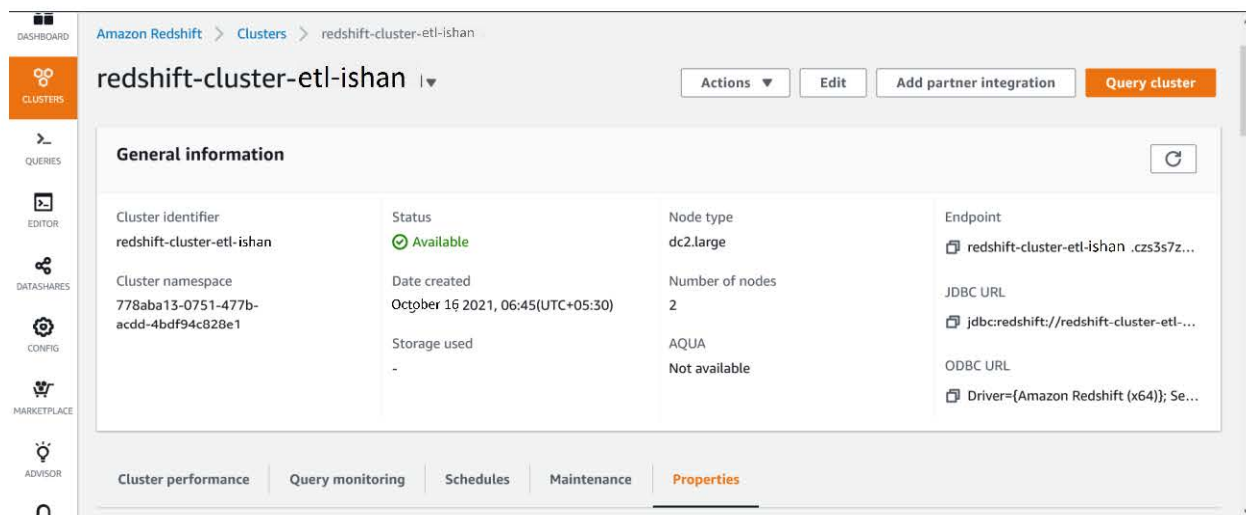


Creation of a RedShift Cluster

Screenshots of the configuration of the RedShift cluster that I have created:

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

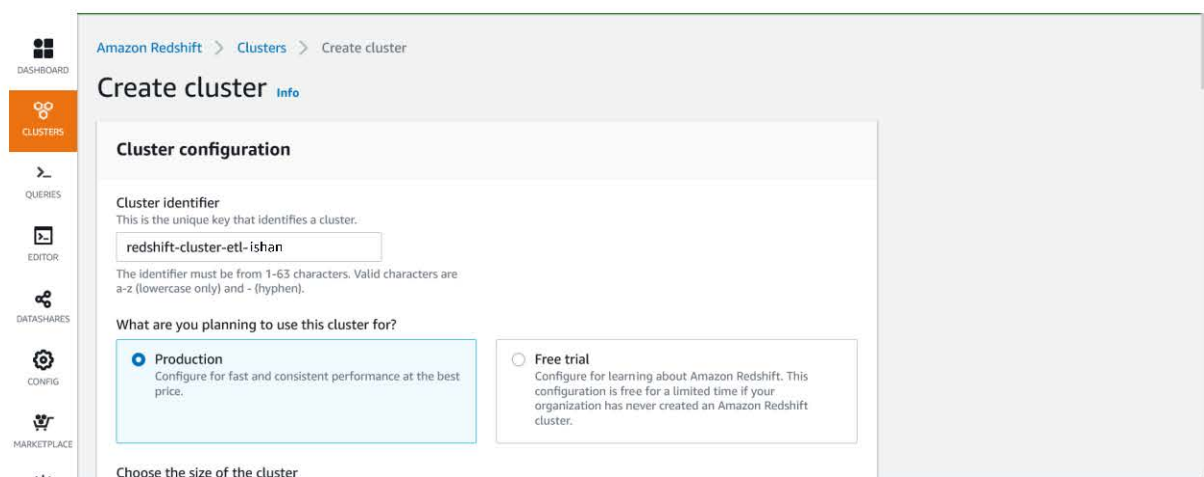


The screenshot shows the Amazon RedShift console interface. The breadcrumb navigation is 'Amazon Redshift > Clusters > redshift-cluster-etl-ishan'. The cluster name 'redshift-cluster-etl-ishan' is displayed with a dropdown arrow. Action buttons include 'Actions', 'Edit', 'Add partner integration', and 'Query cluster'. The 'General information' tab is active, showing the following details:

General information			
Cluster identifier	Status	Node type	Endpoint
redshift-cluster-etl-ishan	Available	dc2.large	redshift-cluster-etl-ishan.czs3s7z...
Cluster namespace	Date created	Number of nodes	JDBC URL
778aba13-0751-477b-acdd-4bdf94c828e1	October 16 2021, 06:45(UTC+05:30)	2	jdbc:redshift://redshift-cluster-etl-...
	Storage used	AQUA	ODBC URL
	-	Not available	Driver=(Amazon Redshift (x64)); Se...

