

# Creation of a Redshift Cluster

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

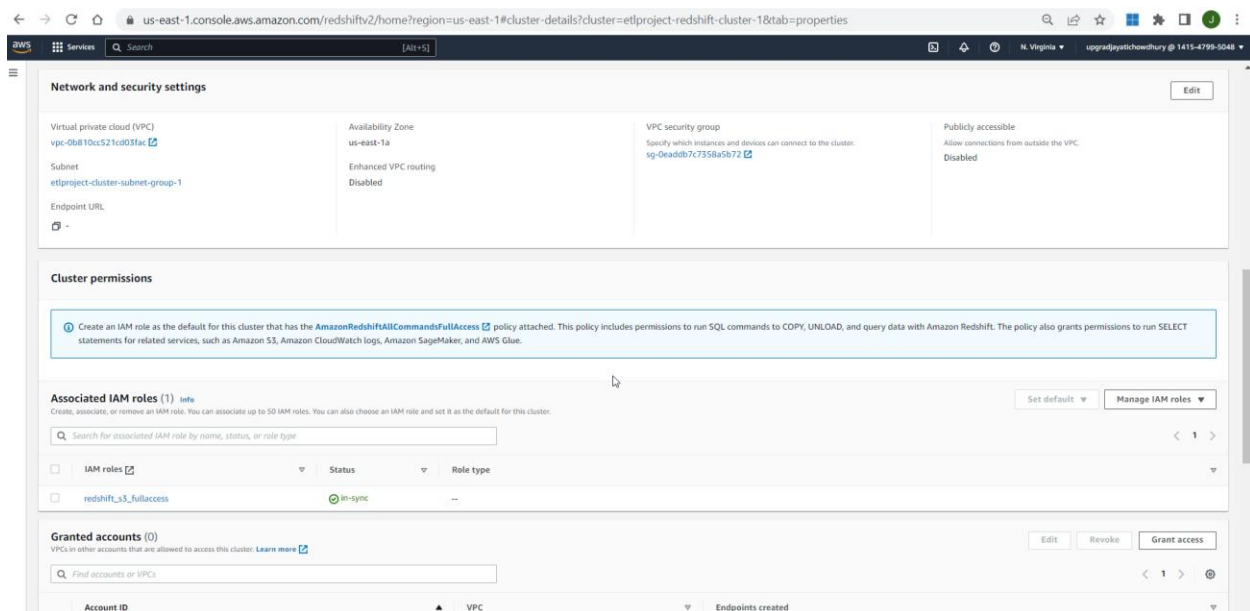
## Screenshot of the type of machine used along with number of nodes

The first screenshot shows the AWS Redshift console 'Clusters' page. It displays a table with one cluster, 'etlproject-redshift-cluster-1', which is in an 'Available' state. The cluster configuration shows it is a 'dc2.large' instance type with 2 nodes. The 'General information' section provides details about the cluster's status, creation date, storage usage, and multi-AZ configuration.

Cluster	Status	Cluster namespace	Availability Zone	Multi-AZ	Storage capacity us...	CPU utilization	Snapshots	Notificati...	Tags
etlproject-redshift-cluster-1 dc2.large   2 nodes   320 GB	Available	c6f7f74b-eb88-4763-...	us-east-1a	No	< 1%	8%	-		

