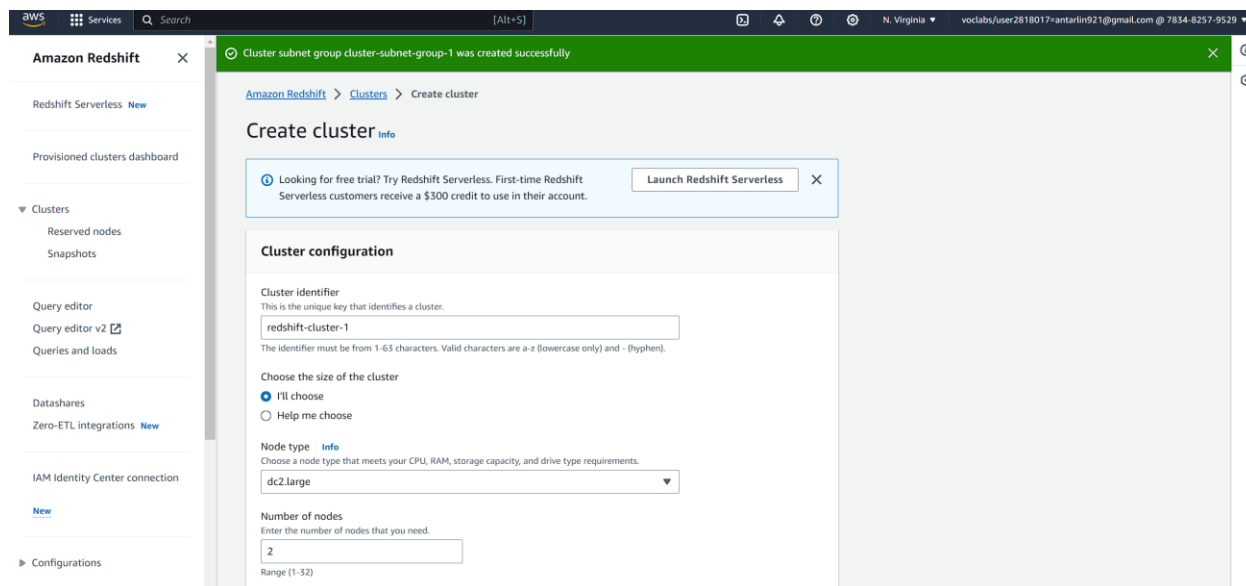


## Creation of a Redshift Cluster

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

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



Amazon Redshift

Cluster subnet group cluster-subnet-group-1 was created successfully

Amazon Redshift > Clusters > Create cluster

Create cluster [Info](#)

Looking for free trial? Try Redshift Serverless. First-time Redshift Serverless customers receive a \$300 credit to use in their account. [Launch Redshift Serverless](#)

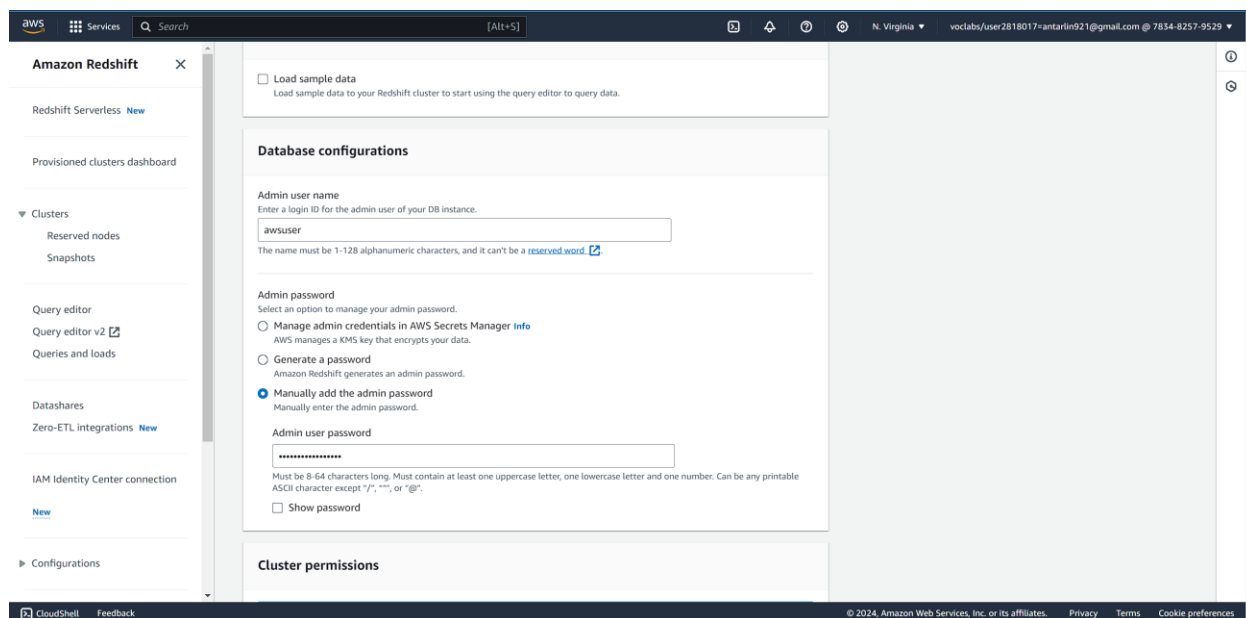
**Cluster configuration**

**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).

**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.

**Number of nodes**  
Enter the number of nodes that you need.  
  
Range (1-32)



Amazon Redshift

☐ Load sample data  
Load sample data to your Redshift cluster to start using the query editor to query data.