At the bottom, there are tabs for 'Cluster performance', 'Query monitoring', 'Schedules', 'Maintenance', and 'Properties' (which is currently selected).

Screenshots of the various configurations associated with cluster creation:



The screenshot shows the 'Create cluster' page in the Amazon RedShift console. The breadcrumb navigation is 'Amazon Redshift > Clusters > Create cluster'. The page title is 'Create cluster' with an 'Info' link. The 'Cluster configuration' section includes:

- Cluster identifier:** A text box containing 'redshift-cluster-etl-ishan'. Below it, a note states: '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?** Two radio button options:
 - Production** (selected): '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.'

Below these options is a section titled 'Choose the size of the cluster'.

ADVISOR

ALARMS

EVENTS

WHAT'S NEW

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

Nodes

Enter the number of nodes that you need.

2

Range (1-32)

Configuration summary

Info

dc2.large | 2 nodes

\$360.00/month

Estimated on-demand compute price

Save more than 60% of your costs

320 GB

Total compressed storage

The total storage capacity for the cluster if you deploy the number

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 "@".

► **Cluster permissions**

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**

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

vpc-016d4002e55745e04

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

cloudera ✕
sg-0b0c2604316b85322

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

Availability Zone

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

us-east-1f

Enhanced VPC routing

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

☒ Disabled
 ☐ Enabled

Publicly accessible

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.

etl-ishah

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

▼ Database configurations

Database name

Specify a database name to create an additional database.

etl-ishah

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.

55555

The port must be numeric (1150-65535).

Parameter groups

Defines database parameter and query queues for all the databases.

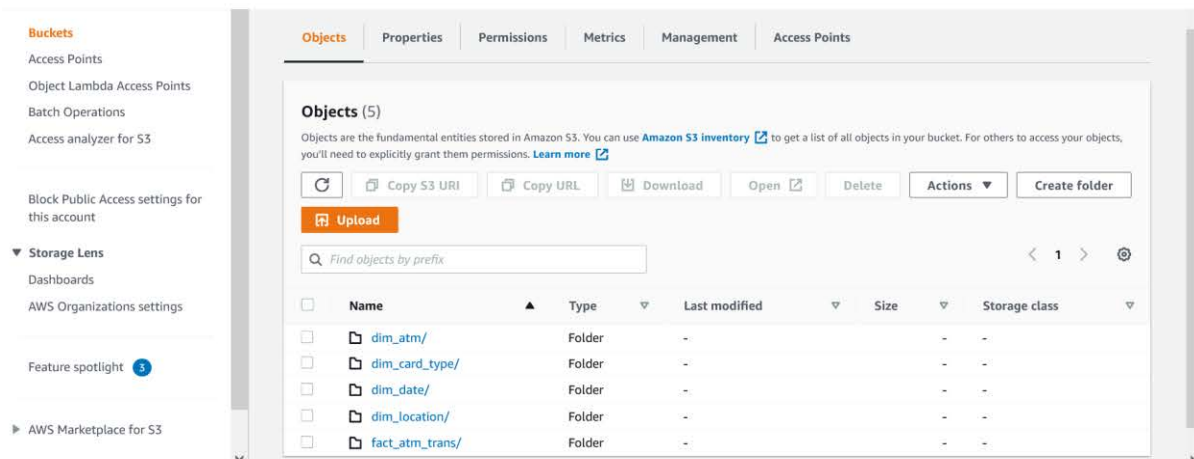
Encryption

Encrypt all data on your cluster.

