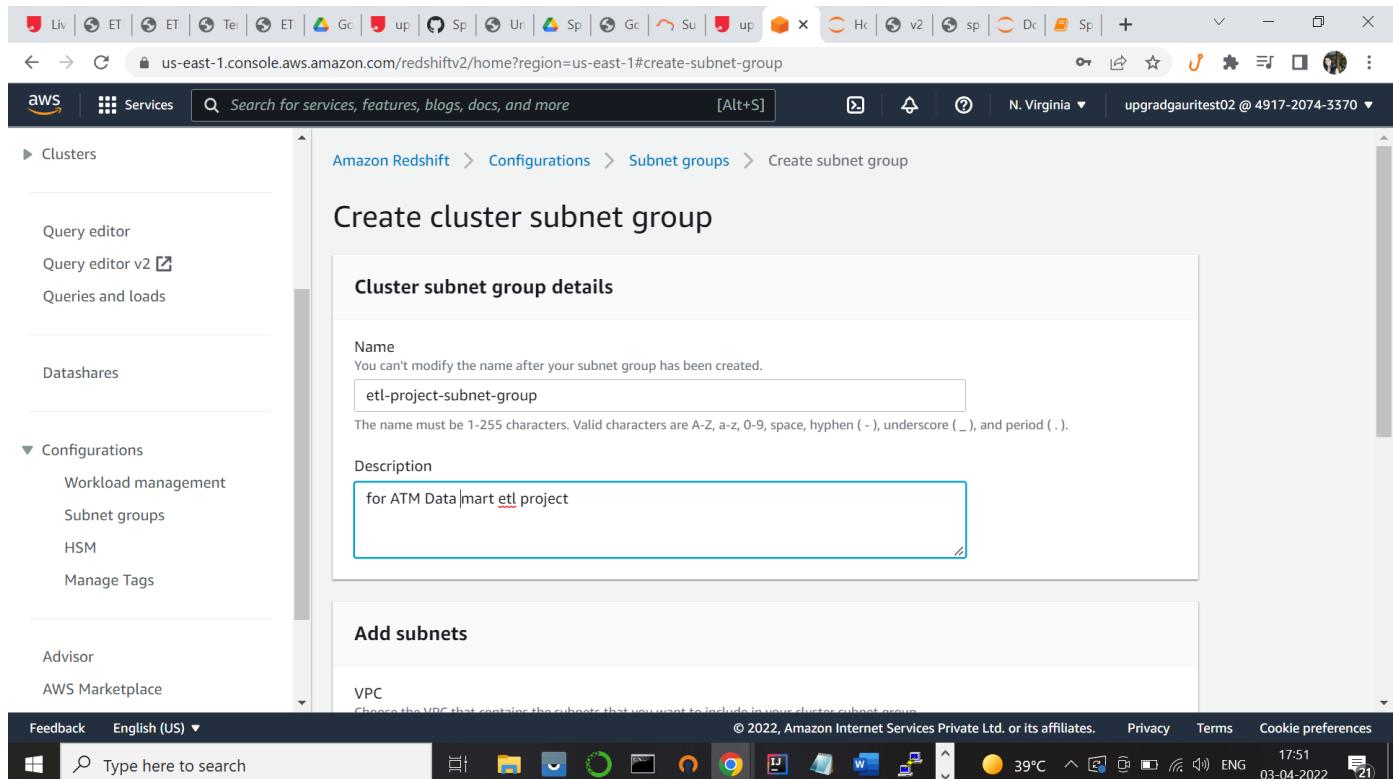


## Creation of a Redshift Cluster

### Screenshots of the configuration of the created Redshift cluster:

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

#### 1. Creating Subnet group:



The screenshot shows the 'Create cluster subnet group' page in the AWS Amazon Redshift console. The left sidebar shows navigation links for Clusters, Query editor, Datashares, Configurations (selected), Workload management, Subnet groups, HSM, Manage Tags, Advisor, and AWS Marketplace. The main content area has two sections: 'Cluster subnet group details' and 'Add subnets'. In 'Cluster subnet group details', the 'Name' field is set to 'etl-project-subnet-group' and the 'Description' field contains 'for ATM Data | mart etl project'. In 'Add subnets', there is a 'VPC' section with a note: 'Choose the VPC that contains the subnets that you want to include in your cluster subnet group.' The bottom of the screen shows the Windows taskbar with various pinned icons and system status information.

Screenshot of the AWS Redshift console showing the creation of a cluster subnet group.

**Add subnets**

**VPC**  
Choose the VPC that contains the subnets that you want to include in your cluster subnet group.  
new\_vpc  
vpc-005e937b3ae428543

**Add all the subnets for this VPC**

**Availability Zone** us-east-1d (1)    **Subnet** subnet-0b12ea2623ee66f7b    **Add subnet**

**Subnets in this cluster subnet group (1)**

Availability Zone	Subnet ID	CIDR block	Action
us-east-1d	subnet-0b12ea2623ee66f7b	10.0.0.0/24	<b>Remove</b>

**Create cluster subnet group**

Screenshot of the AWS Redshift console showing the successful creation of a cluster subnet group.

**Cluster subnet group etl-project-subnet-group was created successfully**

**Amazon Redshift > Configurations > Subnet groups**

**Cluster subnet groups (2)**

Name	Status	VPC ID	Description	Tags
etl-project-subnet-group	Complete	vpc-005e937b3ae428543	for ATM Data m...	
upgrad-cluster-subnet-group	Complete	vpc-005e937b3ae428543	for redshift	

## 2. Creating a Redshift cluster:

Screenshot 1: Initial 'Create cluster' configuration page.

The 'Cluster identifier' field contains 'etl-project-redshift-cluster'. The 'What are you planning to use this cluster for?' section shows 'Production' selected. The 'Choose the size of the cluster' section has 'I'll choose' selected.

Screenshot 2: Configuration summary page.

The 'Node type' dropdown is set to 'dc2.large'. The 'Number of nodes' input field shows '2'. The 'Configuration summary' section displays the cost '\$360.00/month' and storage capacity '320 GB'.

us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#create-cluster

Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia upgradgauritest02 @ 4917-2074-3370

### Amazon Redshift provisioned clusters

- Redshift serverless [New](#)
- Provisioned clusters dashboard
- Clusters
  - Reserved nodes
  - Snapshots
- Query editor
- Query editor v2 [Edit](#)
- Queries and loads
- Datasources

### 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

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.

Feedback English (US) © 2022, Amazon Internet Services Private Ltd. or its affiliates. Privacy Terms Cookie preferences 17:58 03-04-2022

us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#create-cluster

Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia upgradgauritest02 @ 4917-2074-3370

### Amazon Redshift provisioned clusters

- Redshift serverless [New](#)
- Provisioned clusters dashboard
- Clusters
  - Reserved nodes
  - Snapshots
- Query editor
- Query editor v2 [Edit](#)
- Queries and loads
- Datasources

### Network and security

Virtual private cloud (VPC)  
This VPC defines the virtual networking environment for this cluster.  
  