**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](#).

**Admin password**  
Select an option to manage your admin password.  
☐ Manage admin credentials in AWS Secrets Manager [Info](#)  
AWS manages a KMS key that encrypts your data.  
☐ Generate a password  
Amazon Redshift generates an admin password.  
☒ Manually add the admin password  
Manually enter the admin 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**

Amazon Redshift

Redshift Serverless [New](#)

Provisioned clusters dashboard

Clusters

- Reserved nodes
- Snapshots

Query editor

Query editor v2 [v2](#)

Queries and loads

Datashares

Zero-ETL integrations [New](#)

IAM Identity Center connection

[New](#)

Configurations

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](#)

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

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 and security [Info](#)

Amazon Redshift

Redshift Serverless [New](#)

Provisioned clusters dashboard

Clusters

- Reserved nodes
- Snapshots

Query editor

Query editor v2 [v2](#)

Queries and loads

Datashares

Zero-ETL integrations [New](#)

IAM Identity Center connection

[New](#)

Configurations

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 and security [Info](#)

Virtual private cloud (VPC)

This VPC defines the virtual networking environment for this cluster.

Default VPC

vpc-04c536d213598b3fa

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 group](#)

Choose one or more security groups

default

sg-023610277daa012f6

Cluster subnet group [Info](#)

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

cluster-subnet-group-1

Availability Zone

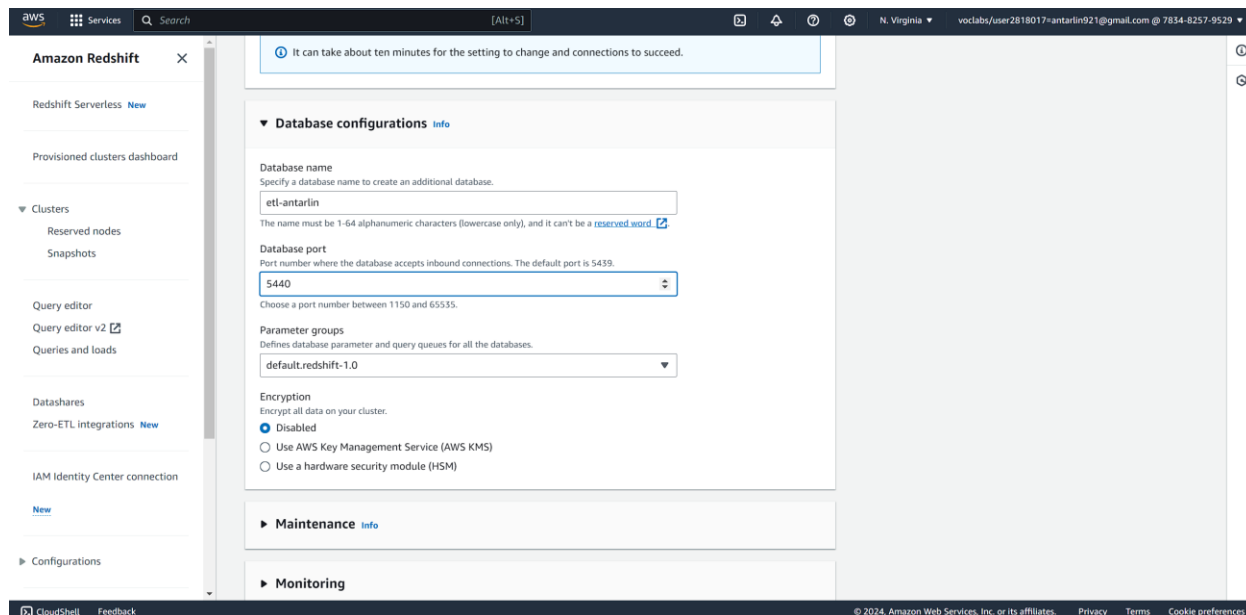
Specify the Availability Zone to create the cluster in. Otherwise, Amazon Redshift chooses an Availability Zone for you.

No preference

Enhanced VPC routing

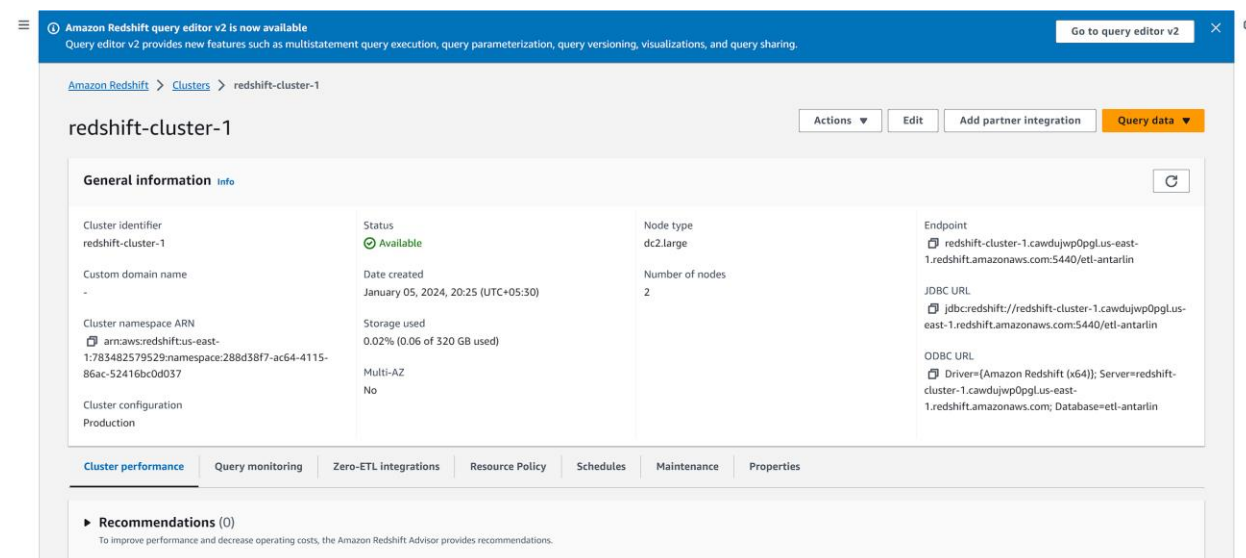
Enabling this option routes network traffic between your cluster and data repositories through a VPC, instead of through the internet. [Learn more about getting started cluster in vpc](#)

Turn off



The screenshot shows the Amazon Redshift console interface. On the left, there is a navigation menu with options like 'Redshift Serverless', 'Provisioned clusters dashboard', 'Clusters', 'Reserved nodes', 'Snapshots', 'Query editor', 'Query editor v2', 'Queries and loads', 'Datashares', 'Zero-ETL integrations', 'IAM Identity Center connection', and 'Configurations'. The main content area displays the 'Database configurations' for a cluster named 'redshift-cluster-1'. The configurations include:
 

