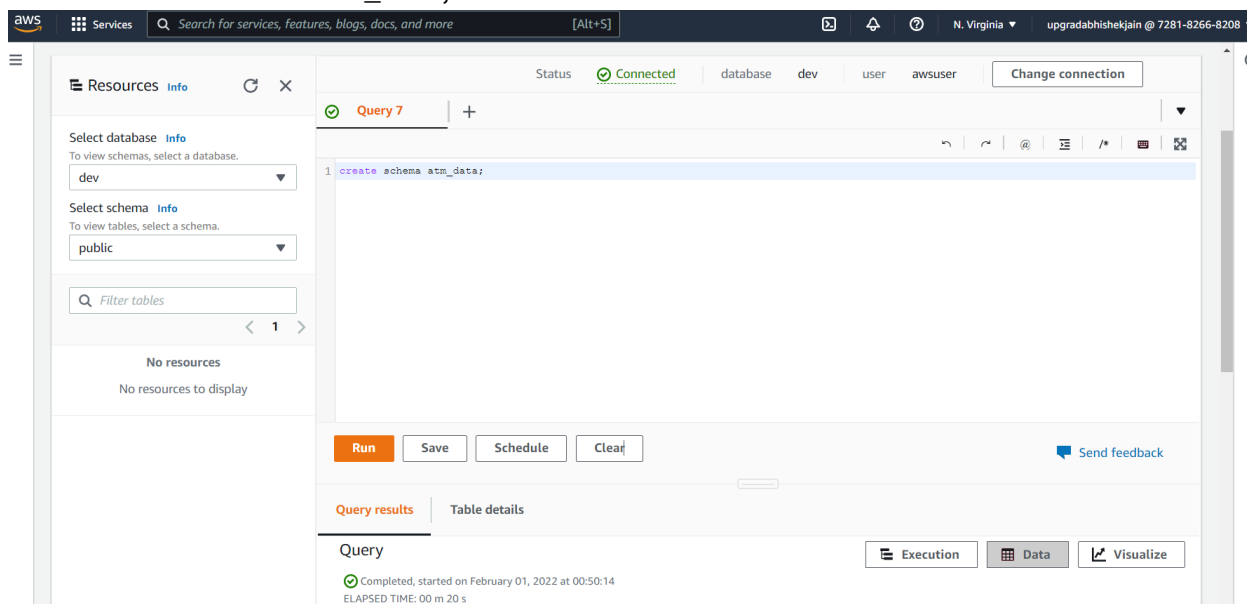
The screenshot shows the Amazon S3 console interface. The left sidebar contains navigation options like Buckets, Access Points, and Storage Lens. The main area displays the 'etlprojectbyabhishek' bucket with a list of 5 objects:

Name	Type	Last modified	Size	Storage class
dim_atm/	Folder	-	-	-
dim_card_type/	Folder	-	-	-
dim_date/	Folder	-	-	-
dim_location/	Folder	-	-	-
fact_atm_trans/	Folder	-	-	-

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

### Create a schema for dimension and fact table

- create schema atm\_data;



The screenshot shows the AWS Redshift console interface. The left sidebar contains navigation options like Resources. The main area displays a query editor with the following SQL query:

```
1 create schema atm_data;
```

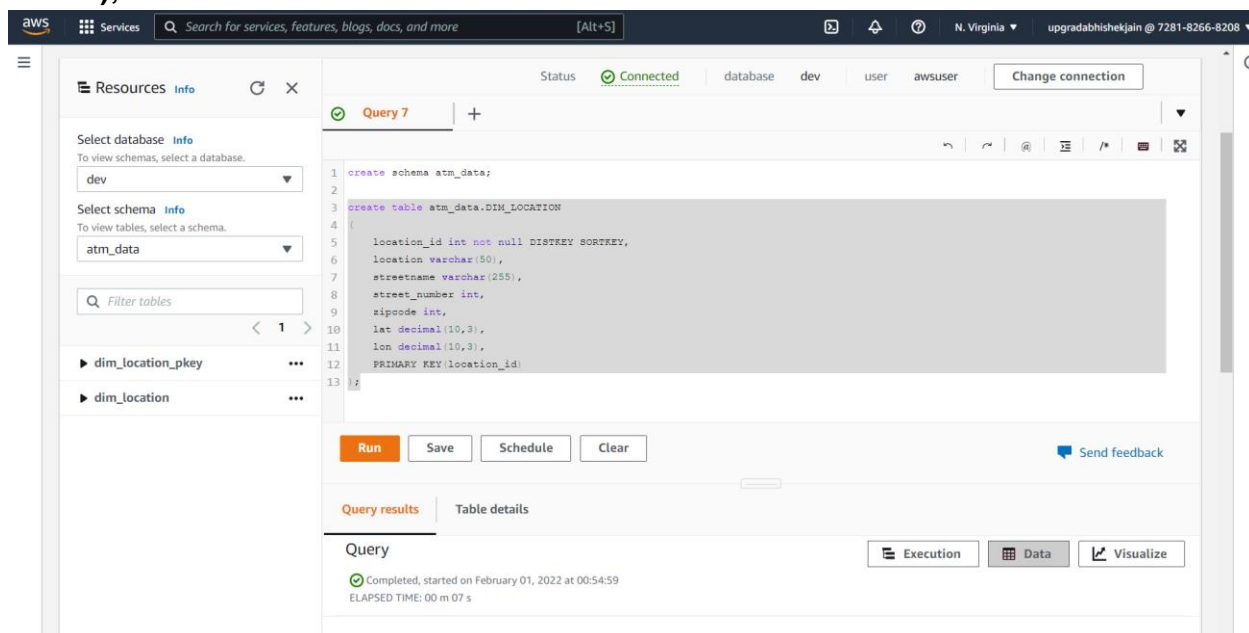
The query is executed successfully, and the results show:

Query results

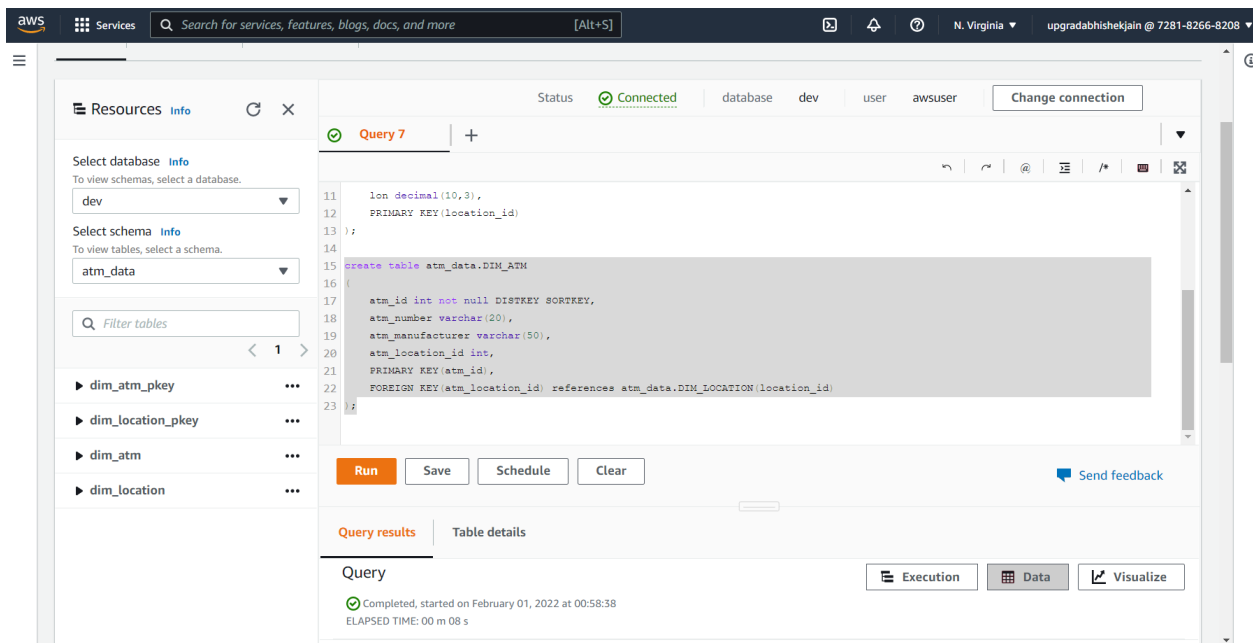
Query

Completed, started on February 01, 2022 at 00:50:14  
ELAPSED TIME: 00 m 20 s

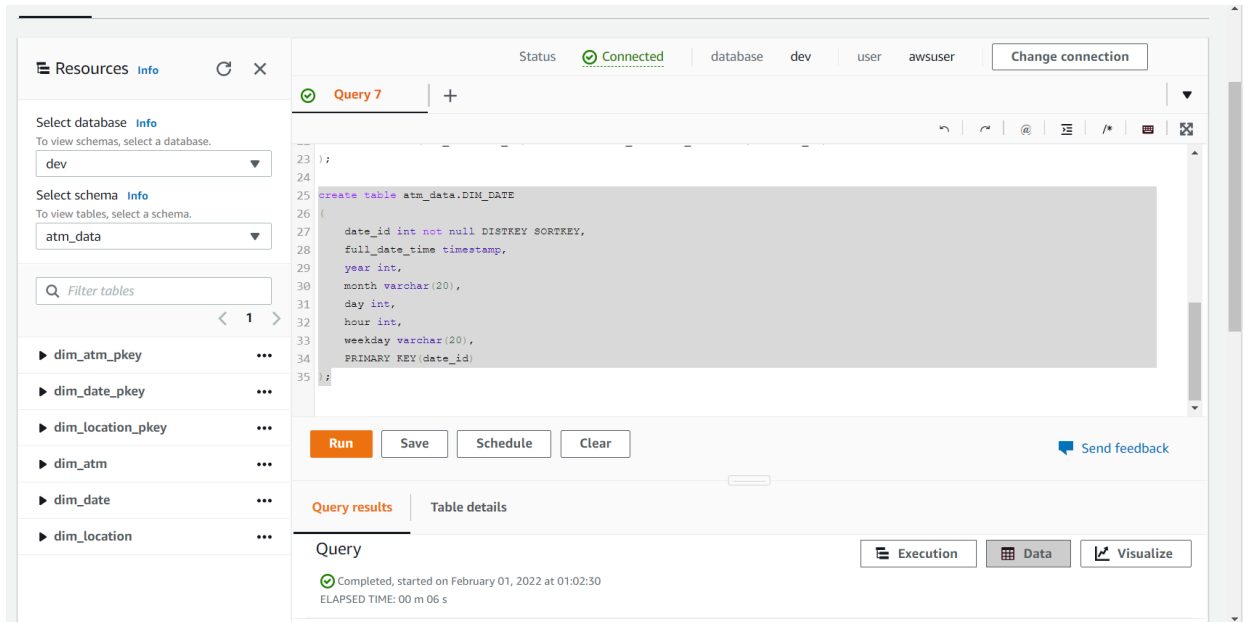
- **Create Dim\_location table**  
**create table atm\_data.DIM\_LOCATION**  
**(**  
     **location\_id int not null DISTKEY SORTKEY,**  
     **location varchar(50),**  
     **streetname varchar(255),**  
     **street\_number int,**  
     **zipcode int,**  
     **lat decimal(10,3),**  
     **lon decimal(10,3),**  
     **PRIMARY KEY(location\_id)**  
**);**