vpc-005e937b3ae428543

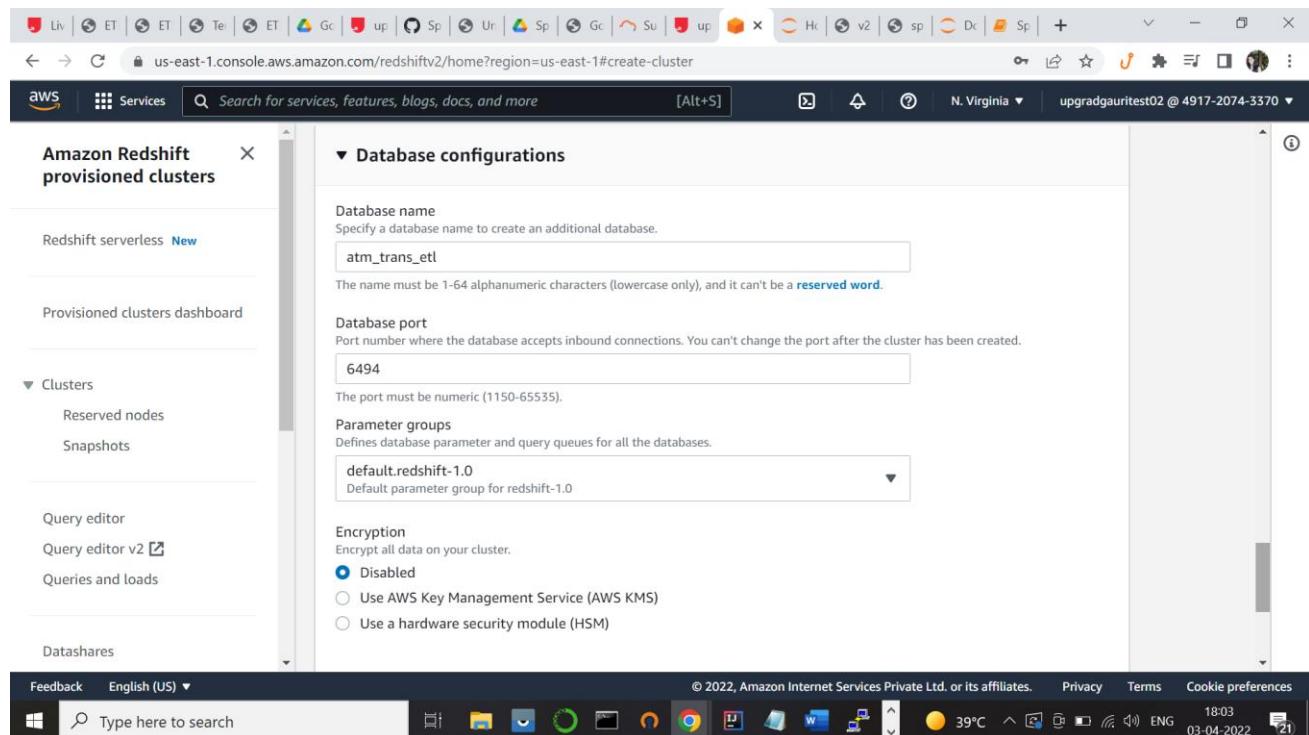
You can't change the VPC associated with this cluster after the cluster has been created. [Learn more](#)

VPC security groups  
This VPC security group defines which subnets and IP ranges the cluster can use in the VPC.  
  
cloudera\_new [X](#)  
sg-075ead2fb7358867

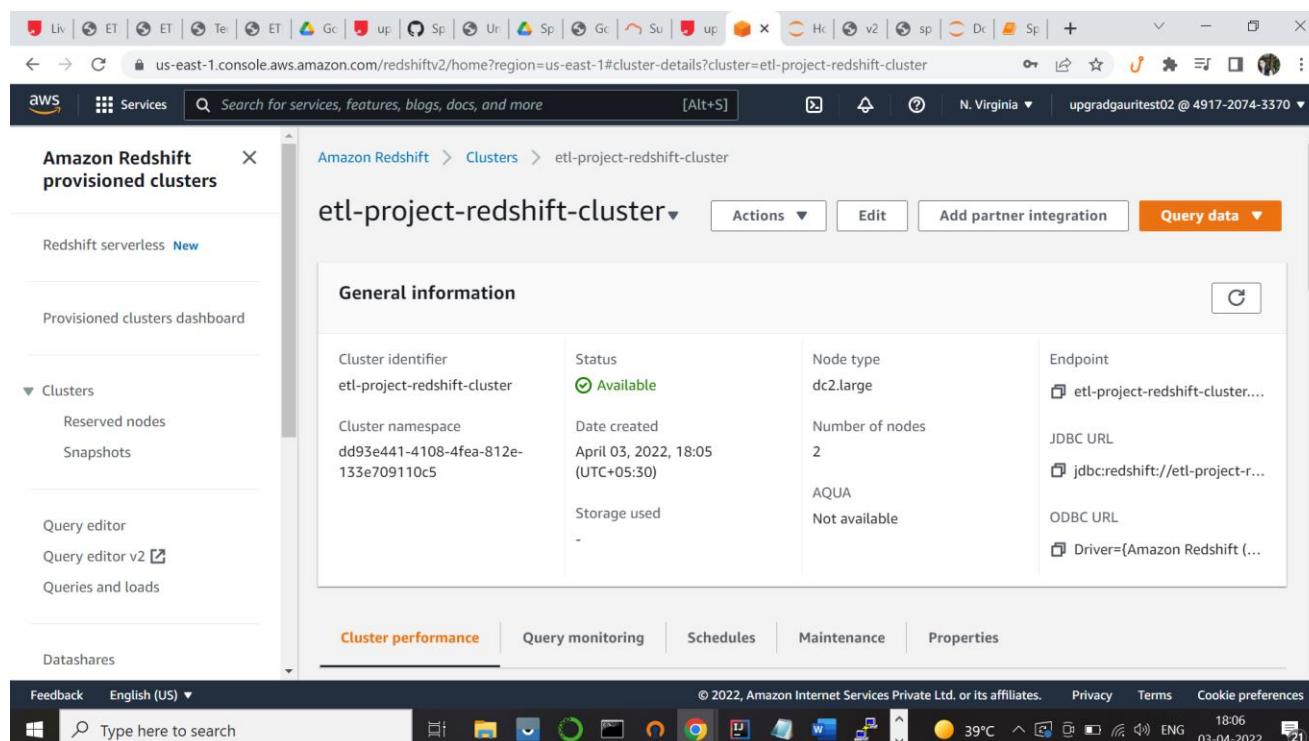
Cluster subnet group  
Choose the Amazon Redshift subnet group to launch the cluster in.

Availability Zone

Feedback English (US) © 2022, Amazon Internet Services Private Ltd. or its affiliates. Privacy Terms Cookie preferences 18:01 03-04-2022

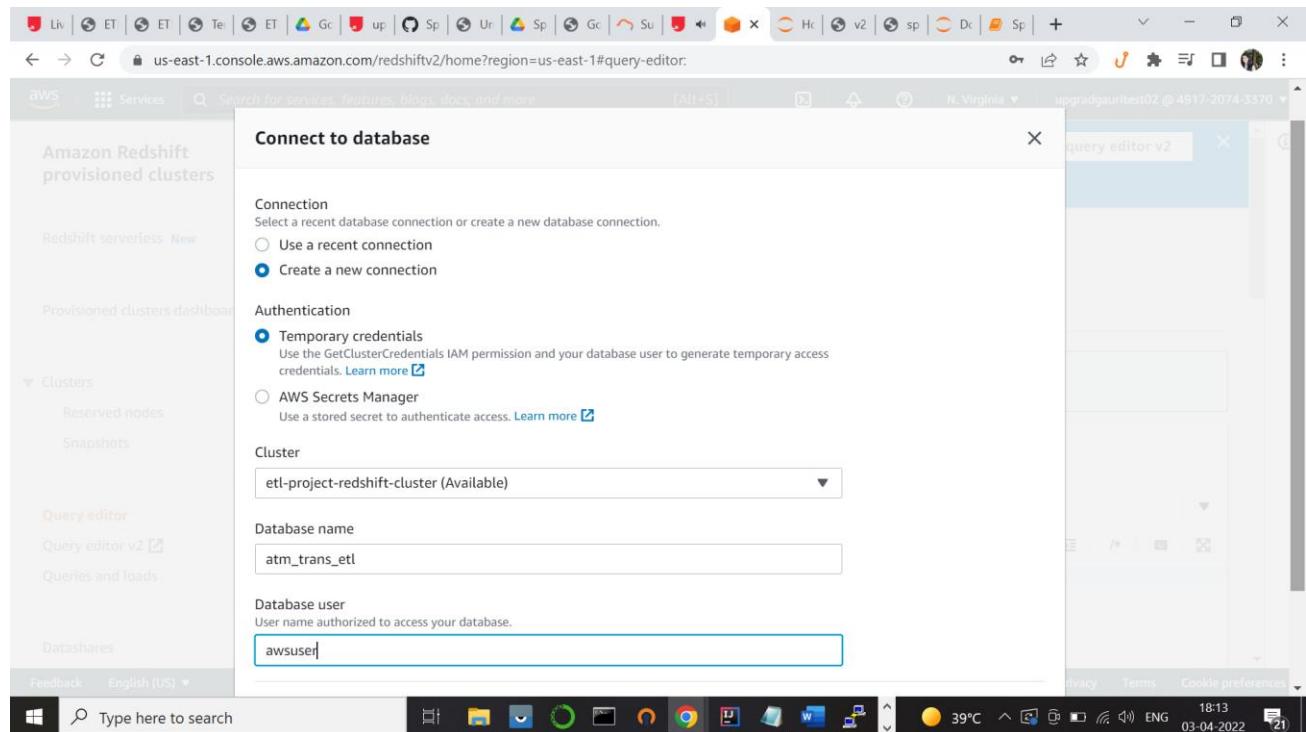


The screenshot shows the 'Database configurations' section for creating a new cluster. The database name is set to 'atm\_trans\_elt'. The database port is set to 6494. The parameter group is set to 'default.redshift-1.0'. The encryption setting is set to 'Disabled'.