- Database name:** etl-antarlin
- Database port:** 5440
- Parameter groups:** default.redshift-1.0
- Encryption:** Disabled
- Maintenance:** (link to maintenance info)
- Monitoring:** (link to monitoring info)

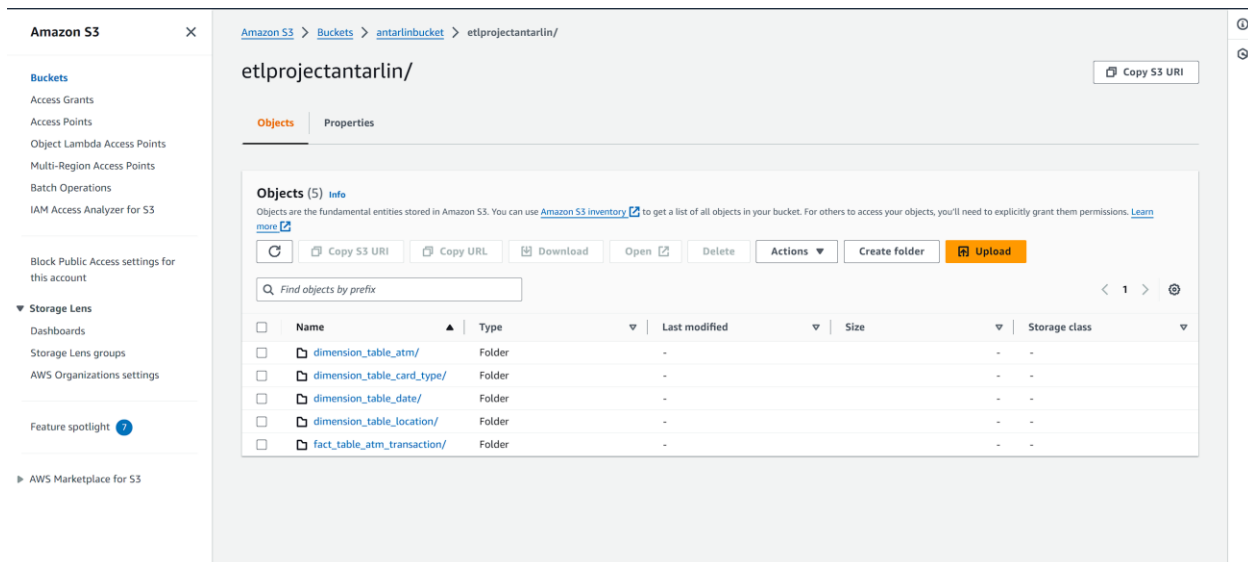


The screenshot shows the 'General information' tab for the 'redshift-cluster-1' cluster. The cluster is in an 'Available' state. The configuration details are as follows:
 

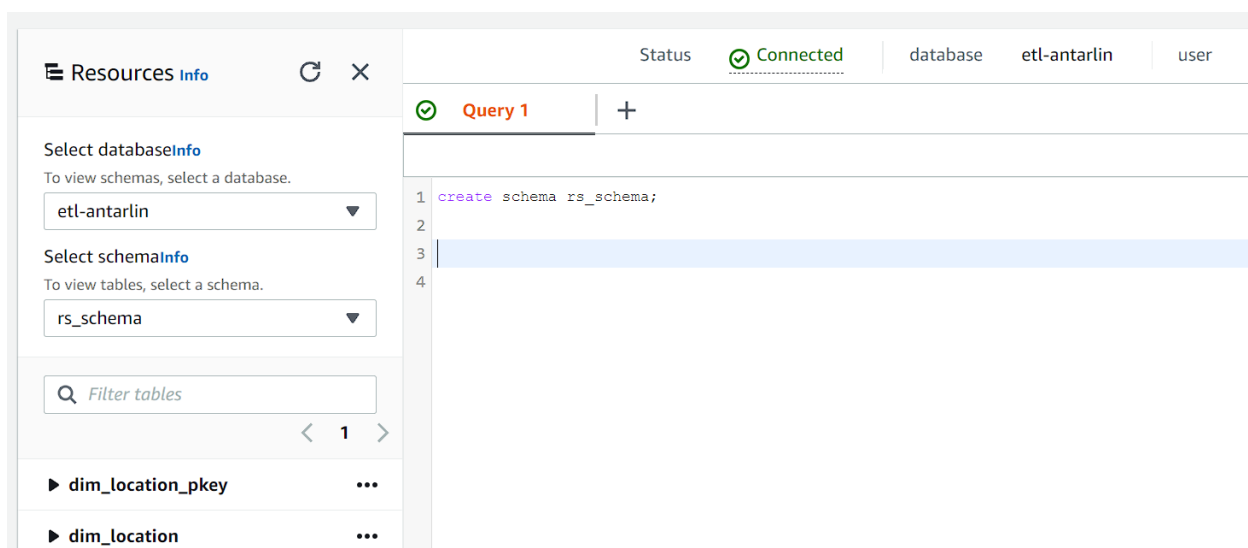
Property	Value
Cluster identifier	redshift-cluster-1
Status	Available
Node type	dc2.large
Endpoint	redshift-cluster-1.cawdujwp0pgl-us-east-1.redshift.amazonaws.com:5440/etl-antarlin
Custom domain name	-
Date created	January 05, 2024, 20:25 (UTC+05:30)
Number of nodes	2
Cluster namespace ARN	arn:aws:redshift:us-east-1:783482579529:namespace:288d38f7-ac64-4115-86ac-52416bc0d037
Storage used	0.02% (0.06 of 320 GB used)
Multi-AZ	No
Cluster configuration	Production
JDBC URL	jdbc:redshift://redshift-cluster-1.cawdujwp0pgl-us-east-1.redshift.amazonaws.com:5440/etl-antarlin
ODBC URL	Driver={Amazon Redshift (x64)}; Server=redshift-cluster-1.cawdujwp0pgl-us-east-1.redshift.amazonaws.com; Database=etl-antarlin