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

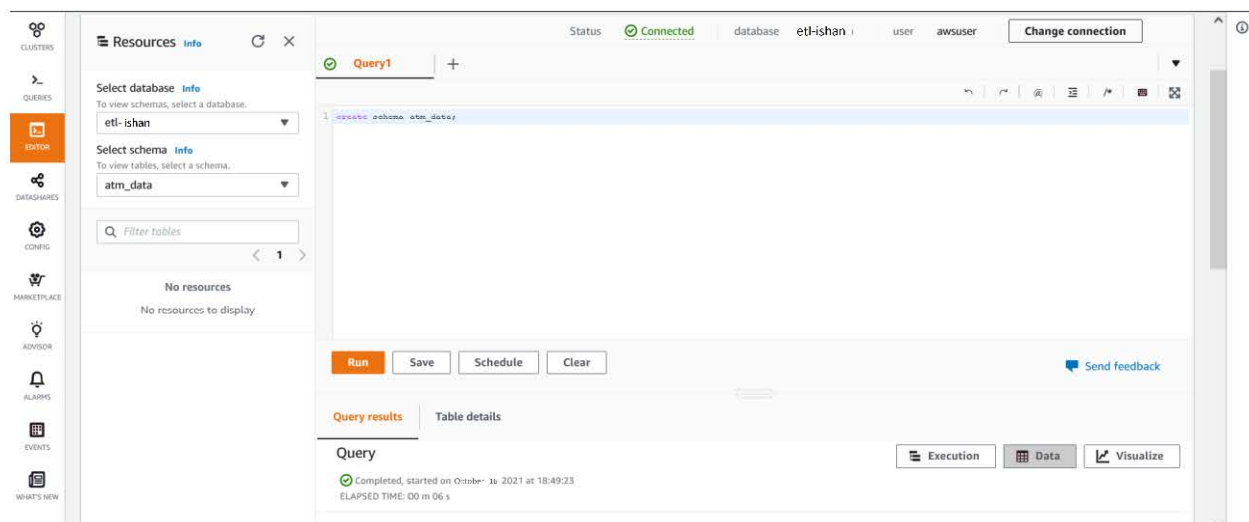
Setting up a database in the RedShift cluster and running queries to create the dimension and fact tables

Viewing all the data in Amazon S3 bucket:



Query to create a schema for the dimension and fact tables:

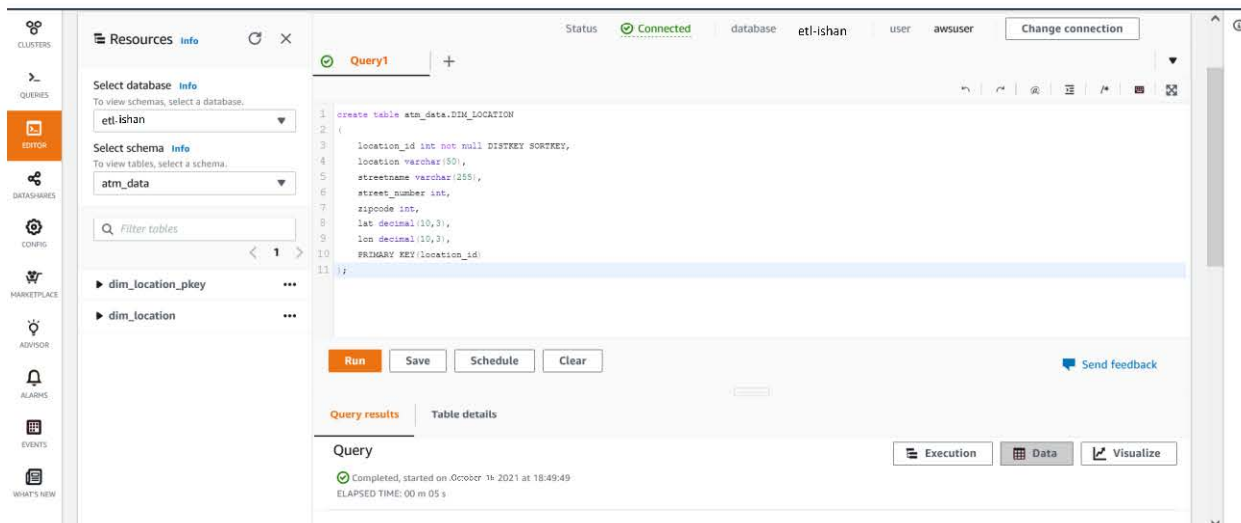
create schema atm_data;