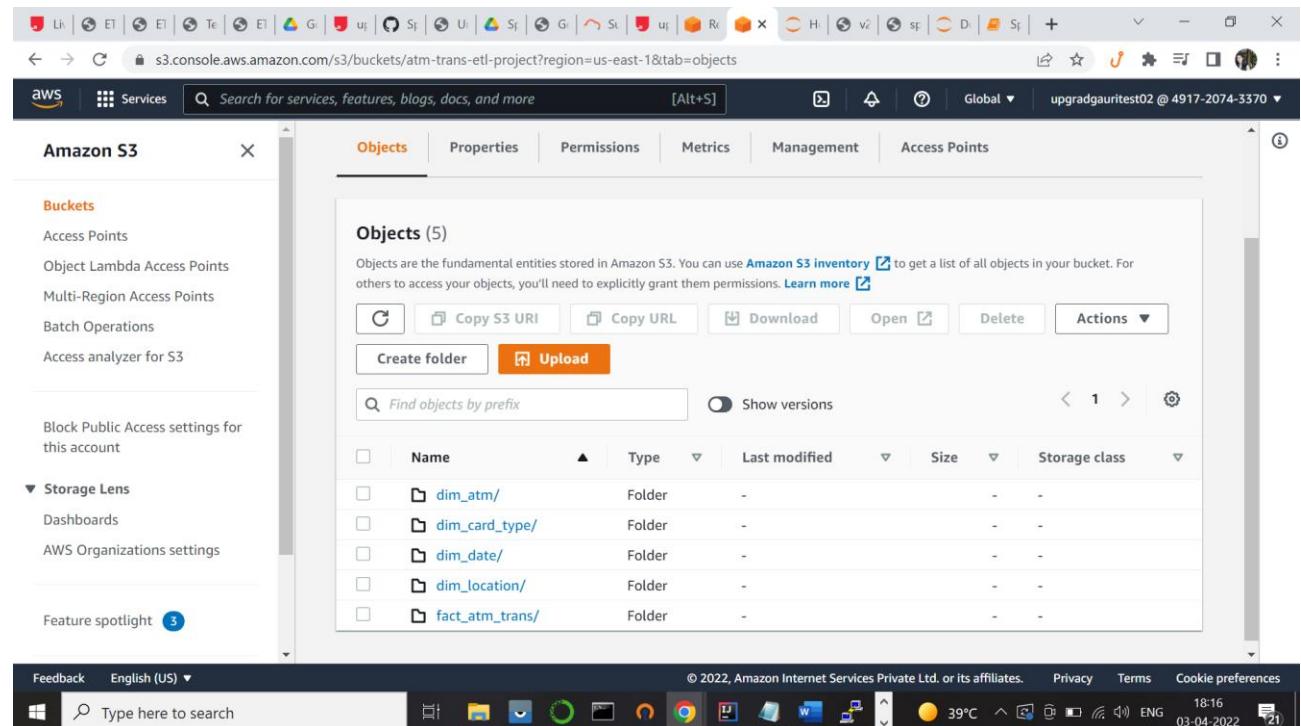


The screenshot shows the 'General information' section for the cluster 'etl-project-redshift-cluster'. The cluster identifier is 'etl-project-redshift-cluster', the status is 'Available', the node type is 'dc2.large', and the endpoint is 'etl-project-redshift-cluster....'. The cluster namespace is 'dd93e441-4108-4fea-812e-133e709110c5'. The date created was April 03, 2022, at 18:05 (UTC+05:30). There are 2 nodes, and the storage used is AQUA. The storage used is 'Not available'.

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



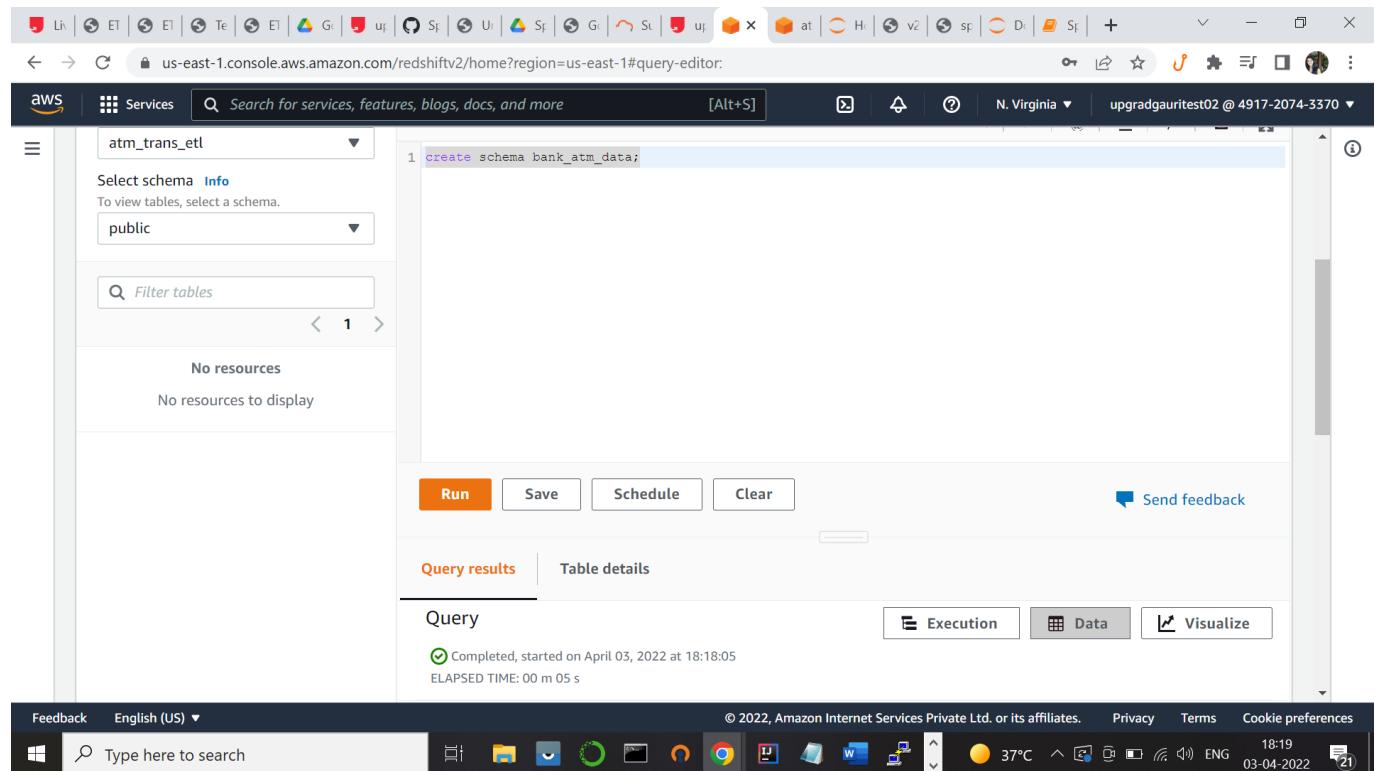
## Amazon S3 Bucket:



The screenshot shows the AWS S3 console interface. The left sidebar has sections for Buckets, Storage Lens, and Feature spotlight. The main area is titled 'Objects (5)' and lists the following 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	-	-	-

## Query to create schema: bank\_atm\_data



The screenshot shows the AWS Redshift Query Editor interface. The query window contains the following SQL command:

```
1 create schema bank_atm_data;
```

The left sidebar shows the schema selection dropdown set to "atm\_trans\_etl" and the "Info" tab selected. Below it, there is a "Select schema" dropdown set to "public" and a "Filter tables" search bar.