We can see redshift configurations above

Lets check the S3 folder with the data we imported from HDFS on EMR



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

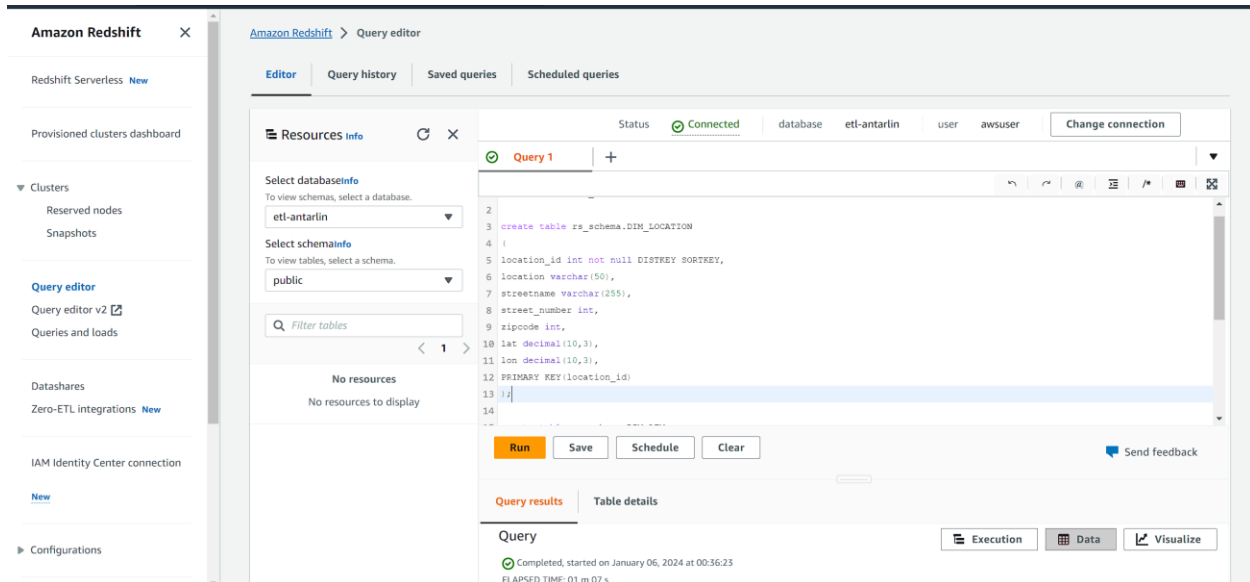


Create schema rs\_schema;

- **Creating location dimension table**

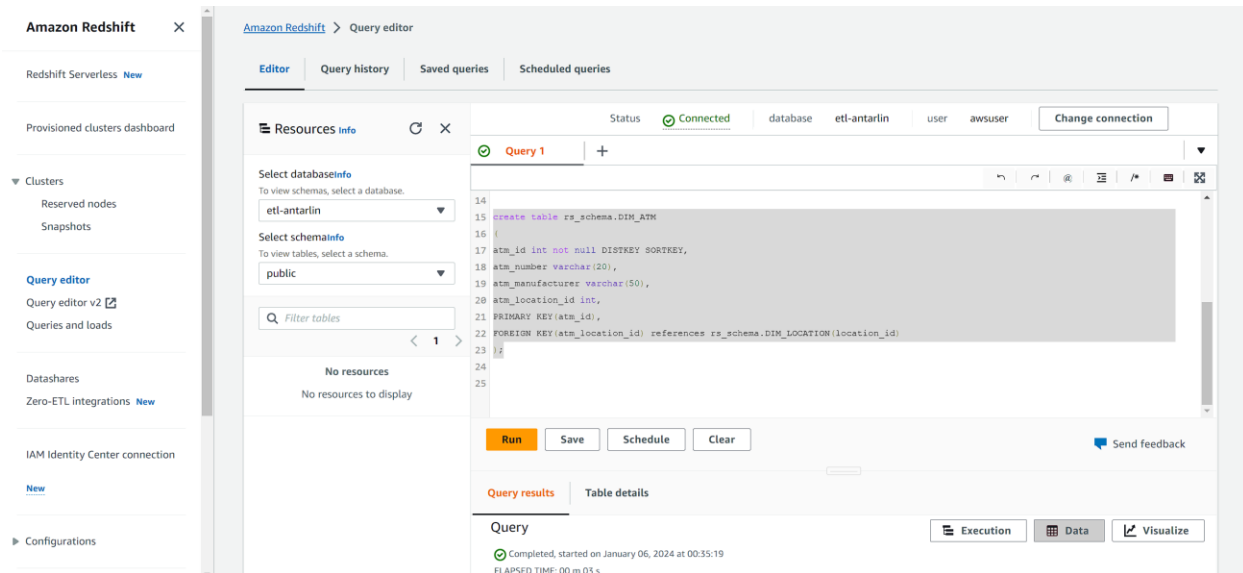
```
create table rs_schema.DIM_LOCATION
(
location_id int not null DISTKEY SORTKEY,
```

```
location varchar(50),
streetname varchar(255),
street_number int,
zipcode int,
lat decimal(10,3),
lon decimal(10,3),
PRIMARY KEY(location_id)
);
```