- **Create Dim\_ATM table**  
**create table atm\_data.DIM\_ATM**  
**(**  
     **atm\_id int not null DISTKEY SORTKEY,**  
     **atm\_number varchar(20),**  
     **atm\_manufacturer varchar(50),**  
     **atm\_location\_id int,**  
     **PRIMARY KEY(atm\_id),**  
     **FOREIGN KEY(atm\_location\_id) references**  
     **atm\_data.DIM\_LOCATION(location\_id)**  
**);**



- **Create Dimension date 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)**  
**);**



The screenshot shows the upGrad SQL IDE interface. On the left, the 'Resources' panel displays the database 'dev' and schema 'atm\_data'. The main editor shows a SQL query to create a table named 'atm\_data.DIM\_DATE' with the following schema:

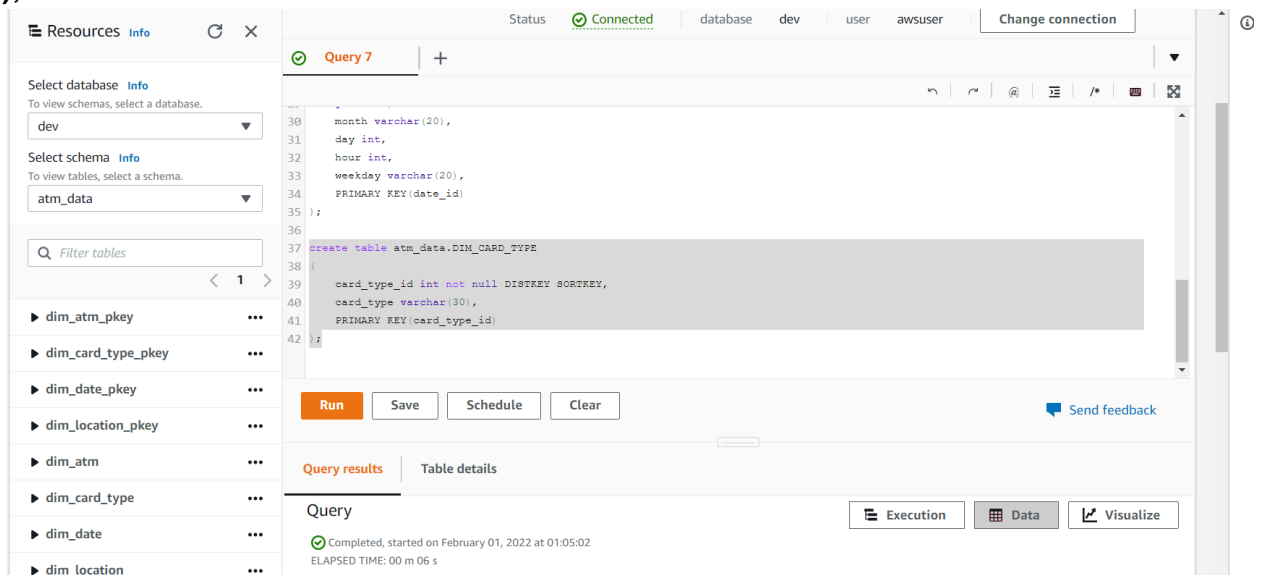
```

23 );
24
25 create table atm_data.DIM_DATE
26 (
27     date_id int not null DISTKEY SORTKEY,
28     full_date_time timestamp,
29     year int,
30     month varchar(20),
31     day int,
32     hour int,
33     weekday varchar(20),
34     PRIMARY KEY(date_id)
35 );

```

Below the editor, the 'Run' button is highlighted. The 'Query results' tab is active, showing a message: 'Completed, started on February 01, 2022 at 01:02:30 ELAPSED TIME: 00 m 06 s'.

- **Create card type dimension table**  
**create table atm\_data.DIM\_CARD\_TYPE**  
**(**  
     **card\_type\_id int not null DISTKEY SORTKEY,**  
     **card\_type varchar(30),**  
     **PRIMARY KEY(card\_type\_id)**  
**);**



The screenshot shows the upGrad SQL IDE interface. On the left, the 'Resources' panel displays the database 'dev' and schema 'atm\_data'. The main editor shows a SQL query to create a table named 'atm\_data.DIM\_CARD\_TYPE' with the following schema:

```

30     month varchar(20),
31     day int,
32     hour int,
33     weekday varchar(20),
34     PRIMARY KEY(date_id)
35 );
36
37 create table atm_data.DIM_CARD_TYPE
38 (
39     card_type_id int not null DISTKEY SORTKEY,
40     card_type varchar(30),
41     PRIMARY KEY(card_type_id)
42 );

```

Below the editor, the 'Run' button is highlighted. The 'Query results' tab is active, showing a message: 'Completed, started on February 01, 2022 at 01:05:02 ELAPSED TIME: 00 m 06 s'.