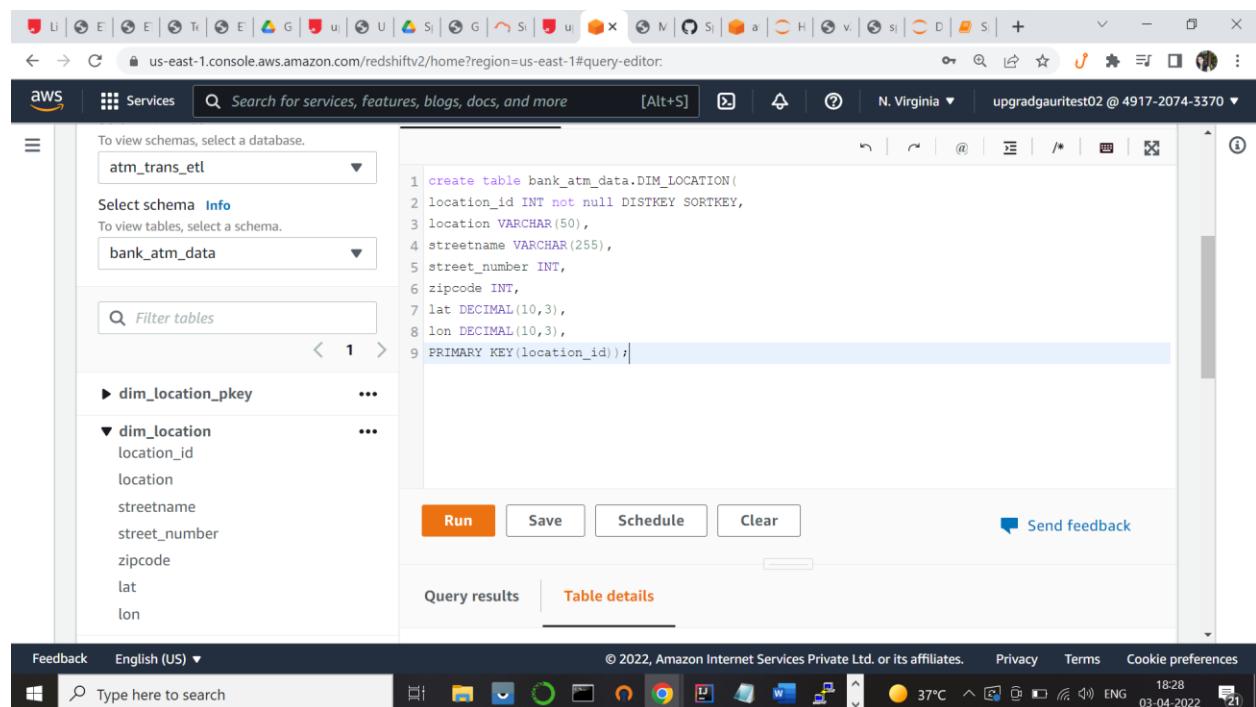
At the bottom of the editor, there are four buttons: "Run", "Save", "Schedule", and "Clear". To the right of these buttons is a "Send feedback" link. Below the editor, there are tabs for "Query results" and "Table details", with "Query results" currently selected. The "Query" tab shows the completed status of the query.

At the very bottom of the screen, a Windows taskbar is visible, showing various open applications like File Explorer, Edge, and a browser.

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

### 1. Query to create DIM\_LOCATION table:

```
create table bank_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));
```

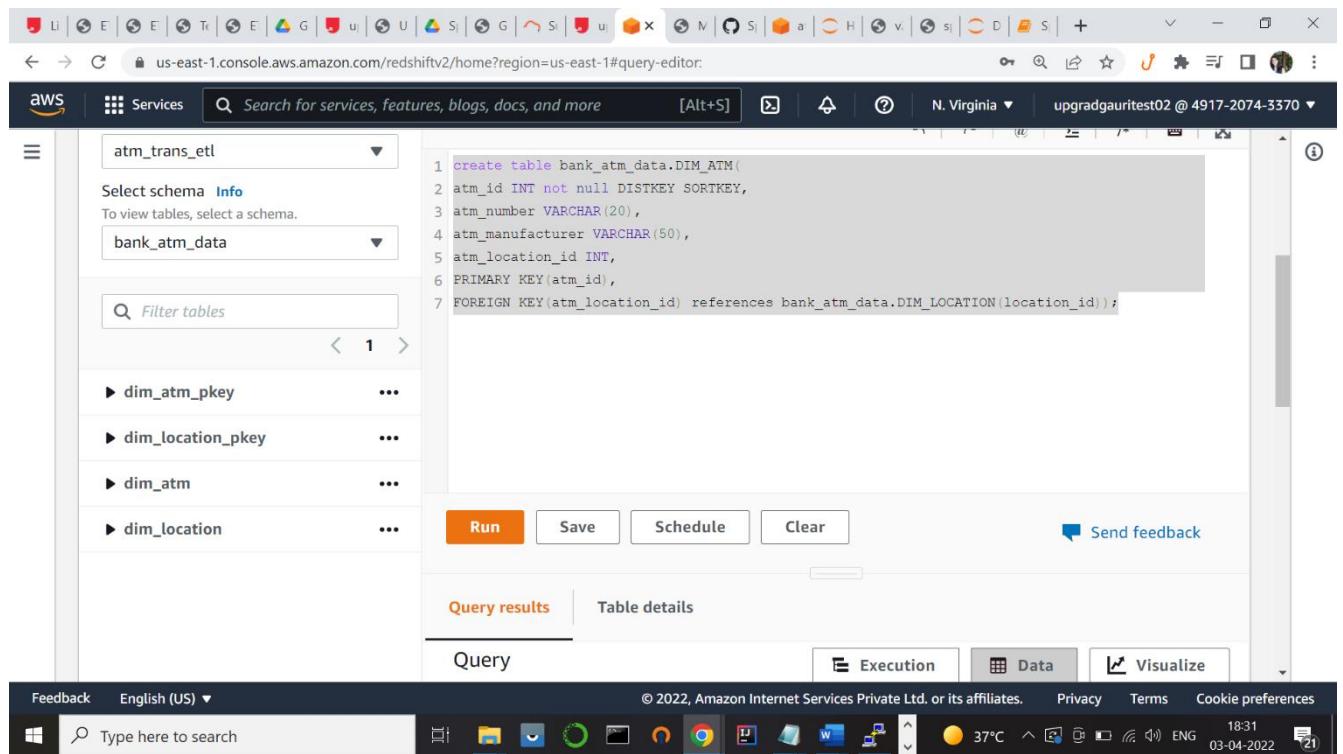


The screenshot shows the AWS Redshift Query Editor interface. The top navigation bar includes links for various AWS services like Lambda, EBS, ETL, etc., and the URL is [us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#query-editor](https://us-east-1.console.aws.amazon.com/redshiftv2/home?region=us-east-1#query-editor). The main area displays the SQL query for creating the DIM\_LOCATION table, which is identical to the one shown in the previous code block. Below the query, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The bottom of the screen shows the Windows taskbar with various pinned icons and system status information.

```
1 create table bank_atm_data.DIM_LOCATION(
2     location_id INT not null DISTKEY SORTKEY,
3     location VARCHAR(50),
4     streetname VARCHAR(255),
5     street_number INT,
6     zipcode INT,
7     lat DECIMAL(10,3),
8     lon DECIMAL(10,3),
9     PRIMARY KEY(location_id));
```

## 2. Query to create DIM\_ATM table:

```
create table bank_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
    bank_atm_data.DIM_LOCATION(location_id));
```



