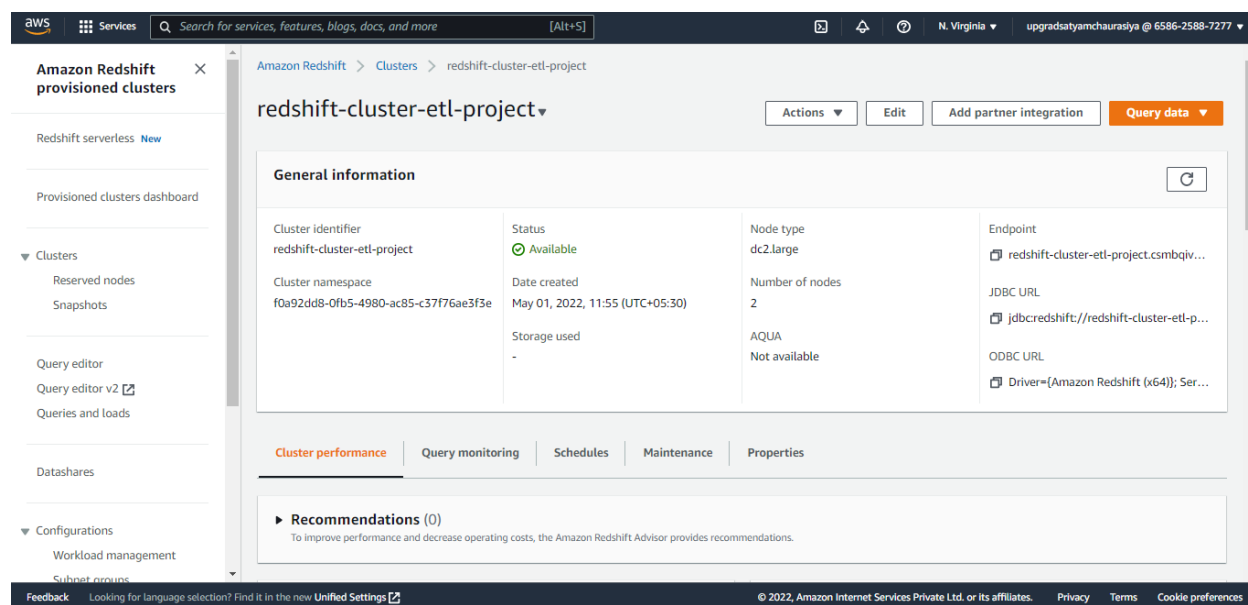
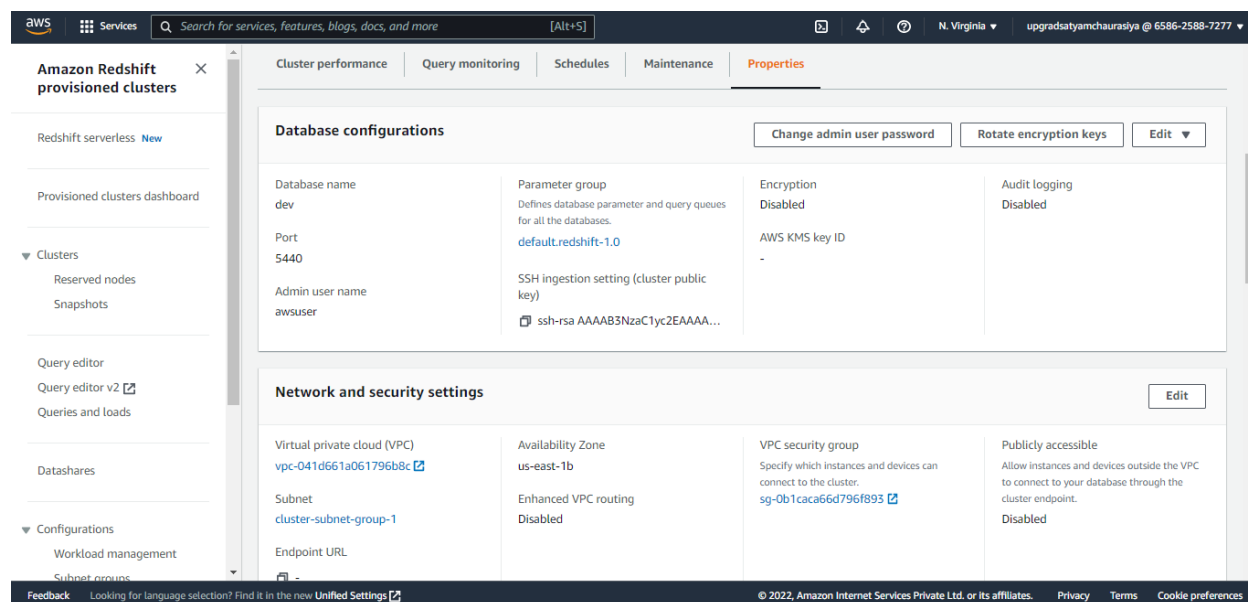