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

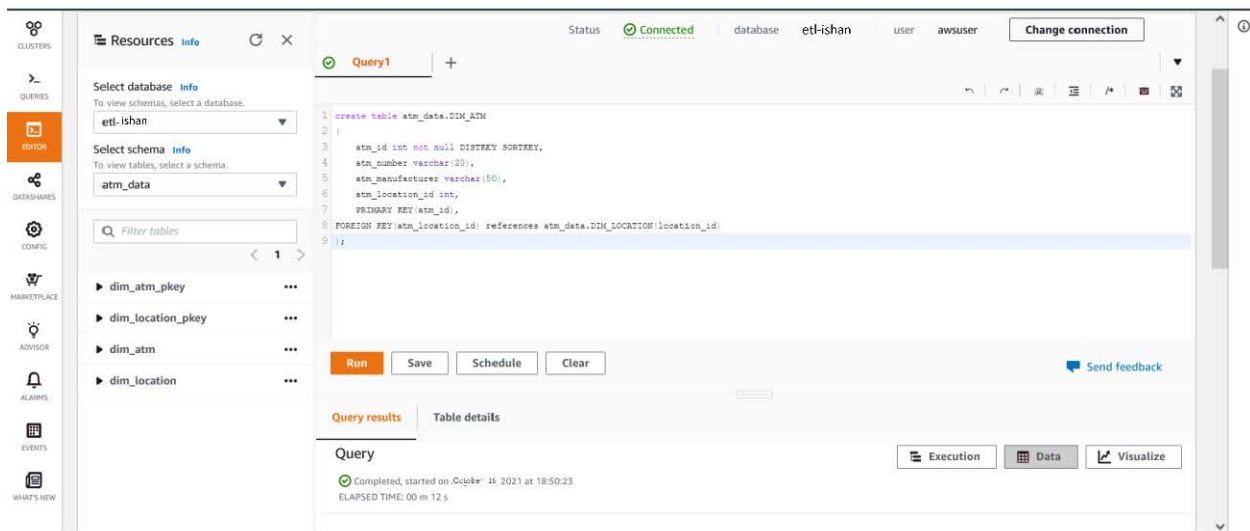
- **Creating location dimension table**

```
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)
);
```



- **Creating atm dimension 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)
);
```



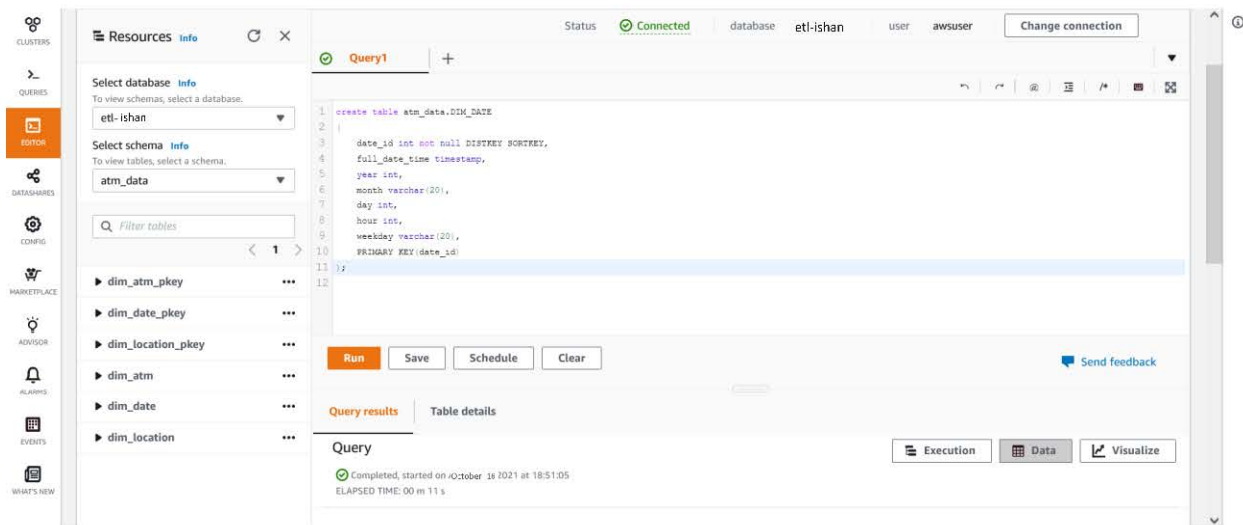
- **Creating date dimension table**

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)

);



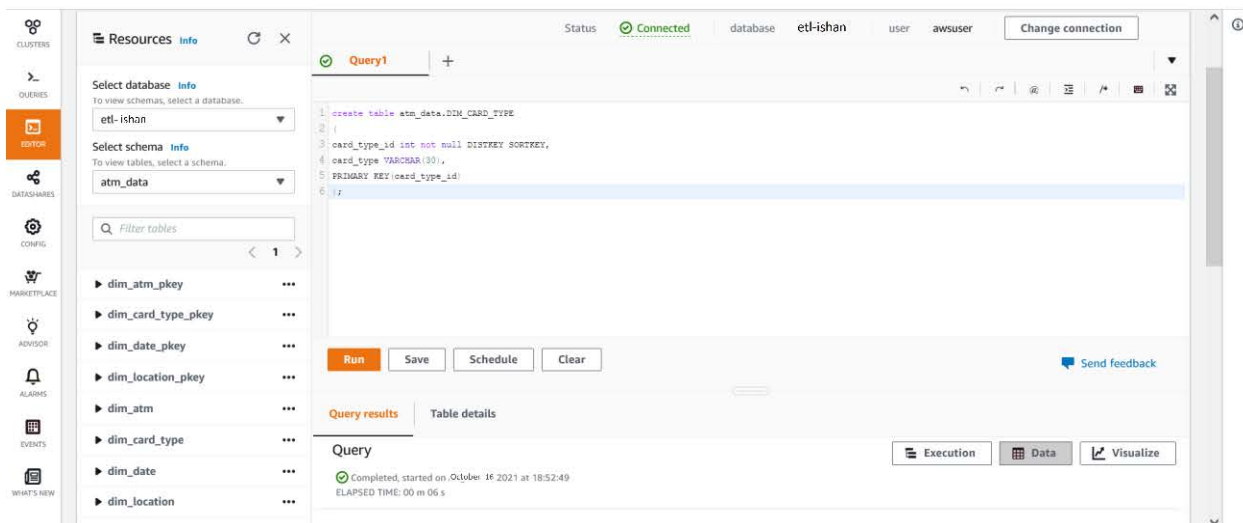
The screenshot shows the AWS Glue console interface. On the left, the 'Resources' sidebar lists various components like Clusters, Queries, Editor, Databases, Config, Marketplace, Advisor, Alarms, and Events. The main panel displays the 'Query1' editor. The 'Select database' dropdown is set to 'etl-ishan' and the 'Select schema' dropdown is set to 'atm_data'. The query editor contains the following SQL code:

```
1 create table atm_data.DIM_DATE
2
3   date_id int not null DISTKEY SORTKEY,
4   full_date_time timestamp,
5   year int,
6   month varchar(20),
7   day int,
8   hour int,
9   weekday varchar(20),
10  PRIMARY KEY(date_id)
11 ;
12
```

Below the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' tab is active, showing a message: 'Query Completed, started on October 16 2021 at 18:51:05. ELAPSED TIME: 00 m 11 s.' There are also buttons for 'Execution', 'Data', and 'Visualize'.

- **Creating 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 AWS Glue console interface. On the left, the 'Resources' sidebar lists various components like Clusters, Queries, Editor, Databases, Config, Marketplace, Advisor, Alarms, and Events. The main panel displays the 'Query1' editor. The 'Select database' dropdown is set to 'etl-ishan' and the 'Select schema' dropdown is set to 'atm_data'. The query editor contains the following SQL code:

```
1 create table atm_data.DIM_CARD_TYPE
2
3   card_type_id int not null DISTKEY SORTKEY,
4   card_type varchar(30),
5   PRIMARY KEY(card_type_id)
6 ;
```

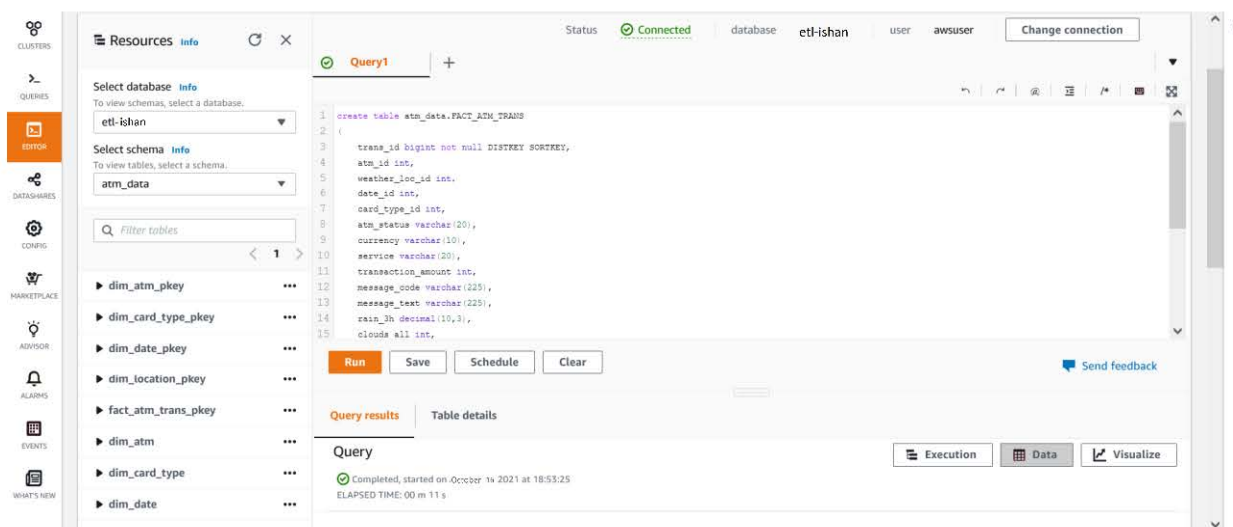
Below the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' tab is active, showing a message: 'Query Completed, started on October 16 2021 at 18:52:49. ELAPSED TIME: 00 m 06 s.' There are also buttons for 'Execution', 'Data', and 'Visualize'.

- **Creating atm transactions 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(255),
PRIMARY KEY(trans_id),
FOREIGN KEY(weather_loc_id) references atm_data.DIM_LOCATION(location_id),
FOREIGN KEY(atm_id) references atm_data.DIM_DATA(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)
);
```