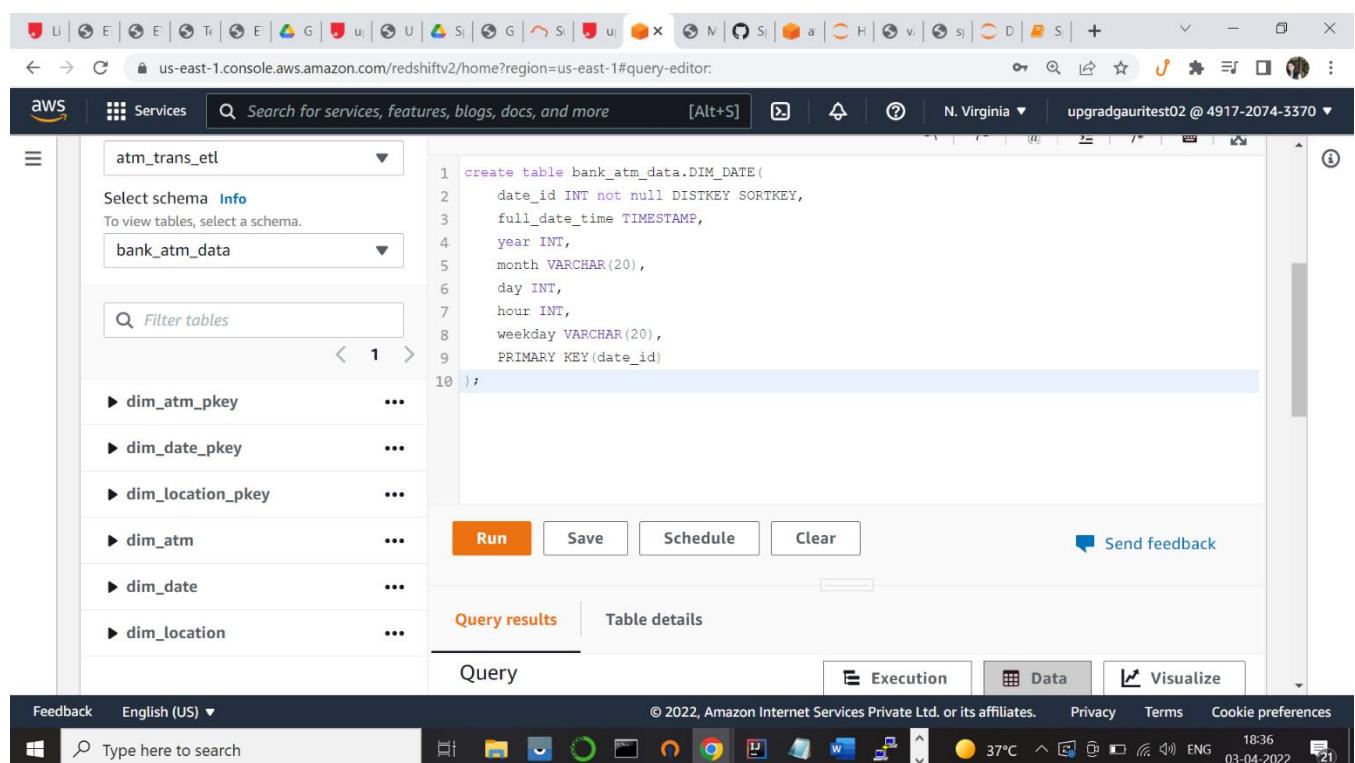
The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar with a tree view showing database schemas: 'atm\_trans\_etl' (selected), 'bank\_atm\_data' (selected), and 'dim\_atm\_pkey', 'dim\_location\_pkey', 'dim\_atm', 'dim\_location'. The main area is a query editor with the following content:

```
1 create table bank_atm_data.DIM_ATM(
2     atm_id INT not null DISTKEY SORTKEY,
3     atm_number VARCHAR(20),
4     atm_manufacturer VARCHAR(50),
5     atm_location_id INT,
6     PRIMARY KEY(atm_id),
7     FOREIGN KEY(atm_location_id) references bank_atm_data.DIM_LOCATION(location_id));
```

Below the query, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. At the bottom, tabs for 'Query results' and 'Table details' are visible, along with buttons for 'Execution', 'Data', and 'Visualize'. The status bar at the bottom shows system information like 'Feedback English (US)', '© 2022, Amazon Internet Services Private Ltd. or its affiliates.', 'Privacy Terms Cookie preferences', and system metrics like '37°C', '18:31', '03-04-2022', and battery level.

### 3. Query to create DIM\_DATE table:

```
create table bank_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)
);
```



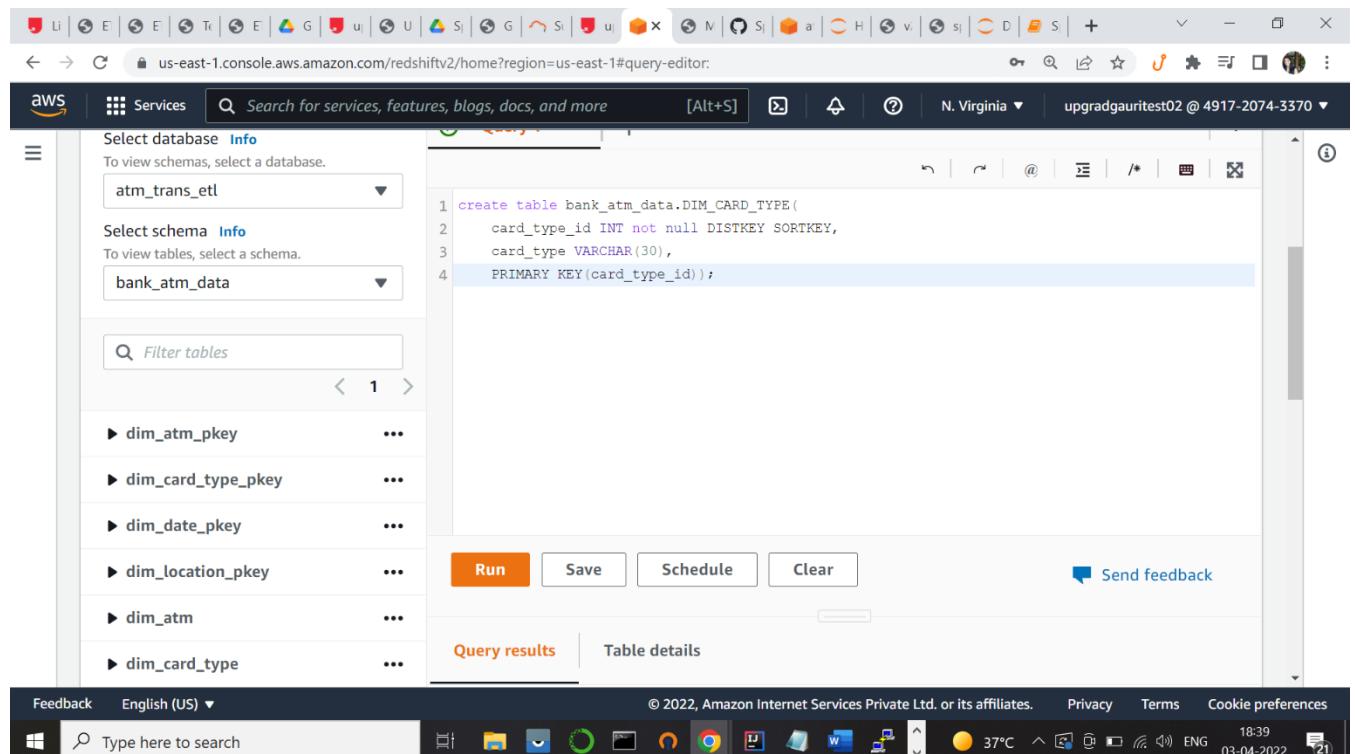
The screenshot shows the AWS Redshift Query Editor interface. The query editor window displays the SQL code for creating the DIM\_DATE table. The code is as follows:

```
1 create table bank_atm_data.DIM_DATE(
2     date_id INT not null DISTKEY SORTKEY,
3     full_date_time TIMESTAMP,
4     year INT,
5     month VARCHAR(20),
6     day INT,
7     hour INT,
8     weekday VARCHAR(20),
9     PRIMARY KEY(date_id)
10 );
```

The interface includes a sidebar with schema navigation (atm\_trans\_etl, bank\_atm\_data), a search bar, and a toolbar with buttons for Run, Save, Schedule, and Clear. Below the query editor, there are tabs for Query results and Table details, and buttons for Execution, Data, and Visualize. The bottom of the screen shows the Windows taskbar with various pinned icons and system status information.

#### 4. Query to create DIM\_CARD\_TYPE table:

```
create table bank_atm_data.DIM_CARD_TYPE(
    card_type_id INT not null DISTKEY SORTKEY,
    card_type VARCHAR(30),
    PRIMARY KEY(card_type_id));
```