- Create atm transaction fact table

create table atm\_data.FACT\_ATM\_TRANS

(

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(225),

PRIMARY KEY(trans\_id),

FOREIGN KEY(weather\_loc\_id) references

atm\_data.DIM\_LOCATION(location\_id),

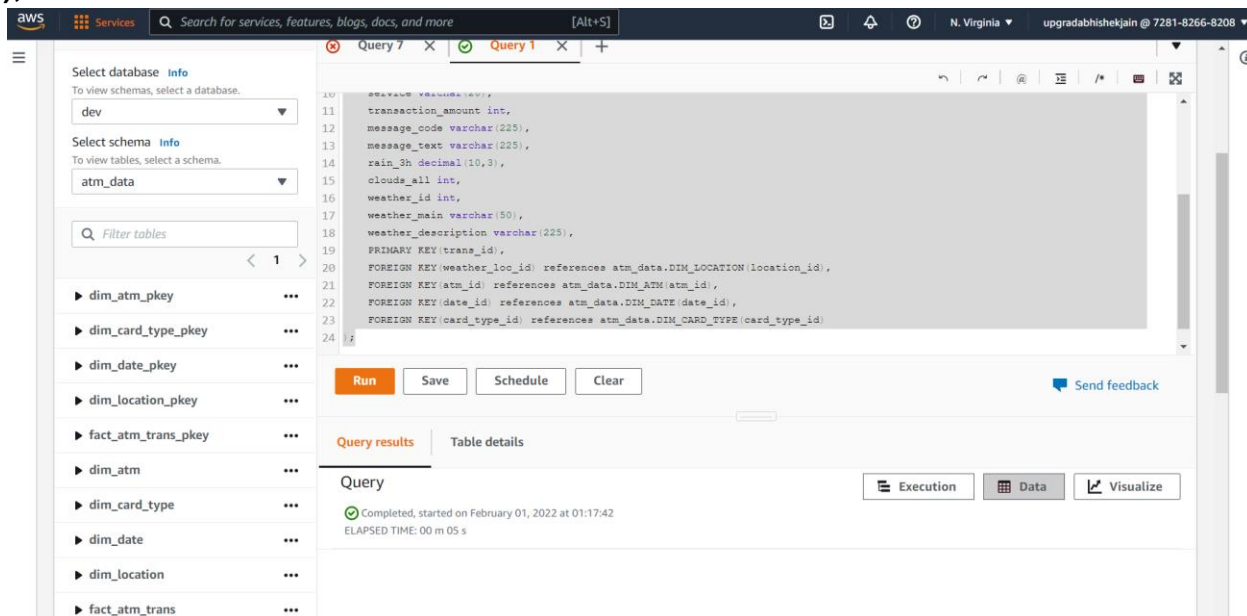
FOREIGN KEY(atm\_id) references atm\_data.DIM\_ATM(atm\_id),

FOREIGN KEY(date\_id) references atm\_data.DIM\_DATE(date\_id),

FOREIGN KEY(card\_type\_id) references

atm\_data.DIM\_CARD\_TYPE(card\_type\_id)

);



The screenshot shows the AWS Glue console interface. On the left, there's a sidebar with 'Select database' (dev) and 'Select schema' (atm\_data). Below this is a list of tables: dim\_atm\_pkey, dim\_card\_type\_pkey, dim\_date\_pkey, dim\_location\_pkey, fact\_atm\_trans\_pkey, dim\_atm, dim\_card\_type, dim\_date, dim\_location, and fact\_atm\_trans. The main area displays a SQL query in the 'Query 1' editor. The query is a CREATE TABLE statement for 'atm\_data.FACT\_ATM\_TRANS' with various columns and foreign key constraints. Below the query editor are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' tab is active, showing a 'Query' section with a green checkmark indicating the query was completed successfully on February 01, 2022, at 01:17:42, with an elapsed time of 00 m 05 s. There are also buttons for 'Execution', 'Data', and 'Visualize'.

```

10 create table atm_data.FACT_ATM_TRANS
11 transaction_amount int,
12 message_code varchar(225),
13 message_text varchar(225),
14 rain_3h decimal(10,3),
15 clouds_all int,
16 weather_id int,
17 weather_main varchar(50),
18 weather_description varchar(225),
19 PRIMARY KEY(trans_id),
20 FOREIGN KEY(weather_loc_id) references atm_data.DIM_LOCATION(location_id),
21 FOREIGN KEY(atm_id) references atm_data.DIM_ATM(atm_id),
22 FOREIGN KEY(date_id) references atm_data.DIM_DATE(date_id),
23 FOREIGN KEY(card_type_id) references atm_data.DIM_CARD_TYPE(card_type_id)
24 ;
  
```



## 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

- Copy data to DIM\_LOCATION and DIM\_ATM table

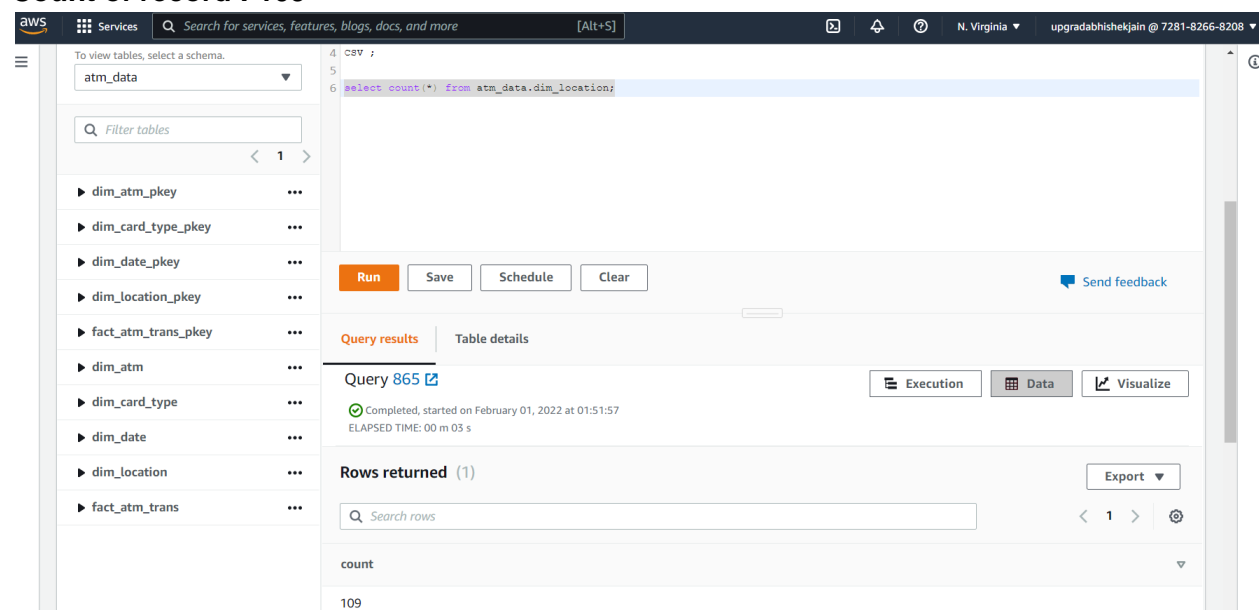
copy atm\_data.dim\_location from 's3://etlprojectbyabhishek/dim\_location/part-00000-2c3f4932-5d7f-4a0f-b030-9b3283149f77-c000.csv'