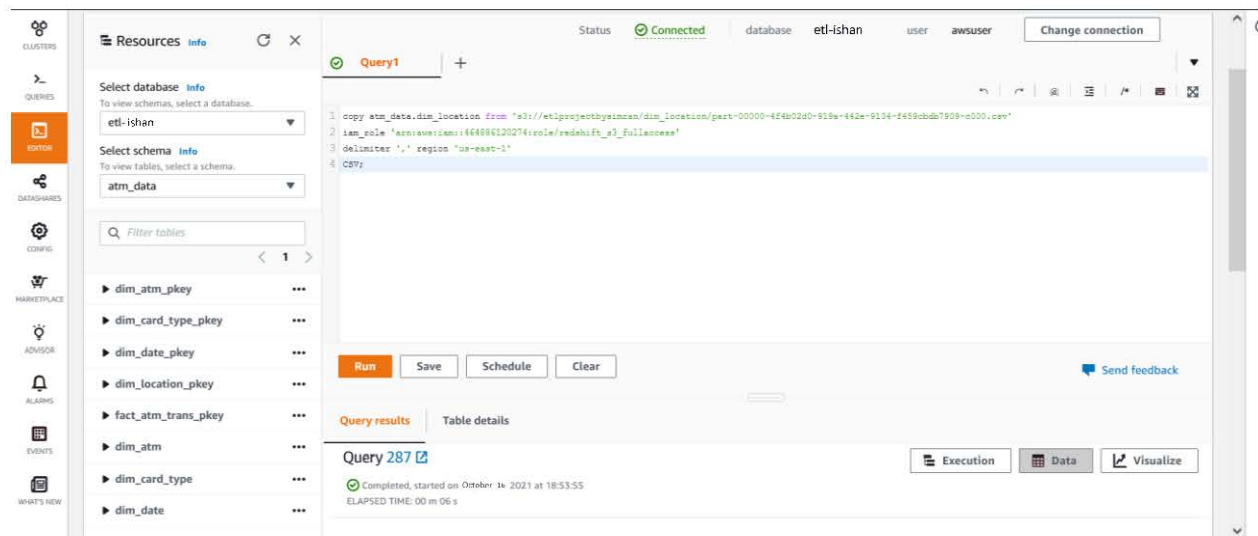


Loading data into a RedShift cluster from Amazon S3 bucket

Queries to copy the data from S3 bucket to the RedShift cluster in the appropriate tables:

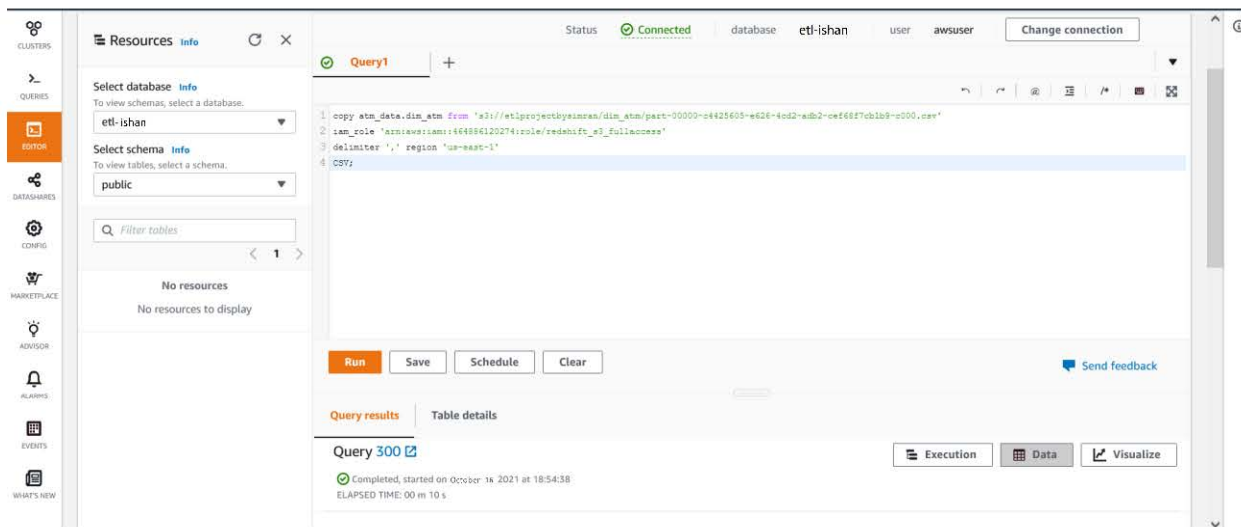
- Copying the data to dim_location table

```
copy atm_data.dim_location from 's3://etlprojectbyishan/dim_location/
part-00000-4f4b02d0-919a-442e-9134-f459cbdb7909-c000.csv'
iam_role 'arn:aws:iam::464886120274:role/redshift_s3_fullaccess'
delimiter ',' region 'us-east-1'
CSV;
```



- Copying the data to dim_atm table

```
copy atm_data.dim_atm from 's3://etlprojectbyishan/dim_atm/part-00000-c4425605-
e626-4cd2-adb2-cef68f7cb1b9-c000.csv'
iam_role 'arn:aws:iam::464886120274:role/redshift_s3_fullaccess'
delimiter ',' region 'us-east-1'
CSV;
```



The screenshot shows the AWS Redshift console interface. On the left, there's a sidebar with navigation options like CLUSTERS, QUERIES, EDITOR, DATASHARES, CONFIG, MARKETPLACE, ADVISOR, ALARMS, EVENTS, and WHAT'S NEW. The main area is titled 'Resources info' and shows a list of tables under the 'public' schema. The 'Query1' editor is open, displaying the following SQL query:

```
1 copy atm_data.dim_atm from 's3://etlprojectbyishan/dim_atm/part-00000-c4425605-e626-4ed2-ad2-cef667cblb8-c000.csv'
2 iam_role 'arn:aws:iam::464886120274:role/redshift_s3_fullaccess'
3 delimiter ',' region 'us-east-1'
4 CSV;
```

Below the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' section shows 'Query 300' with a status of 'Completed, started on October 18, 2021 at 18:54:38' and an 'ELAPSED TIME: 00 m 10 s'.

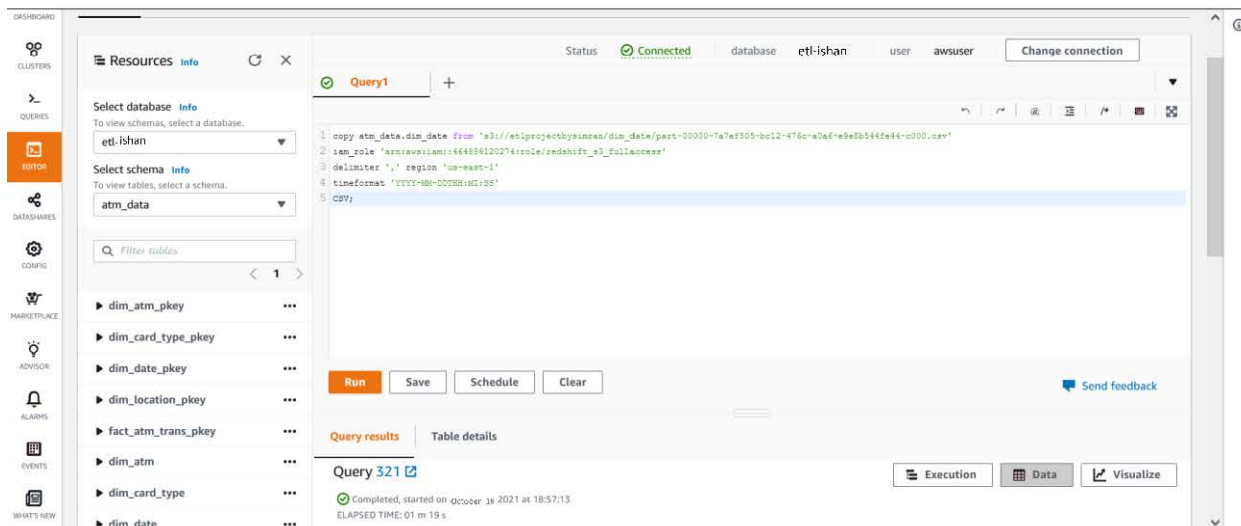
- Copying the data to dim_date table

copy atm_data.dim_date from ' s3://etlprojectbyishan/dim_date/part-00000-7a7ef505-bc12-476c-a0a6-e9e8b544fe44-c000.csv'

iam_role 'arn:aws:iam::464886120274:role/redshift_s3_fullaccess'