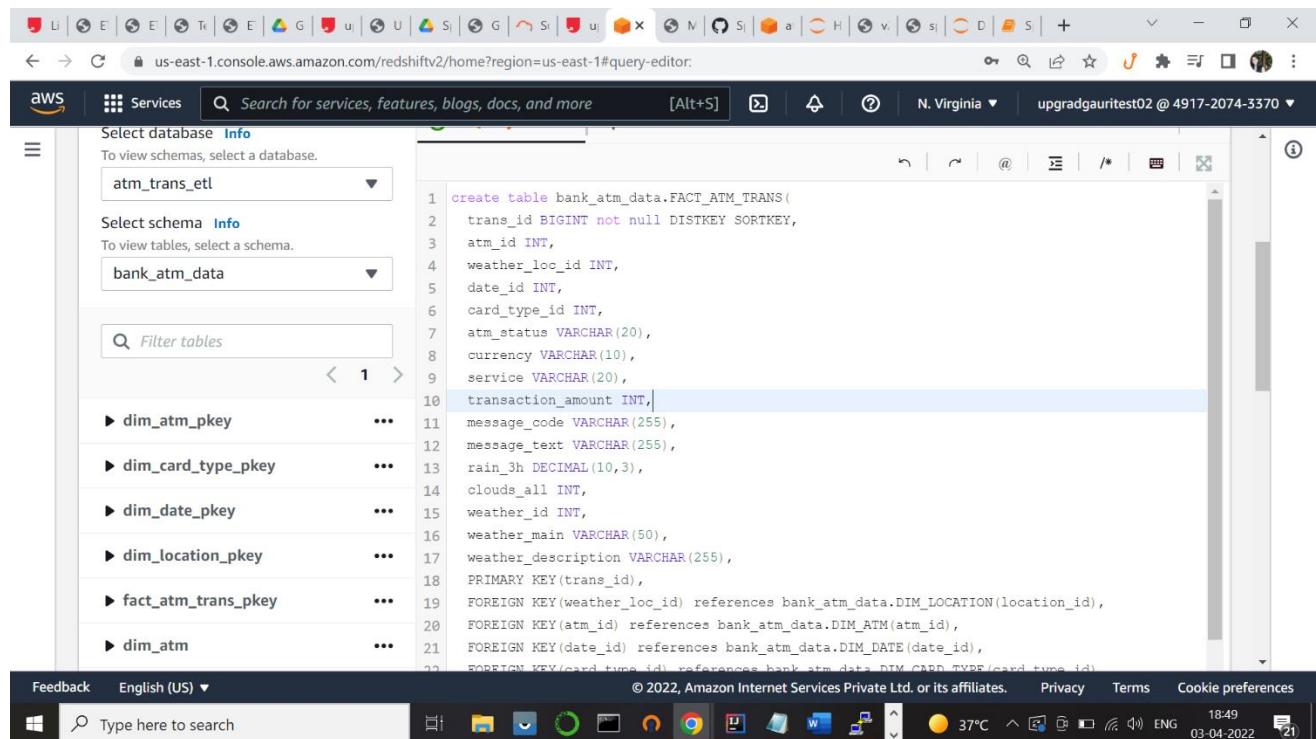


The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar for selecting the database (atm\_trans\_elt) and schema (bank\_atm\_data). Below that is a list of tables: dim\_atm\_pkey, dim\_card\_type\_pkey, dim\_date\_pkey, dim\_location\_pkey, dim\_atm, and dim\_card\_type. The main area contains the SQL query for creating the table. At the bottom, there are buttons for Run, Save, Schedule, and Clear, along with a Send feedback link. The status bar at the bottom right shows the date (03-04-2022), time (18:39), and system temperature (37°C).

```
1 create table bank_atm_data.DIM_CARD_TYPE(
2     card_type_id INT not null DISTKEY SORTKEY,
3     card_type VARCHAR(30),
4     PRIMARY KEY(card_type_id));
```

**5. Query to create FACT\_ATM\_TRANS table:**

```
create table bank_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(255),
    message_text VARCHAR(255),
    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 bank_atm_data.DIM_LOCATION(location_id),
    FOREIGN KEY(atm_id) references bank_atm_data.DIM_ATM(atm_id),
    FOREIGN KEY(date_id) references bank_atm_data.DIM_DATE(date_id),
    FOREIGN KEY(card_type_id) references bank_atm_data.DIM_CARD_TYPE(card_type_id)
);
```



The screenshot shows the AWS Redshift Query Editor interface. The top navigation bar includes links for various AWS services like Lambda, CloudWatch Metrics, and S3. The main header says "Search for services, features, blogs, docs, and more". Below it, the database and schema are selected: "Select database **Info**" and "Select schema **bank\_atm\_data**". A search bar "Filter tables" is present. On the left, a sidebar lists tables: "dim\_atm\_pkey", "dim\_card\_type\_pkey", "dim\_date\_pkey", "dim\_location\_pkey", "fact\_atm\_trans\_pkey", and "dim\_atm". The main content area displays a SQL script for creating a table:

```
1 create table bank_atm_data.FACT_ATM_TRANS (
2     trans_id BIGINT not null DISTKEY SORTKEY,
3     atm_id INT,
4     weather_loc_id INT,
5     date_id INT,
6     card_type_id INT,
7     atm_status VARCHAR(20),
8     currency VARCHAR(10),
9     service VARCHAR(20),
10    transaction_amount INT,
11    message_code VARCHAR(255),
12    message_text VARCHAR(255),
13    rain_3h DECIMAL(10,3),
14    clouds_all INT,
15    weather_id INT,
16    weather_main VARCHAR(50),
17    weather_description VARCHAR(255),
18    PRIMARY KEY(trans_id),
19    FOREIGN KEY(weather_loc_id) REFERENCES bank_atm_data.DIM_LOCATION(location_id),
20    FOREIGN KEY(atm_id) REFERENCES bank_atm_data.DIM_ATM(atm_id),
21    FOREIGN KEY(date_id) REFERENCES bank_atm_data.DIM_DATE(date_id),
22    FOREIGN KEY(card_type_id) REFERENCES bank_atm_data.DIM_CARD_TYPE(card_type_id)
```

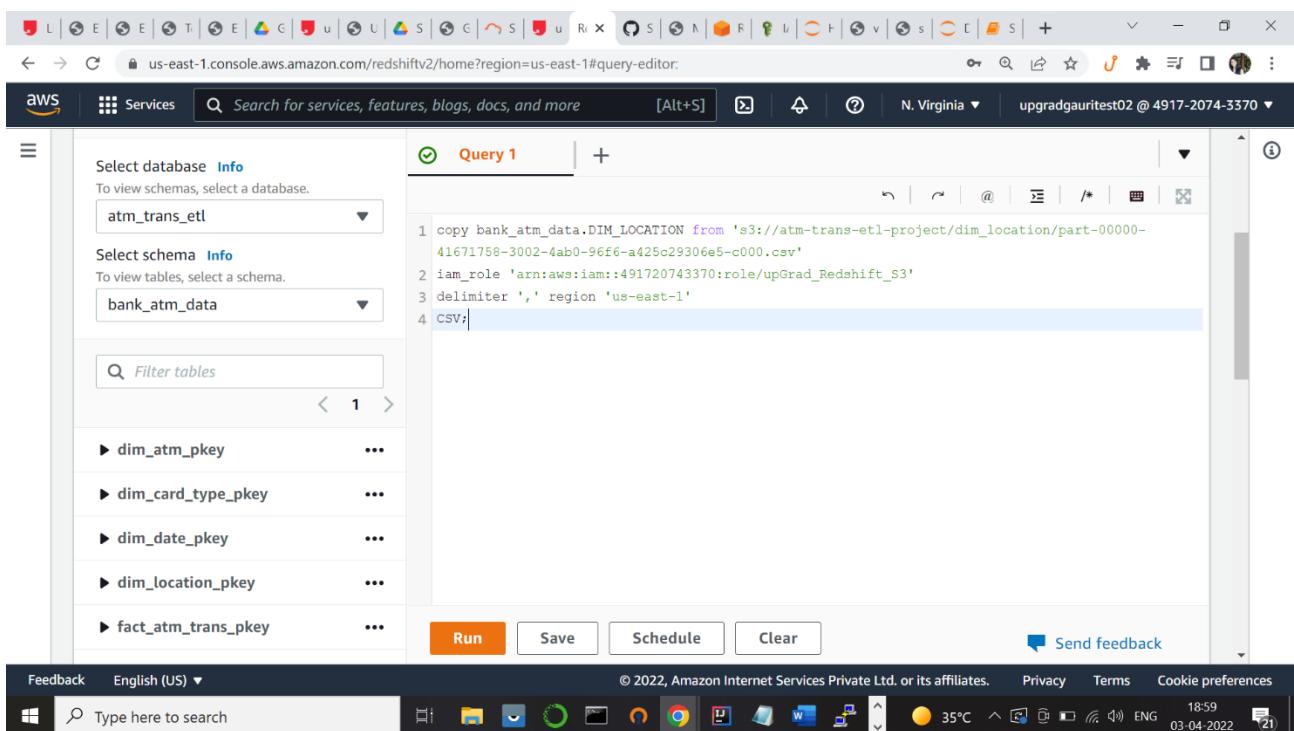
The bottom of the screen shows the Windows taskbar with icons for File Explorer, Task View, Start, Taskbar settings, and other system icons. The system tray shows the date (03-04-2022), time (18:49), battery level (21%), and temperature (37°C).