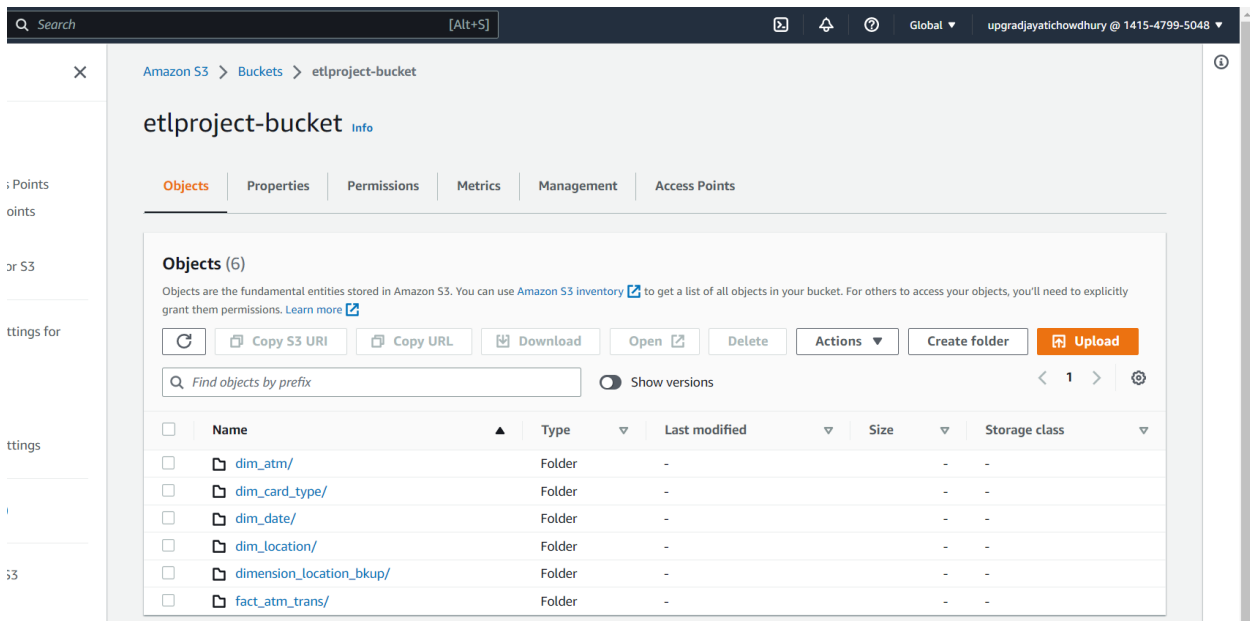
The second screenshot shows the 'etlproject-redshift-cluster-1' details page. It provides a comprehensive overview of the cluster's configuration, including the node type, number of nodes, and various URLs for connecting to the cluster.

General information			
Cluster identifier etlproject-redshift-cluster-1	Status Available	Node type dc2.large	Endpoint etlproject-redshift-cluster-1.c7eaugivv4j.us-east-1.redshift.amaz...
Cluster namespace c6f7f74b-eb88-4763-8b14-4d4e94b9ba31	Date created February 25, 2023, 09:05 (UTC+01:00)	Number of nodes 2	JDBC URL jdbc:redshift://etlproject-redshift-cluster-1.c7eaugivv4j.us-east-1...
Cluster configuration Production	Storage used 0.25% (0.81 of 320 GB used)		ODBC URL Driver=(Amazon Redshift (x64)); Server=etlproject-redshift-cluster...
	Multi-AZ No		



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

## Viewing all data in Amazon S3 Bucket



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

### Creating Schema for dimension and fact tables

create schema atm\_data;

Amazon Redshift > Query editor

Editor | Query history | Saved queries | Scheduled queries

Resources [Info](#)

Select database [Info](#)  
To view schemas, select a database.  
devetl

Select schema [Info](#)  
To view tables, select a schema.  
atm\_data

Filter tables

No resources  
No resources to display

Query 1

```
1 create schema atm_data
```

Run Save Schedule Clear

Query results | Table details

Query

Completed, started on February 25, 2023 at 09:13:28  
ELAPSED TIME: 00 m 57 s

## 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)  
);
```

The screenshot displays the Amazon Redshift Query Editor interface. At the top, the breadcrumb navigation shows 'Amazon Redshift > Query editor'. Below this is a tabbed interface with 'Editor', 'Query history', 'Saved queries', and 'Scheduled queries'. The 'Editor' tab is active, showing a left-hand sidebar with 'Resources' and 'Info' sections. The 'Resources' section includes dropdowns for 'Select database' (set to 'devetl') and 'Select schema' (set to 'public'), along with a 'Filter tables' search bar. The main area on the right shows a list of queries, with 'Query 2' selected and highlighted in orange. The SQL code for 'Query 2' is displayed in a text editor, showing the creation of the 'atm\_data.DIM\_LOCATION' table. Below the code editor are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. At the bottom, the 'Query results' tab is active, showing a status message: 'Completed, started on February 25, 2023 at 09:57:39' and 'ELAPSED TIME: 00 m 03 s'.

Amazon Redshift > Query editor

Editor | Query history | Saved queries | Scheduled queries

Resources Info

Select database Info  
To view schemas, select a database.  
devetl

Select schema Info  
To view tables, select a schema.  
public

Filter tables

No resources  
No resources to display

Query 1 Query 2 +