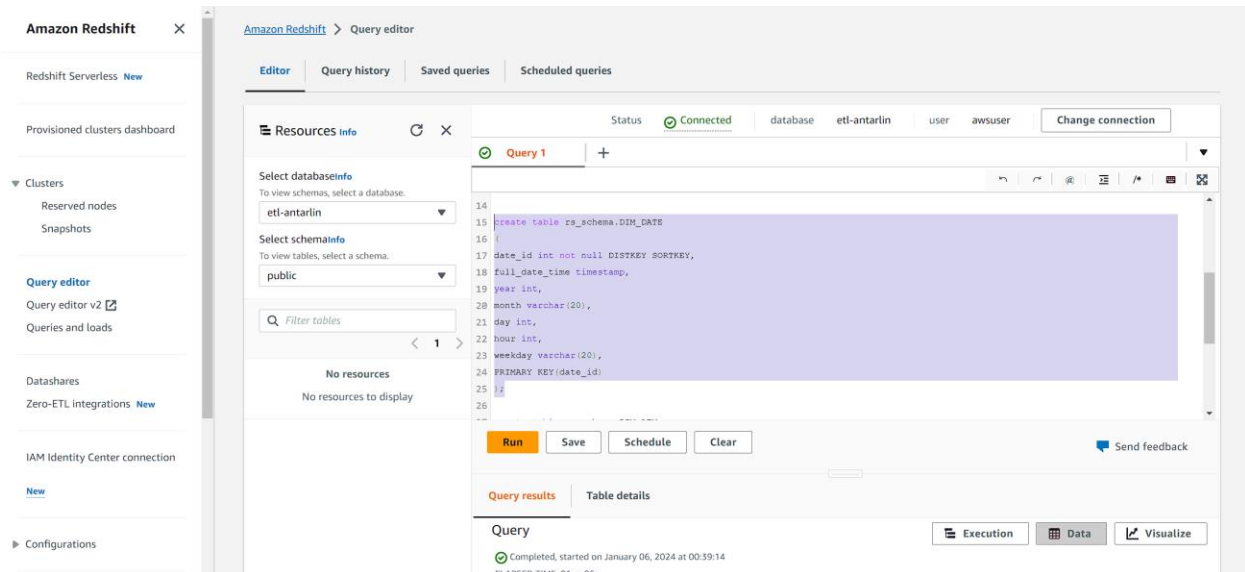
- **Creating atm dimension table**

```
create table rs_schema.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 rs_schema.DIM_LOCATION(location_id)
);
```



- **Creating date dimension table**

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



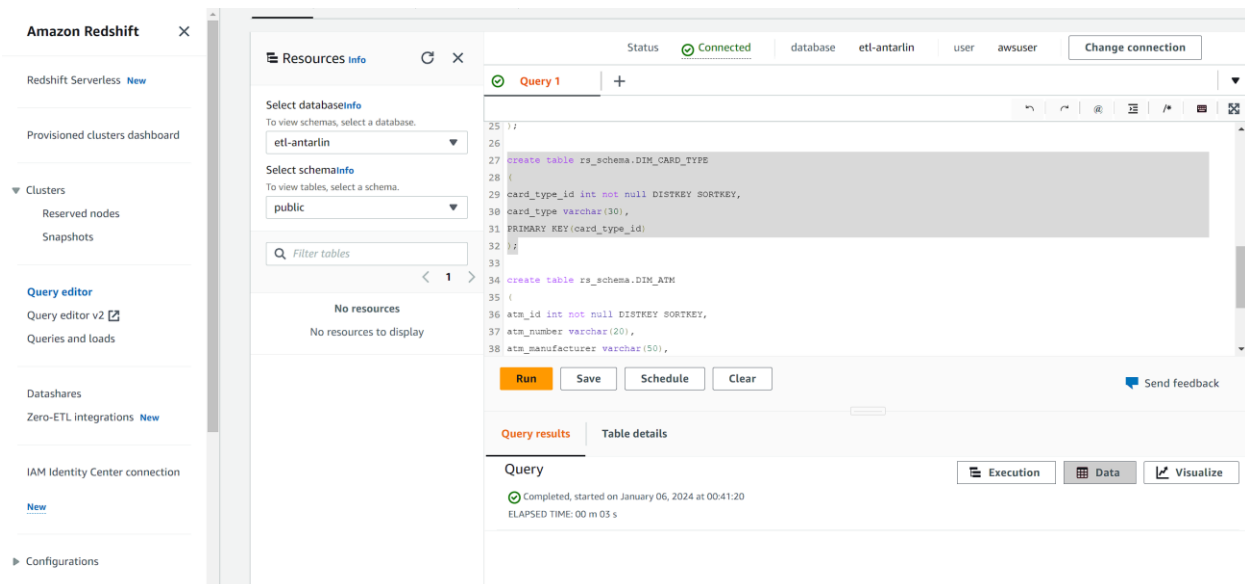
The screenshot shows the Amazon Redshift Query Editor interface. On the left, the navigation pane includes sections for Redshift Serverless, Provisioned clusters dashboard, Clusters (Reserved nodes, Snapshots), Query editor (Query editor v2, Queries and loads), Datashares, Zero-ETL integrations, IAM Identity Center connection, and Configurations. The main editor area is titled 'Query editor' and shows a SQL query being executed. The query is as follows:

```
14
15 create table rs_schema.DIM_DATE
16
17 date_id int not null DISTKEY SORTKEY,
18 full_date_time timestamp,
19 year int,
20 month varchar(20),
21 day int,
22 hour int,
23 weekday varchar(20),
24 PRIMARY KEY(date_id)
25
26
```

The query has been executed successfully, as indicated by the 'Query results' section showing 'Query Completed, started on January 06, 2024 at 00:39:14'. The 'Table details' section is also visible.