## 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. Query to load data from S3 bucket to DIM\_LOCATION:

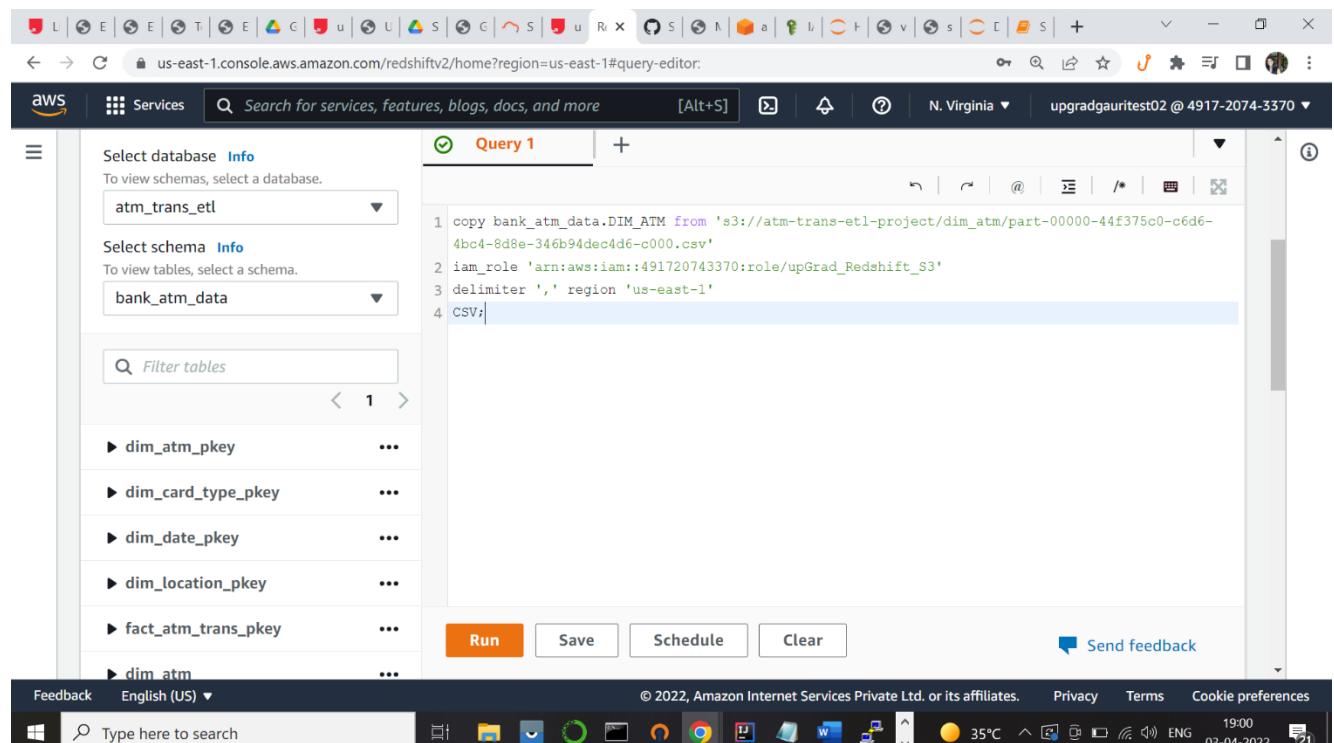
```
copy bank_atm_data.DIM_LOCATION from 's3://atm-trans-etl-project/dim_location/part-00000-41671758-3002-4ab0-96f6-a425c29306e5-c000.csv'
iam_role 'arn:aws:iam::491720743370:role/upGrad_Redshift_S3'
delimiter ',' region 'us-east-1'
CSV;
```



The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar with database and schema selection dropdowns set to 'atm\_trans\_etl' and 'bank\_atm\_data'. Below that is a table list with entries like 'dim\_atm\_pkey', 'dim\_card\_type\_pkey', etc. The main area is titled 'Query 1' and contains the four-line SQL query provided above. At the bottom of the editor are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The status bar at the bottom shows the date and time as '03-04-2022 18:59'.

## 2. Query to load data from S3 bucket to DIM\_ATM:

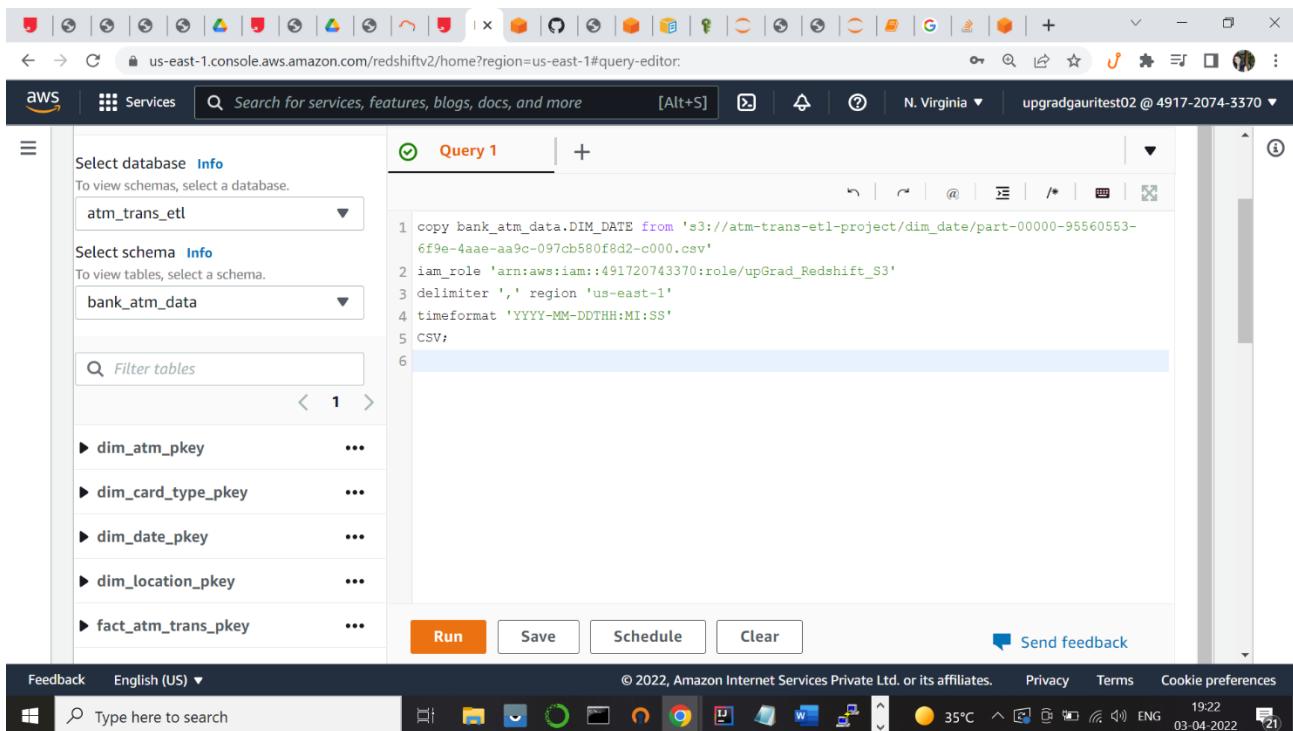
```
copy bank_atm_data.DIM_ATM from 's3://atm-trans-etl-project/dim_atm/part-00000-44f375c0-c6d6-4bc4-8d8e-346b94dec4d6-c000.csv'  
iam_role 'arn:aws:iam::491720743370:role/upGrad_Redshift_S3'  
delimiter ',' region 'us-east-1'  
CSV;
```



The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar for selecting a database ('atm\_trans\_etl') and schema ('bank\_atm\_data'). The main area is titled 'Query 1' and contains the four-line SQL copy command shown above. Below the query are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. At the bottom of the editor, there's a feedback button and links for 'Feedback', 'English (US)', 'Privacy', 'Terms', and 'Cookie preferences'. The status bar at the bottom right shows system information like temperature (35°C), time (19:00), and date (03-04-2022).