iam\_role 'arn:aws:iam::728182668208:role/upgrad\_redshift\_s3'

delimiter ',' region 'us-east-1'

CSV;

Count of record : 109



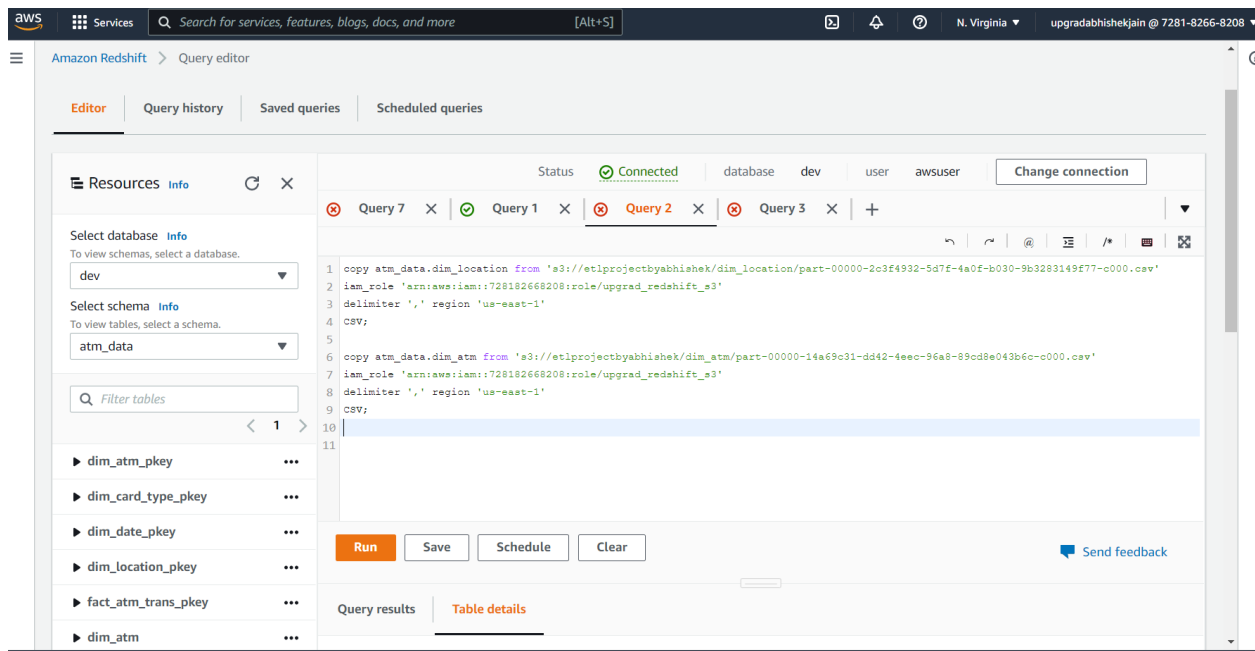
The screenshot shows the AWS Redshift console interface. On the left, there is a sidebar with a list of tables under the 'atm\_data' schema, including 'dim\_atm\_pkey', 'dim\_card\_type\_pkey', 'dim\_date\_pkey', 'dim\_location\_pkey', 'fact\_atm\_trans\_pkey', 'dim\_atm', 'dim\_card\_type', 'dim\_date', 'dim\_location', and 'fact\_atm\_trans'. The main area displays a SQL query: 'select count(\*) from atm\_data.dim\_location;'. Below the query, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' tab is active, showing 'Query 865' which is 'Completed, started on February 01, 2022 at 01:51:57' with an 'ELAPSED TIME: 00 m 03 s'. The 'Rows returned' section shows '(1)' row. The table has one column named 'count' with a value of '109'.

copy atm\_data.dim\_atm from 's3://etlprojectbyabhishek/dim\_atm/part-00000-14a69c31-dd42-4eec-96a8-89cd8e043b6c-c000.csv'

iam\_role 'arn:aws:iam::728182668208:role/upgrad\_redshift\_s3'

delimiter ',' region 'us-east-1'

CSV;