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

CSV;



The screenshot shows the AWS Redshift console interface. On the left, there's a sidebar with navigation options like CLUSTERS, QUERIES, EDITOR, DATASHARES, CONFIG, MARKETPLACE, ADVISOR, ALARMS, EVENTS, and WHAT'S NEW. The main area is titled 'Resources info' and shows a list of tables under the 'public' schema. The 'Query1' editor is open, displaying the following SQL query:

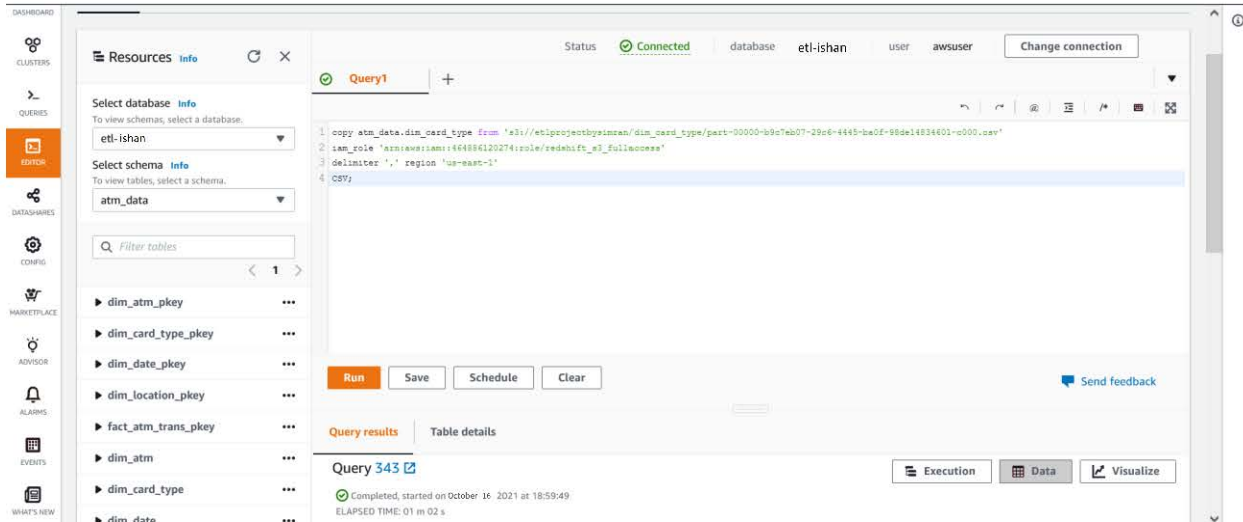
```
1 copy atm_data.dim_date from 's3://etlprojectbyishan/dim_date/part-00000-7a7ef505-bc12-476c-a0a6-e9e8b544fe44-c000.csv'
2 iam_role 'arn:aws:iam::464886120274:role/redshift_s3_fullaccess'
3 delimiter ',' region 'us-east-1'
4 timeformat 'YYYY-MM-DDTHH:MM:SS'
5 CSV;
```

Below the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' section shows 'Query 321' with a status of 'Completed, started on October 18, 2021 at 18:57:13' and an 'ELAPSED TIME: 01 m 19 s'.

- Copying the data to dim_card_type table

copy atm_data.dim_card_type from ' s3://etlprojectbyishan/dim_card_type/part-00000-b9c7eb07-29c6-4445-ba0f-98de14834601-c000.csv'

```
iam_role 'arn:aws:iam::464886120274:role/redshift_s3_fullaccess'
delimiter ',' region 'us-east-1'
CSV;
```



- Copying the data to fact_atm_trans table

```
copy atm_data.fact_atm_trans from ' s3://etlprojectbyishan/fact_atm_trans/
part-00000-978dd709-2ef2-4145-8ab5-9981558a8c60-c000.csv'
iam_role 'arn:aws:iam::464886120274:role/redshift_s3_fullaccess'
delimiter ',' region 'us-east-1'
CSV;
```