### 3. Query to load data from S3 bucket to DIM\_DATE:

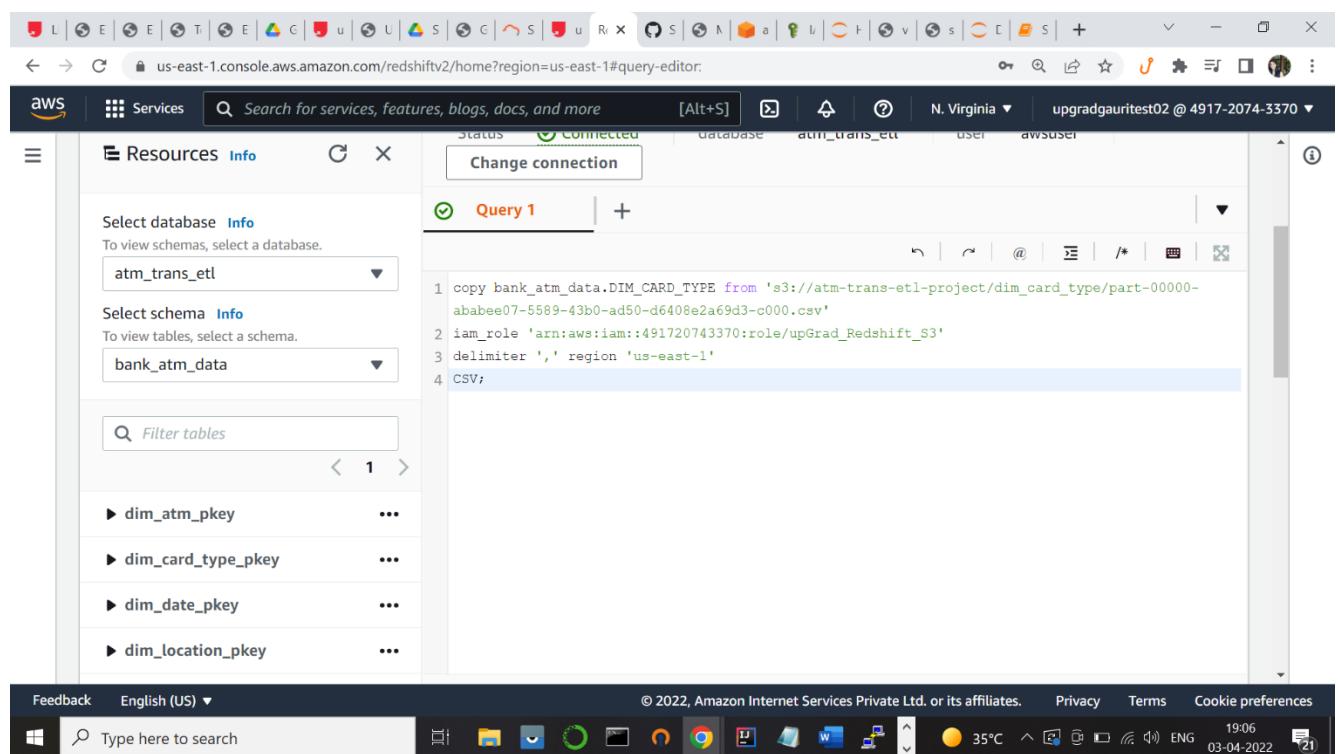
```
copy bank_atm_data.DIM_DATE from 's3://atm-trans-etl-project/dim_date/part-00000-95560553-6f9e-4aae-aa9c-097cb580f8d2-c000.csv'  
iam_role 'arn:aws:iam::491720743370:role/upGrad_Redshift_S3'  
delimiter ',' region 'us-east-1'  
timeformat 'YYYY-MM-DDTHH:MI:SS'  
CSV;
```



The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar for selecting a database ('atm\_trans\_etl') and schema ('bank\_atm\_data'). The main area is titled 'Query 1' and contains the SQL code provided in the previous block. Below the code are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. At the bottom, there's a search bar, system status icons, and a footer with copyright information and system metrics.

#### 4. Query to load data from S3 bucket to DIM\_CARD\_TYPE:

```
copy bank_atm_data.DIM_CARD_TYPE from 's3://atm-trans-etl-project/dim_card_type/part-00000-ababee07-5589-43b0-ad50-d6408e2a69d3-c000.csv'  
iam_role 'arn:aws:iam::491720743370:role/upGrad_Redshift_S3'  
delimiter ',' region 'us-east-1'  
CSV;
```



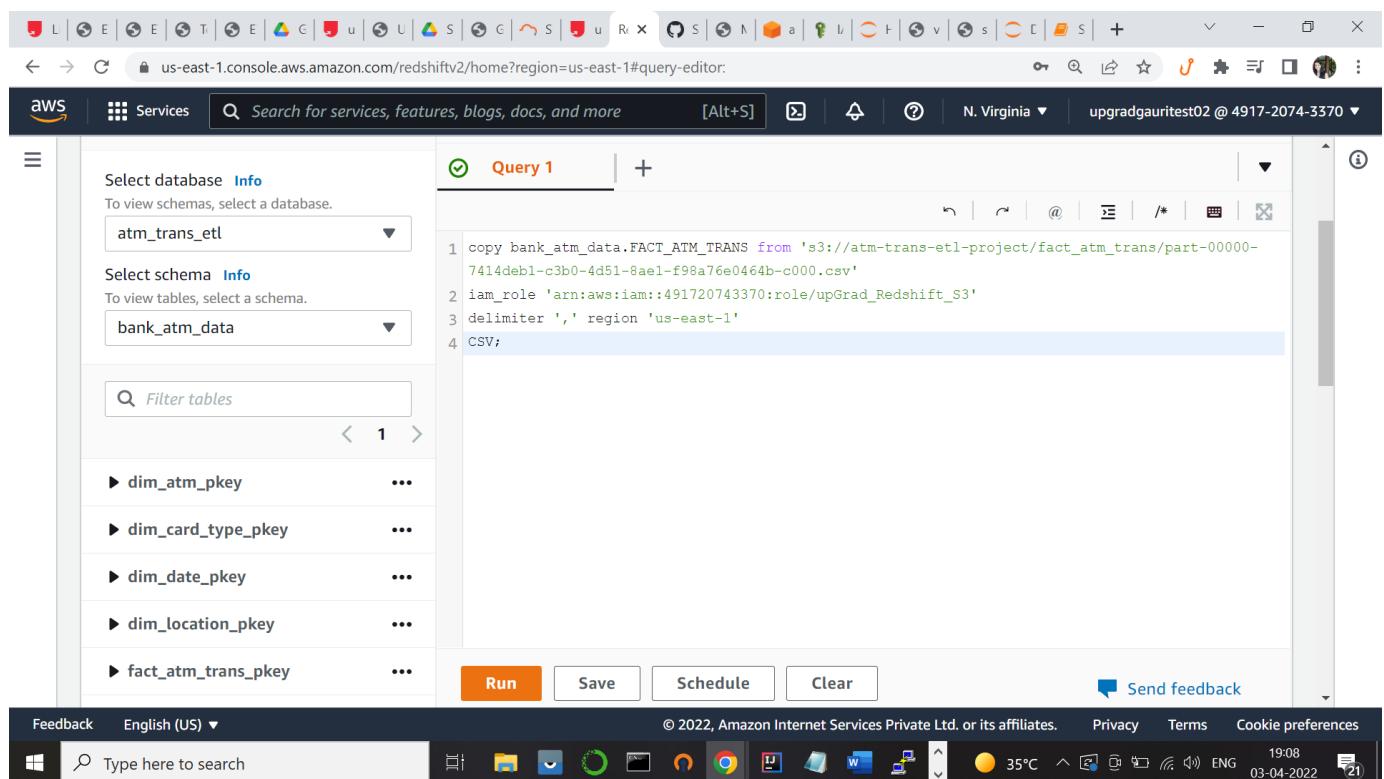
The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar for 'Resources' and 'Info' showing the selected database ('atm\_trans\_etl') and schema ('bank\_atm\_data'). The main area has a 'Query 1' tab open with the following SQL code:

```
copy bank_atm_data.DIM_CARD_TYPE from 's3://atm-trans-etl-project/dim_card_type/part-00000-ababee07-5589-43b0-ad50-d6408e2a69d3-c000.csv'  
iam_role 'arn:aws:iam::491720743370:role/upGrad_Redshift_S3'  
delimiter ',' region 'us-east-1'  
CSV;
```

The interface includes various buttons for running queries, viewing logs, and managing connections. The bottom of the screen shows the Windows taskbar with several open application icons.

## 5. Query to load data from S3 bucket to FACT\_ATM\_TRANS:

```
copy bank_atm_data.FACT_ATM_TRANS from 's3://atm-trans-etl-project/fact_atm_trans/part-00000-7414deb1-c3b0-4d51-8ae1-f98a76e0464b-c000.csv'  
iam_role 'arn:aws:iam::491720743370:role/upGrad_Redshift_S3'  
delimiter ',' region 'us-east-1'  
CSV;
```



The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar for selecting a database (atm\_trans\_etl) and schema (bank\_atm\_data). The main area is titled "Query 1" and contains the four-line SQL copy command shown above. Below the query text are buttons for "Run", "Save", "Schedule", and "Clear". At the bottom of the editor, there's a feedback section and a taskbar with various icons.