Creation of a Redshift Cluster

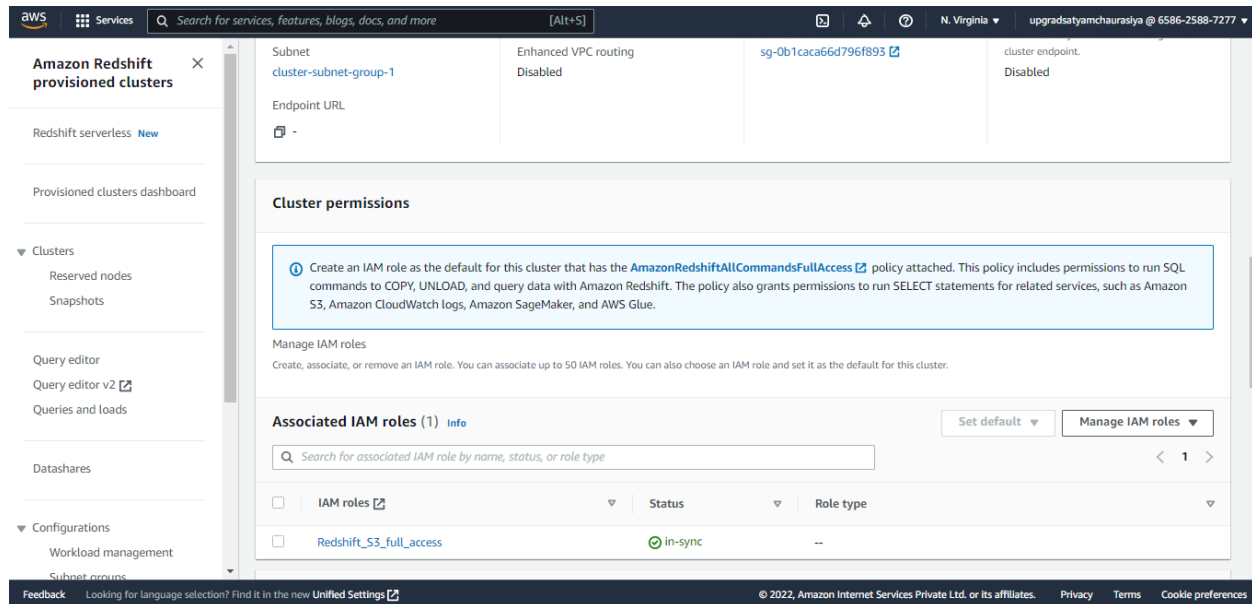
Screenshots of the configuration of the Redshift cluster that I have created:



The screenshot shows the Amazon Redshift console interface. The left sidebar contains navigation options like 'Amazon Redshift provisioned clusters', 'Redshift serverless', 'Provisioned clusters dashboard', 'Clusters', 'Reserved nodes', 'Snapshots', 'Query editor', 'Query editor v2', 'Queries and loads', 'Datashares', 'Configurations', 'Workload management', and 'Subnet groups'. The main content area displays the 'redshift-cluster-etl-project' cluster details under the 'General information' tab. The cluster is in an 'Available' status. Key details include: Cluster identifier (redshift-cluster-etl-project), Cluster namespace (f0a92dd8-0fb5-4980-ac85-c37f76ae3f3e), Node type (dc2.large), Number of nodes (2), and Storage used (-). The endpoint is redshift-cluster-etl-project.csmbqiv... and the JDBC URL is jdbc:redshift://redshift-cluster-etl-p... The ODBC URL is Driver=(Amazon Redshift (x64)); Ser... Below the general information, there are tabs for 'Cluster performance', 'Query monitoring', 'Schedules', 'Maintenance', and 'Properties'. A 'Recommendations' section is also visible, stating that the Amazon Redshift Advisor provides recommendations to improve performance and decrease operating costs.