```
1 create table atm_data.DIM_LOCATION
2 (
3 location_id int not null DISTKEY SORTKEY,
4 location varchar(50),
5 streetname varchar(255),
6 street_number int,
7 zipcode int,
8 lat decimal(10,3),
9 lon decimal(10,3),
10 PRIMARY KEY(location_id)
11 );
```

Run Save Schedule Clear

Query results | Table details

Query

Completed, started on February 25, 2023 at 09:57:39  
ELAPSED TIME: 00 m 03 s

## 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)  
);
```

The screenshot displays a database management interface. On the left, a sidebar titled 'Resources' shows the database 'devetl' and schema 'atm\_data'. Below this, a list of tables is visible: 'dim\_atm\_pkey', 'dim\_location\_pkey', 'dim\_atm', and 'dim\_location'. The main area on the right shows a SQL query editor with the following code:

```
1 create table atm_data.DIM_ATM
2 (
3   atm_id int not null DISTKEY SORTKEY,
4   atm_number varchar(20),
5   atm_manufacturer varchar(50),
6   atm_location_id int,
7   PRIMARY KEY(atm_id),
8   FOREIGN KEY(atm_location_id) references atm_data.DIM_LOCATION(location_id)
9 );
10
```

Below the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Run' button is highlighted in orange. Below these buttons, there are tabs for 'Query results' and 'Table details'. The 'Query results' tab is active, showing a status message: 'Completed, started on February 25, 2023 at 10:03:27' and 'ELAPSED TIME: 00 m 08 s'.

## 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)  
);
```

Amazon Redshift > Query editor

Editor | Query history | Saved queries | Scheduled queries

Resources [Info](#)

Select database [Info](#)  
To view schemas, select a database.  
devetl

Select schema [Info](#)  
To view tables, select a schema.  
atm\_data

Filter tables

< 1 >

- ▶ dim\_atm\_pkey ...
- ▶ dim\_date\_pkey ...
- ▶ dim\_location\_pkey ...
- ▶ dim\_atm ...
- ▶ dim\_date ...
- ▶ dim\_location ...

Query 1 × | Query 2 × | Query 3 × | Query 4 × | +

```
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 };
```

Run Save Schedule Clear

Query results | Table details

Query

✓ Completed, started on February 25, 2023 at 10:05:18  
ELAPSED TIME: 00 m 06 s

## 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)  
);
```

Amazon Redshift > Query editor

Editor | Query history | Saved queries | Scheduled queries

Resources [Info](#)

Select database [Info](#)  
To view schemas, select a database.  
devetl

Select schema [Info](#)  
To view tables, select a schema.  
atm\_data

Filter tables

< 1 >

- ▶ dim\_atm\_pkey ...
- ▶ dim\_card\_type\_pkey ...
- ▶ dim\_date\_pkey ...
- ▶ dim\_location\_pkey ...
- ▶ dim\_atm ...
- ▶ dim\_card\_type ...
- ▶ dim\_date ...
- ▶ dim\_location ...

Query 1 x | Query 2 x | Query 3 x | Query 4 x | Query 5 x

```
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 );
```