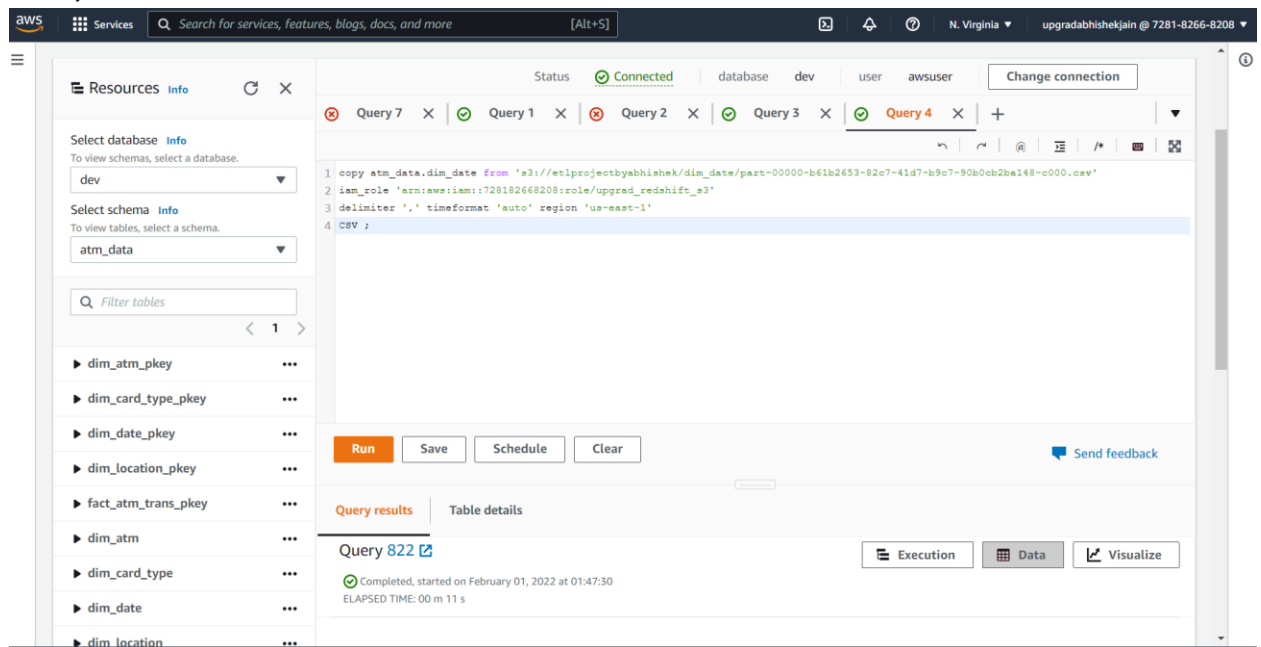
Amazon Redshift Query editor interface showing Query 2. The query is a COPY statement to load data from an S3 bucket into the dim\_location table.

```

1 copy atm_data.dim_location from 's3://etlprojectbyabhishek/dim_location/part-00000-2c3f4932-5d7f-4a0f-b030-9b3283149f77-c000.csv'
2 iam_role 'arn:aws:iam::728182668208:role/upgrad_redshift_s3'
3 delimiter ',' region 'us-east-1'
4 CSV;
5
6 copy atm_data.dim_atm from 's3://etlprojectbyabhishek/dim_atm/part-00000-14a69c31-dd42-4eec-96a8-89cd8e043b6c-c000.csv'
7 iam_role 'arn:aws:iam::728182668208:role/upgrad_redshift_s3'
8 delimiter ',' region 'us-east-1'
9 CSV;
10
11

```

- **Copy data to DIM\_DATE table**  
**copy atm\_data.dim\_date from 's3://etlprojectbyabhishek/dim\_date/part-00000-b61b2653-82c7-41d7-b9c7-90b0cb2ba148-c000.csv'**  
**iam\_role 'arn:aws:iam::728182668208:role/upgrad\_redshift\_s3'**  
**delimiter ',' timeformat 'auto' region 'us-east-1'**  
**CSV ;**



Amazon Redshift Query editor interface showing Query 4. The query is a COPY statement to load data from an S3 bucket into the dim\_date table.

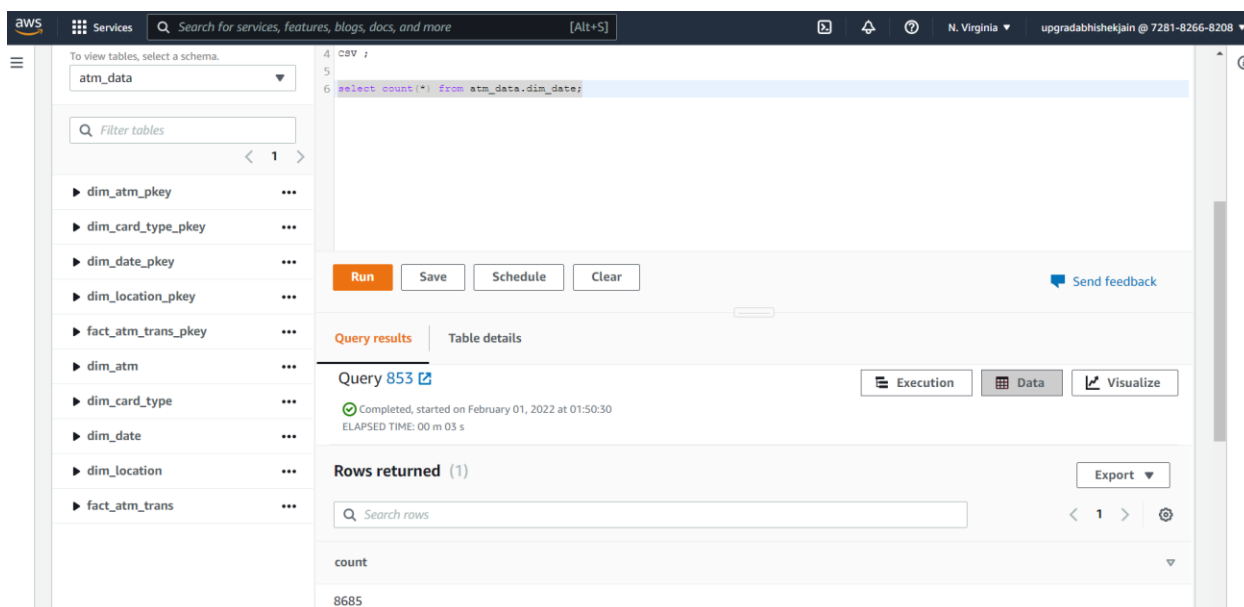
```

1 copy atm_data.dim_date from 's3://etlprojectbyabhishek/dim_date/part-00000-b61b2653-82c7-41d7-b9c7-90b0cb2ba148-c000.csv'
2 iam_role 'arn:aws:iam::728182668208:role/upgrad_redshift_s3'
3 delimiter ',' timeformat 'auto' region 'us-east-1'
4 CSV ;

```

Query results: Query 822. Completed, started on February 01, 2022 at 01:47:30. ELAPSED TIME: 00 m 11 s.