- **Creating card type dimension table**

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



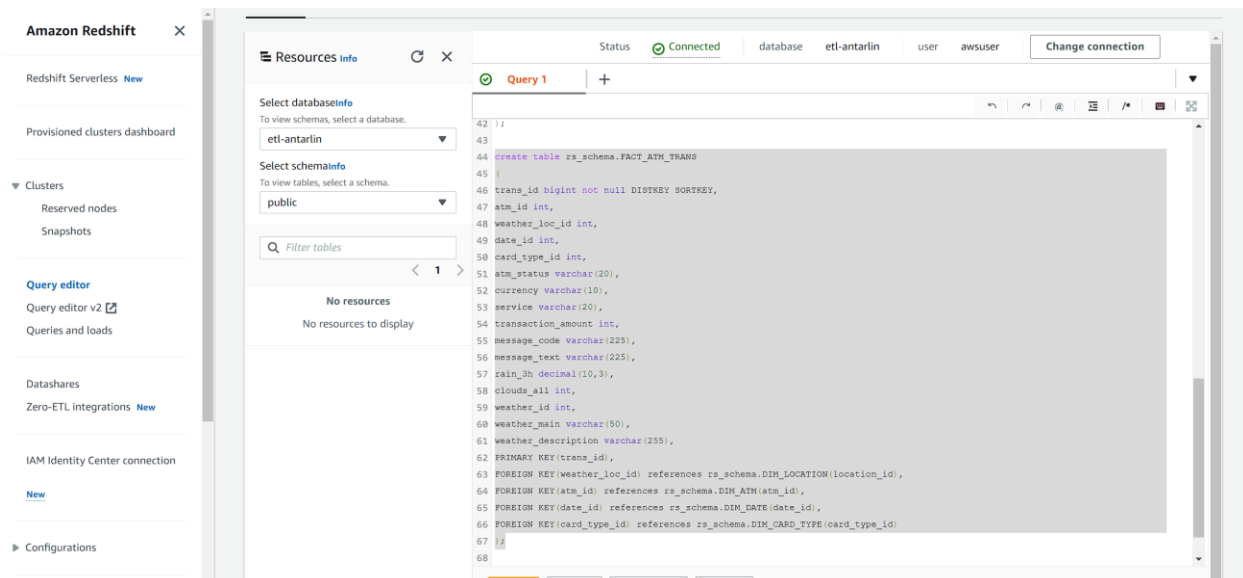
The screenshot shows the Amazon Redshift Query Editor interface with a new SQL query being executed. The query is as follows:

```
25 //
26
27 create table rs_schema.DIM_CARD_TYPE
28 (
29 card_type_id int not null DISTKEY SORTKEY,
30 card_type varchar(30),
31 PRIMARY KEY(card_type_id)
32 )
33
34 create table rs_schema.DIM_ATM
35 (
36 atm_id int not null DISTKEY SORTKEY,
37 atm_number varchar(20),
38 atm_manufacturer varchar(50),
39
```

The query has been executed successfully, as indicated by the 'Query results' section showing 'Query Completed, started on January 06, 2024 at 00:41:20'. The 'Table details' section is also visible.

- **Creating atm transactions fact table**

```
create table rs_schema.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(255),
PRIMARY KEY(trans_id),
FOREIGN KEY(weather_loc_id) references rs_schema.DIM_LOCATION(location_id),
FOREIGN KEY(atm_id) references rs_schema.DIM_ATM(atm_id),
FOREIGN KEY(date_id) references rs_schema.DIM_DATE(date_id),
FOREIGN KEY(card_type_id) references rs_schema.DIM_CARD_TYPE(card_type_id)
);
```

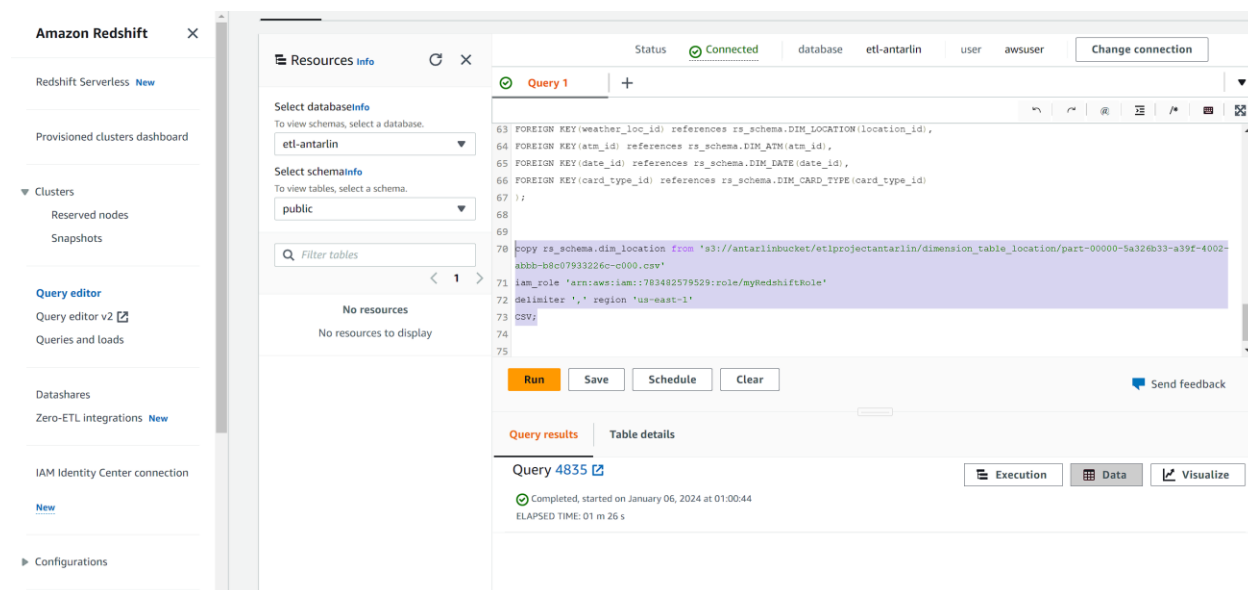




## Queries to copy the data from S3 buckets to the Redshift cluster in the appropriate tables

- Copying the data to dim\_location table

copy rs\_schema.dim\_location from  
's3://antarlinbucket/etlprojectantarlin/dimension\_table\_location/part-00000-5a326b33-a39f-4002-abbb-b8c07933226c-c000.csv'  
iam\_role 'arn:aws:iam::783482579529:role/myRedshiftRole'  
delimiter ',' region 'us-east-1'  
CSV;



The screenshot shows the Amazon Redshift Query Editor interface. The left sidebar contains navigation options like 'Redshift Serverless', 'Provisioned clusters dashboard', 'Clusters', 'Query editor v2', and 'IAM Identity Center connection'. The main area is titled 'Query 1' and shows a SQL query for copying data from an S3 bucket to the 'dim\_location' table. The query includes foreign key references and IAM role information. The 'Run' button is highlighted.

```

63 FOREIGN KEY (weather_loc_id) references rs_schema.DIM_LOCATION(location_id),
64 FOREIGN KEY (atm_id) references rs_schema.DIM_ATM(atm_id),
65 FOREIGN KEY (date_id) references rs_schema.DIM_DATE(date_id),
66 FOREIGN KEY (card_type_id) references rs_schema.DIM_CARD_TYPE(card_type_id)
67 ;
68
69
70 copy rs_schema.dim_location from 's3://antarlinbucket/etlprojectantarlin/dimension_table_location/part-00000-5a326b33-a39f-4002-
71 abbb-b8c07933226c-c000.csv'
72 iam_role 'arn:aws:iam::783482579529:role/myRedshiftRole'
73 delimiter ',' region 'us-east-1'
74 CSV;
75