The screenshot shows the Amazon Redshift console interface, specifically the 'Properties' tab for the 'redshift-cluster-etl-project' cluster. The left sidebar is the same as in the previous screenshot. The main content area displays the 'Database configurations' and 'Network and security settings' sections. The 'Database configurations' section includes: Database name (dev), Port (5440), Admin user name (awsuser), Parameter group (default.redshift-1.0), SSH ingestion setting (ssh-rsa AAAAB3NzaC1yc2EAAA...), Encryption (Disabled), and Audit logging (Disabled). The 'Network and security settings' section includes: Virtual private cloud (VPC) (vpc-041d661a061796b8c), Subnet (cluster-subnet-group-1), Availability Zone (us-east-1b), Enhanced VPC routing (Disabled), VPC security group (sg-0b1caca6d796f893), and Publicly accessible (Disabled). The 'Endpoint URL' is also listed. There are buttons for 'Change admin user password', 'Rotate encryption keys', and 'Edit'.

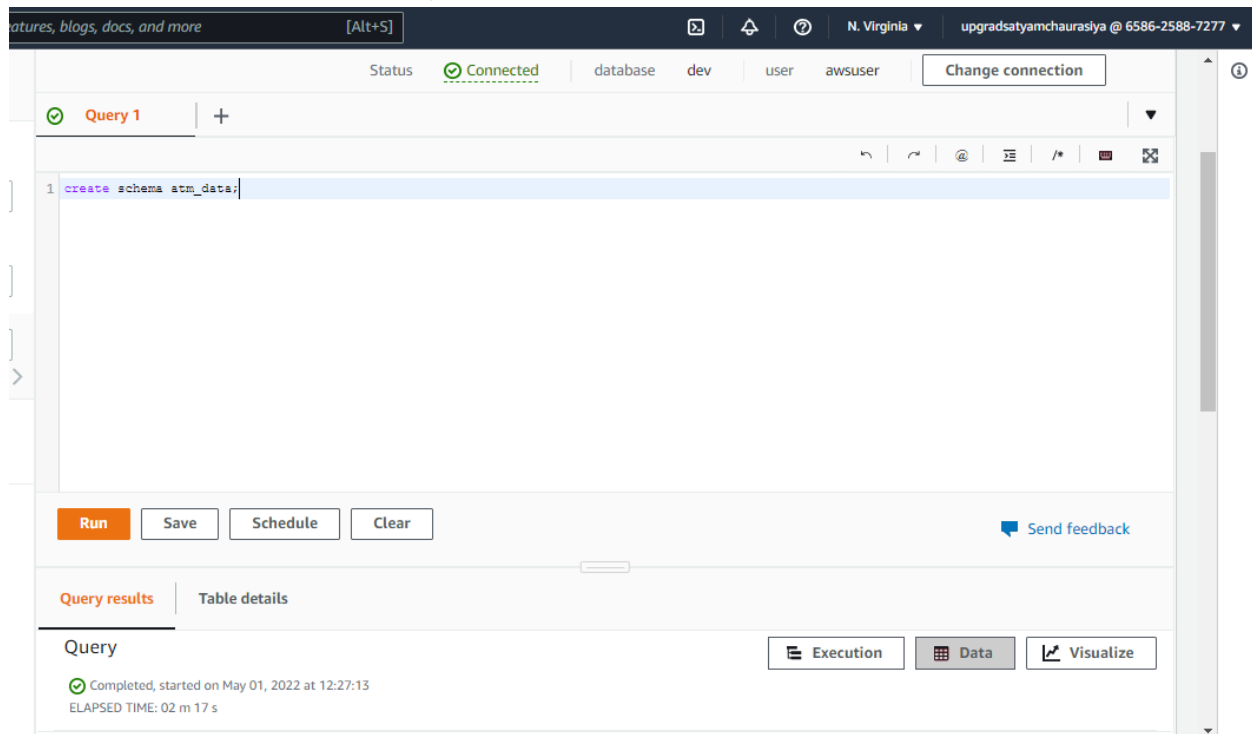


The screenshot shows the Amazon Redshift console interface. On the left, there is a navigation menu with options like 'Amazon Redshift provisioned clusters', 'Redshift serverless', 'Provisioned clusters dashboard', 'Clusters', 'Reserved