**Count of record :8685**



Query 853

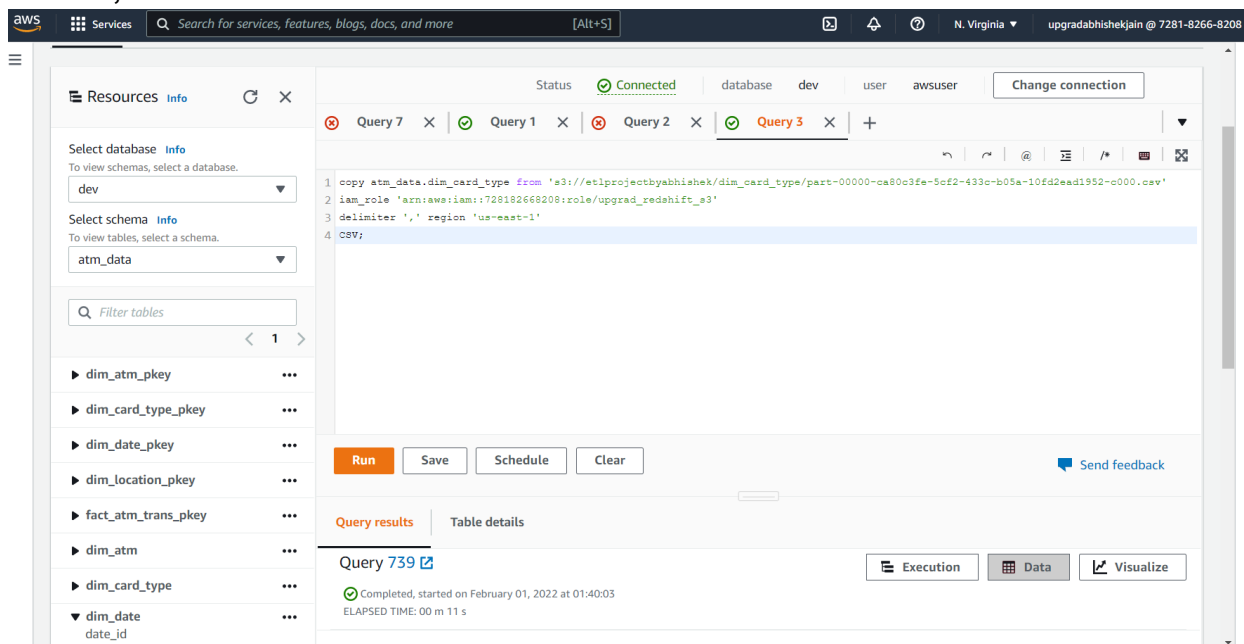
Completed, started on February 01, 2022 at 01:50:30  
ELAPSED TIME: 00 m 03 s

Rows returned (1)

Search rows

count
8685

- Copy data to DIM\_CARD\_TYPE table  
copy atm\_data.dim\_card\_type from 's3://etlprojectbyabhishek/dim\_card\_type/part-00000-ca80c3fe-5cf2-433c-b05a-10fd2ead1952-c000.csv'  
iam\_role 'arn:aws:iam::728182668208:role/upgrad\_redshift\_s3'  
delimiter ',' region 'us-east-1'  
CSV;



Query 739

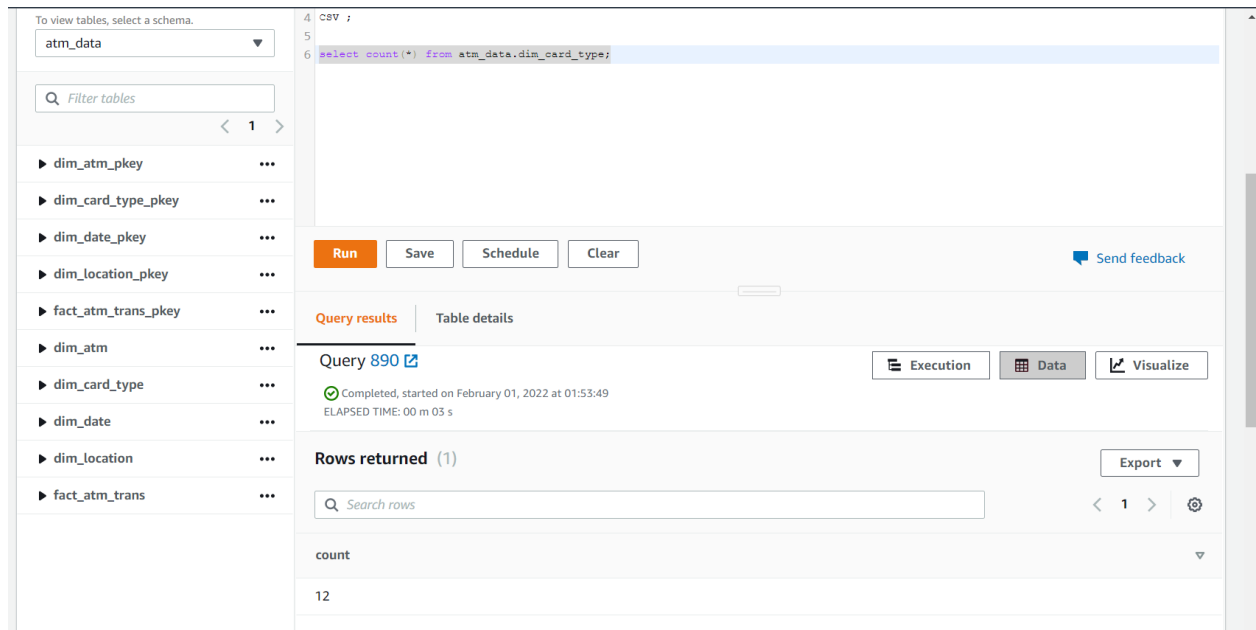
Completed, started on February 01, 2022 at 01:40:03  
ELAPSED TIME: 00 m 11 s