```

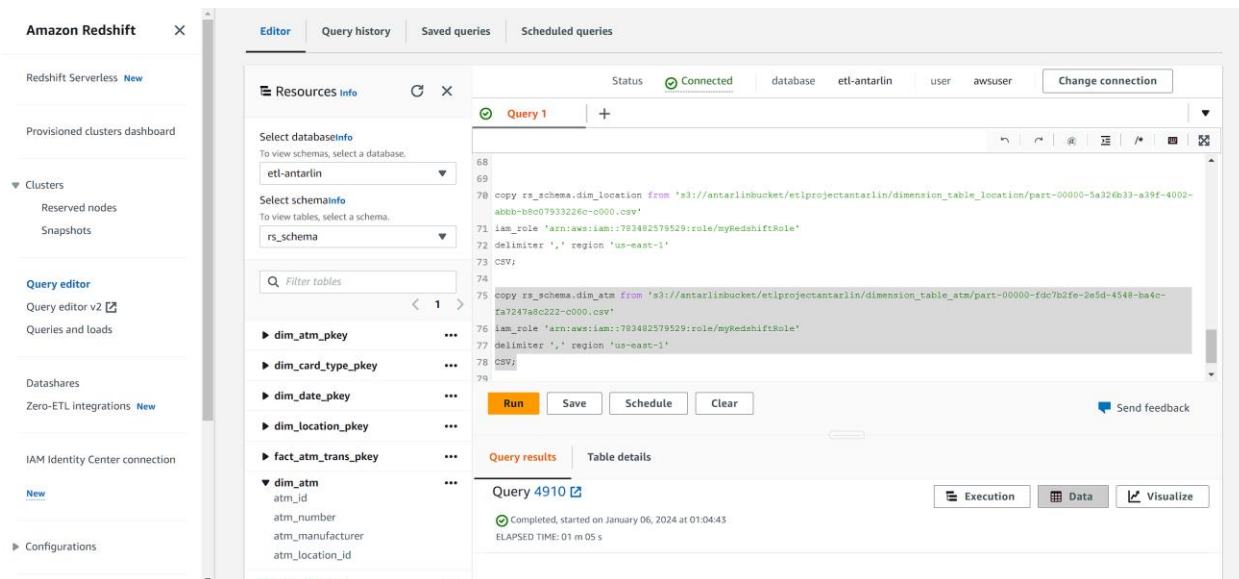
Query results: Table details

Query 4835

Completed, started on January 06, 2024 at 01:00:44  
ELAPSED TIME: 01 m 26 s

- Copying the data to dim\_atm table

copy rs\_schema.dim\_atm from 's3://antarlinbucket/etlprojectantarlin/dimension\_table\_atm/part-00000-fdc7b2fe-2e5d-4548-ba4c-fa7247a8c222-c000.csv'  
iam\_role 'arn:aws:iam::783482579529:role/myRedshiftRole'  
delimiter ',' region 'us-east-1'  
CSV;



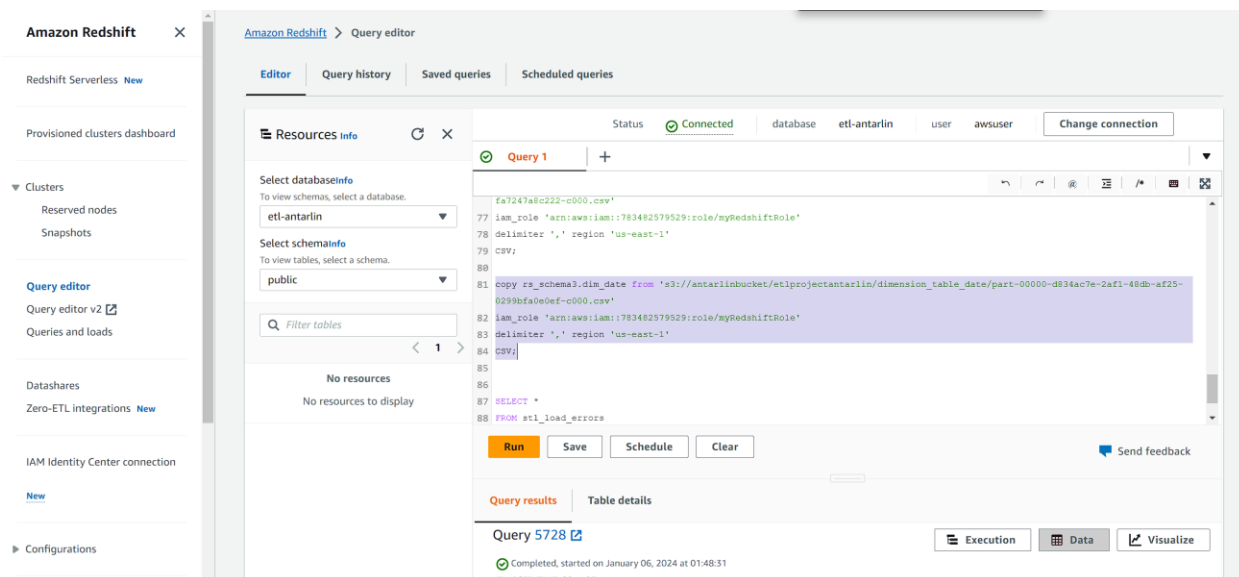
The screenshot shows the Amazon Redshift Query Editor interface. On the left, the 'Resources info' panel is open, showing the selected database 'etl-antarlin' and schema 'rs\_schema'. The 'dim\_atm' table is highlighted in the table list. The main query editor shows a SQL query that copies data from an S3 bucket to the 'dim\_atm' table. The query is as follows:

```
68
69
70 copy rs_schema.dim_location from 's3://antarlinbucket/etlprojectantarlin/dimension_table_location/part-00000-5a326b33-a39f-4002-
71 abbb-b8c07933226c-c000.csv'
72 iam_role 'arn:aws:iam::783482579529:role/myRedshiftRole'
73 delimiter ',' region 'us-east-1'
74 CSV;
75
76 copy rs_schema.dim_atm from 's3://antarlinbucket/etlprojectantarlin/dimension_table_atm/part-00000-fdc7b2fe-2e5d-4548-ba4c-
77 fa7247a8c222-c000.csv'
78 iam_role 'arn:aws:iam::783482579529:role/myRedshiftRole'
79 delimiter ',' region 'us-east-1'
80 CSV;
```