Run Save Schedule Clear

Query results | Table details

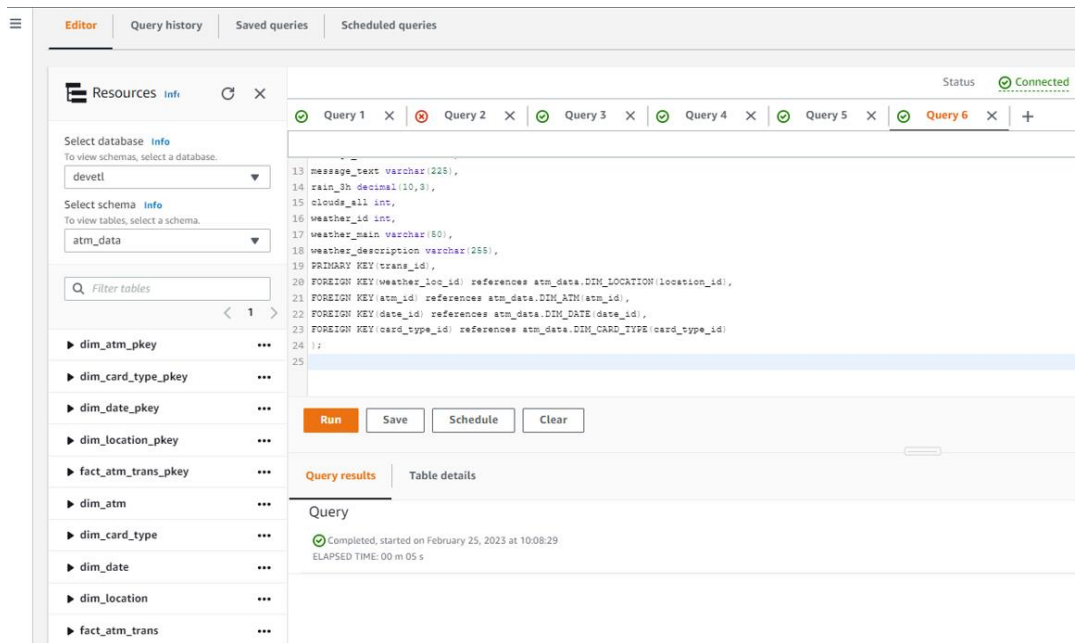
Query

Completed, started on February 25, 2023 at 10:07:05  
ELAPSED TIME: 00 m 04 s

### **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_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)
);
```





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

### Copying the data to dim\_location table

copy atm\_data.dim\_location from 's3://etlproject-bucket/dim\_location/part-00000-7173dbd0-c49d-4c3c-ab8b-abae58b1f91c-c000.csv'

iam\_role 'arn:aws:iam::141547995048:role/redshift\_s3\_fullaccess'

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

CSV;

us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#query-editor:

ServicesSearch[Alt+S]

N. Virginiaupgradjayatichowdhury @ 1415-4799-5048

ResourcesInfo

Select database Info

To view schemas, select a database.

devetl

Select schema Info

To view tables, select a schema.

public

Filter tables

No resources

No resources to display

Query 1Query 2Query 3Query 4Query 7

```
1 copy atm_data.dim_location from 's3://etlproject-bucket/dim_location/part-00000-7173dbd0-c49d-4c3c-ab8b-abae58b1f91c-c000.csv'
2 iam_role 'arn:aws:iam::141547995048:role/redshift_s3_fullaccess'
3 delimiter ',' region 'us-east-1'
4 CSV:
5
```

RunSaveScheduleClear

Send feedback

Query resultsTable details

Query 1000135

ExecutionDataVisualize

Completed, started on February 28, 2023 at 06:30:55

ELAPSED TIME: 00 m 05 s

## Copying the data to dim\_atm table

copy atm\_data.dim\_atm from 's3://etlproject-bucket/dim\_atm/part-00000-09b2500e-0ee1-4902-a6a1-214c331275ac-c000.csv'

iam\_role 'arn:aws:iam::141547995048:role/redshift\_s3\_fullaccess'

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

CSV;

The screenshot shows the AWS Redshift Query Editor interface. The browser address bar displays the URL: `us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#query-editor:`. The interface includes a top navigation bar with tabs for 'Editor', 'Query history', 'Saved queries', and 'Scheduled queries'. On the left, the 'Resources' panel shows the selected database as 'devetl' and the schema as 'atm\_data'. The main editor area contains a SQL query with four lines: `1 copy atm_data.dim_atm from 's3://etlproject-bucket/dim_atm/part-00000-09b2500e-0ee1-4902-a6a1-214c331275ac-c000.csv'`, `2 iam_role 'arn:aws:iam::141547995048:role/redshift_s3_fullaccess'`, `3 delimiter ',' region 'us-east-1'`, and `4 CSV;`. Below the query editor are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The bottom section shows 'Query results' for 'Query 1000287', indicating it is 'Completed, started on February 28, 2023 at 06:40:49' with an 'ELAPSED TIME: 00 m 10 s'. The right sidebar shows the connection status as 'Connected' to the 'database' 'devetl' using the 'awsuser'.

## Copying the data to dim\_date table

copy atm\_data.dim\_date from 's3://etlproject-bucket/dim\_date/part-00000-b0a2f360-e9ce-48fa-8ad7-472676e1aa9e-c000.csv'

iam\_role 'arn:aws:iam::141547995048:role/redshift\_s3\_fullaccess'

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

CSV

TIMEFORMAT 'auto';

The screenshot displays the Amazon Redshift Query Editor interface. The top navigation bar shows the AWS logo, a search bar, and the user's profile. The main interface is divided into several sections:

- Resources:** A sidebar on the left with tabs for 'Select database', 'Select schema', and 'Filter tables'. The 'Select database' dropdown is set to 'devetl', and the 'Select schema' dropdown is set to 'public'. Below these, it states 'No resources' and 'No resources to display'.
- Query Editor:** The central area contains a SQL query in a text editor. The query is as follows:

```
1 copy atm_data.dim_date from 's3://etlproject-bucket/dim_date/part-00000-b0a2f360-e9ce-48fa-8ad7-472676e1aa9e-c000.csv'
2 iam_role 'arn:aws:iam::141547995048:role/redshift_s3_fullaccess'
3 delimiter ',' region 'us-east-1'
4 CSV
5 TIMEFORMAT 'auto';
6
```
- Query Execution:** Below the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Run' button is highlighted in orange.
- Query Results:** At the bottom, the 'Query results' tab is active, showing 'Query 1400145'. It indicates the query is 'Completed, started on February 28, 2023 at 10:00:57' and 'ELAPSED TIME: 00 m 00 s'.

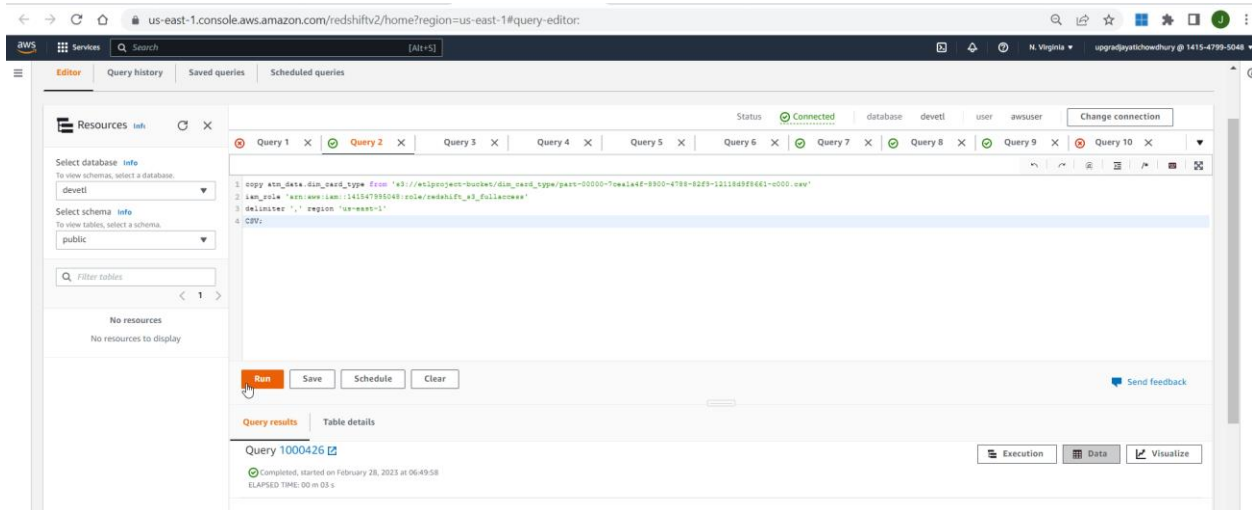
## Copying the data to dim\_card\_type table

copy atm\_data.dim\_card\_type from 's3://etlproject-bucket/dim\_card\_type/part-00000-7cea1a4f-8900-4788-82f9-12118d9f8661-c000.csv'

iam\_role 'arn:aws:iam::141547995048: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://etlproject-bucket/fact\_atm\_trans/part-00000-349988c6-cf3c-4636-8c56-8b0314a2fbc1-c000.csv'

iam\_role 'arn:aws:iam::141547995048:role/redshift\_s3\_fullaccess'

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

CSV;

The screenshot displays the AWS Redshift Query Editor interface. The top navigation bar shows the AWS logo, a search bar, and the current region 'us-east-1'. The main interface is divided into several sections:

- Resources:** A sidebar on the left showing a list of databases and schemas. The 'atm\_data' schema is selected, and a list of tables is displayed, 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'.
- Query Editor:** The central area where the SQL query is written. The query is as follows:

```
1 copy atm_data.fact_atm_trans from 's3://etlproject-bucket/fact_atm_trans/part-00000-349988c6-cf3c-4636-8c56-8b0314a2fbc1-c000.csv'
2 iam_role 'arn:aws:iam::141547995048:role/redshift_s3_fullaccess'
3 delimiter ',' region 'us-east-1'
4 CSV;
5
```
- Query Results:** A section at the bottom showing the execution status of the query. It indicates that the query 'Query 1000476' is 'Completed' and started on February 28, 2023, at 06:53:40. The elapsed time is 00 m 21 s.