Rows returned (12)

Search rows

count
12

Count of record :12



To view tables, select a schema.

atm\_data

Filter tables

1

dim\_atm\_pkey ...

dim\_card\_type\_pkey ...

dim\_date\_pkey ...

dim\_location\_pkey ...

fact\_atm\_trans\_pkey ...

dim\_atm ...

dim\_card\_type ...

dim\_date ...

dim\_location ...

fact\_atm\_trans ...

Run Save Schedule Clear

Send feedback

Query results Table details

Query 890

Completed, started on February 01, 2022 at 01:53:49  
ELAPSED TIME: 00 m 03 s

Rows returned (1)

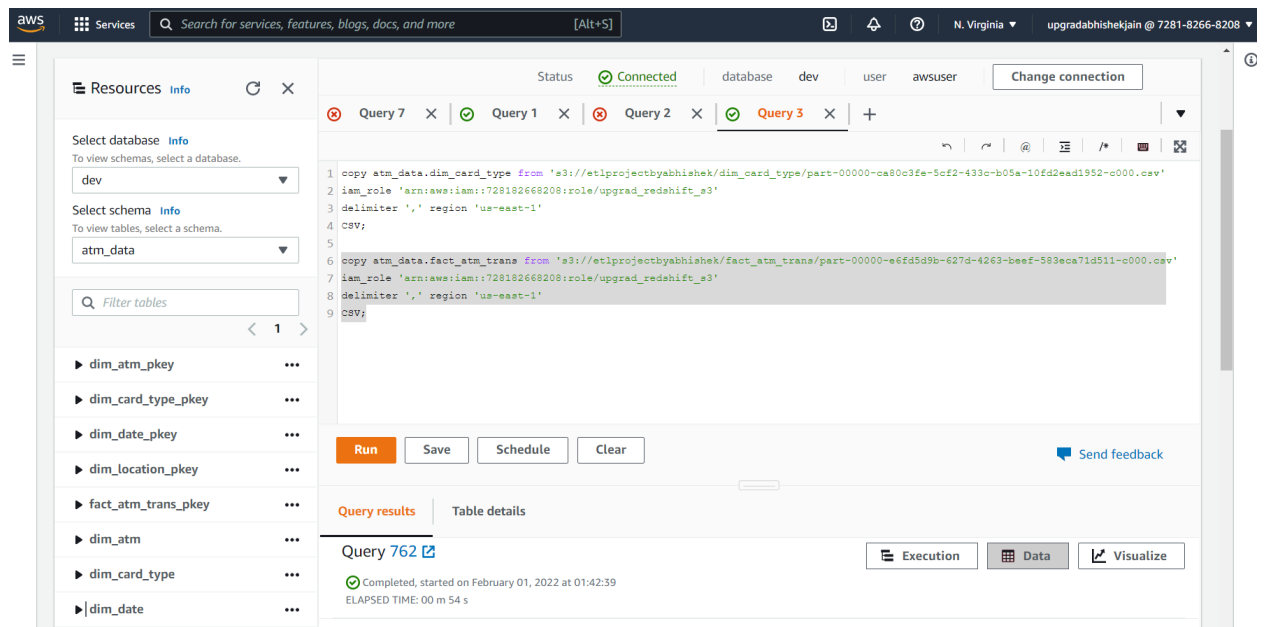
Search rows

1

count

12

- Copy data to FACT\_ATM\_TRANSACTION table  
copy atm\_data.fact\_atm\_trans from  
's3://etlprojectbyabhishek/fact\_atm\_trans/part-00000-e6fd5d9b-627d-4263-beef-583eca71d511-c000.csv'  
iam\_role 'arn:aws:iam::728182668208:role/upgrad\_redshift\_s3'  
delimiter ',' region 'us-east-1'  
CSV;



Resources Info

Select database dev

Select schema atm\_data

Filter tables

1

dim\_atm\_pkey ...

dim\_card\_type\_pkey ...

dim\_date\_pkey ...

dim\_location\_pkey ...

fact\_atm\_trans\_pkey ...

dim\_atm ...

dim\_card\_type ...

dim\_date ...

Status Connected database dev user awsuser Change connection

Query 7 Query 1 Query 2 Query 3

1 copy atm\_data.dim\_card\_type from 's3://etlprojectbyabhishek/dim\_card\_type/part-00000-e6fd5d9b-627d-4263-beef-583eca71d511-c000.csv'

2 iam\_role 'arn:aws:iam::728182668208:role/upgrad\_redshift\_s3'

3 delimiter ',' region 'us-east-1'

4 CSV;

5

6 copy atm\_data.fact\_atm\_trans from 's3://etlprojectbyabhishek/fact\_atm\_trans/part-00000-e6fd5d9b-627d-4263-beef-583eca71d511-c000.csv'

7 iam\_role 'arn:aws:iam::728182668208:role/upgrad\_redshift\_s3'

8 delimiter ',' region 'us-east-1'

9 CSV;

Run Save Schedule Clear

Send feedback

Query results Table details

Query 762

Completed, started on February 01, 2022 at 01:42:39  
ELAPSED TIME: 00 m 54 s

Count of record : 2468572

© Copyright. upGrad Education Pvt. Ltd. All rights reserved