The query has been executed successfully, as indicated by the 'Query results' section showing 'Query 4910' completed on January 06, 2024 at 01:04:43, with an elapsed time of 01 m 05 s.

- Copying the data to dim\_date table

copy rs\_schema.dim\_date from  
's3://antarlinbucket/etlprojectantarlin/dimension\_table\_date/part-00000-d834ac7e-2af1-48db-af25-0299bfa0e0ef-c000.csv'  
iam\_role 'arn:aws:iam::783482579529:role/myRedshiftRole'  
delimiter ',' region 'us-east-1'  
CSV;



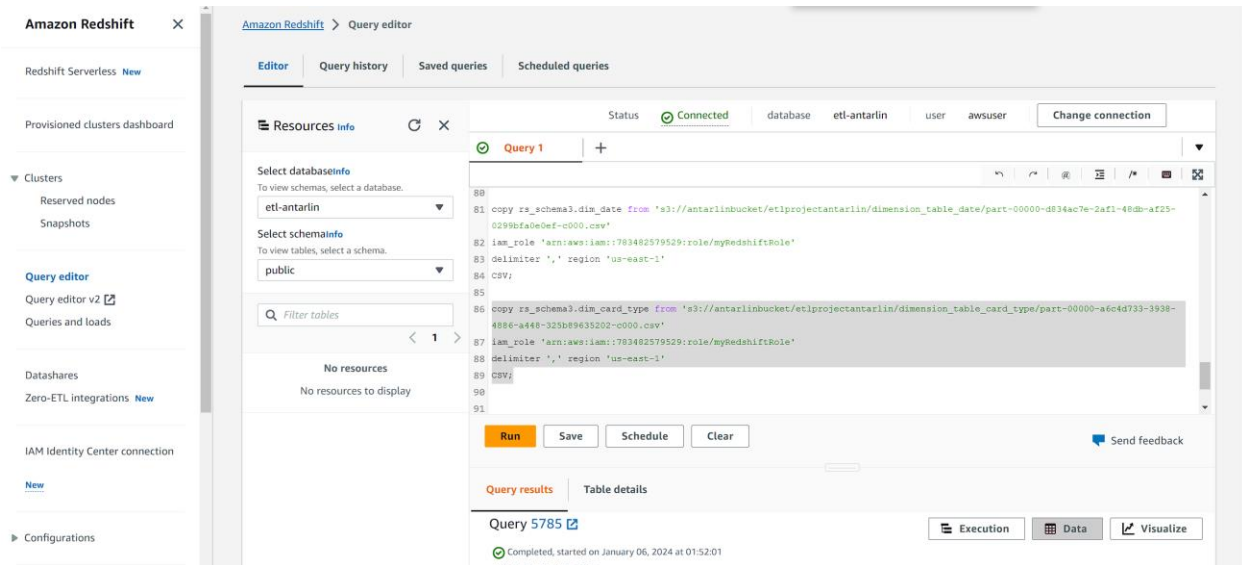
The screenshot shows the Amazon Redshift Query Editor interface. On the left, the 'Resources info' panel is open, showing the selected database 'etl-antarlin' and schema 'public'. The 'dim\_date' table is highlighted in the table list. The main query editor shows a SQL query that copies data from an S3 bucket to the 'dim\_date' table. The query is as follows:

```
81 copy rs_schema3.dim_date from 's3://antarlinbucket/etlprojectantarlin/dimension_table_date/part-00000-d834ac7e-2af1-48db-af25-
82 0299bfa0e0ef-c000.csv'
83 iam_role 'arn:aws:iam::783482579529:role/myRedshiftRole'
84 delimiter ',' region 'us-east-1'
85 CSV;
86
87 SELECT *
88 FROM at1_load_errors
```

The query has been executed successfully, as indicated by the 'Query results' section showing 'Query 5728' completed on January 06, 2024 at 01:48:31, with an elapsed time of 00 m 03 s.

- Copying the data to dim\_card\_type table

```
copy rs_schema.dim_card_type from
's3://antarlinbucket/etlprojectantarlin/dimension_table_card_type/part-00000-a6c4d733-3938-
4886-a448-325b89635202-c000.csv'
iam_role 'arn:aws:iam::783482579529:role/myRedshiftRole'
delimiter ',' region 'us-east-1'
CSV;
```



- **Copying the data to fact\_atm\_trans table**

```
copy rs_schema.fact_atm_trans from
's3://antarlinbucket/etlprojectantarlin/fact_table_atm_transaction/part-00000-ee76c090-09d9-
40a3-ab9b-4b2daad3787e-c000.csv'
iam_role 'arn:aws:iam::783482579529:role/myRedshiftRole'
delimiter ',' region 'us-east-1'
CSV;
```

Amazon Redshift

Redshift Serverless

Provisioned clusters dashboard

Clusters

Reserved nodes

Snapshots

Query editor

Query editor v2

Queries and loads

Datashares

Zero-ETL integrations

IAM Identity Center connection

Configurations

Resources info

Select database info

eti-antarlin

Select schema info

public

Filter tables

No resources

No resources to display

Status Connected

database eti-antarlin

user user

awsuser

Change connection

Query 1

Run

Save

Schedule

Clear

Send feedback

Query results

Table details

Query 5825

Completed, started on January 06, 2024 at 01:54:48

ELAPSED TIME: 00 m 30 s

Execution

Data

Visualize

The End