nodes', 'Snapshots', 'Query editor', 'Query editor v2', 'Queries and loads', 'Databases', 'Configurations', and 'Workload management'. The main content area displays the 'Cluster permissions' section for a specific cluster. It includes a warning box about creating an IAM role with the 'AmazonRedshiftAllCommandsFullAccess' policy. Below this, there is a section for 'Associated IAM roles' with a table showing one role named 'Redshift_S3_full_access' with a status of 'in-sync'.

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

Query for creating schema:

```
create schema atm_data;
```

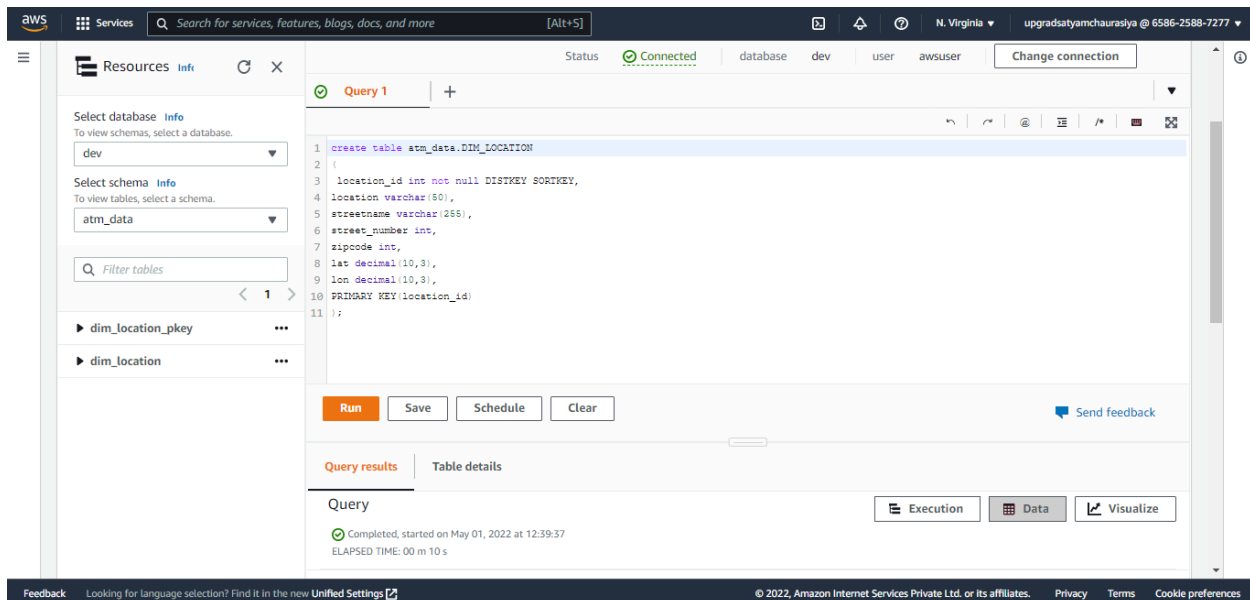


The screenshot shows the Amazon Redshift Query Editor interface. At the top, there is a status bar indicating 'Connected' and 'database: dev'. Below this, there is a text area for writing SQL queries. The query 'create schema atm_data;' is entered and highlighted. At the bottom, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. Below the query editor, there is a section for 'Query results' and 'Table details'. The 'Query results' section shows the query status as 'Completed' and the execution time as '02 m 17 s'.

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

- Creating location dimension table

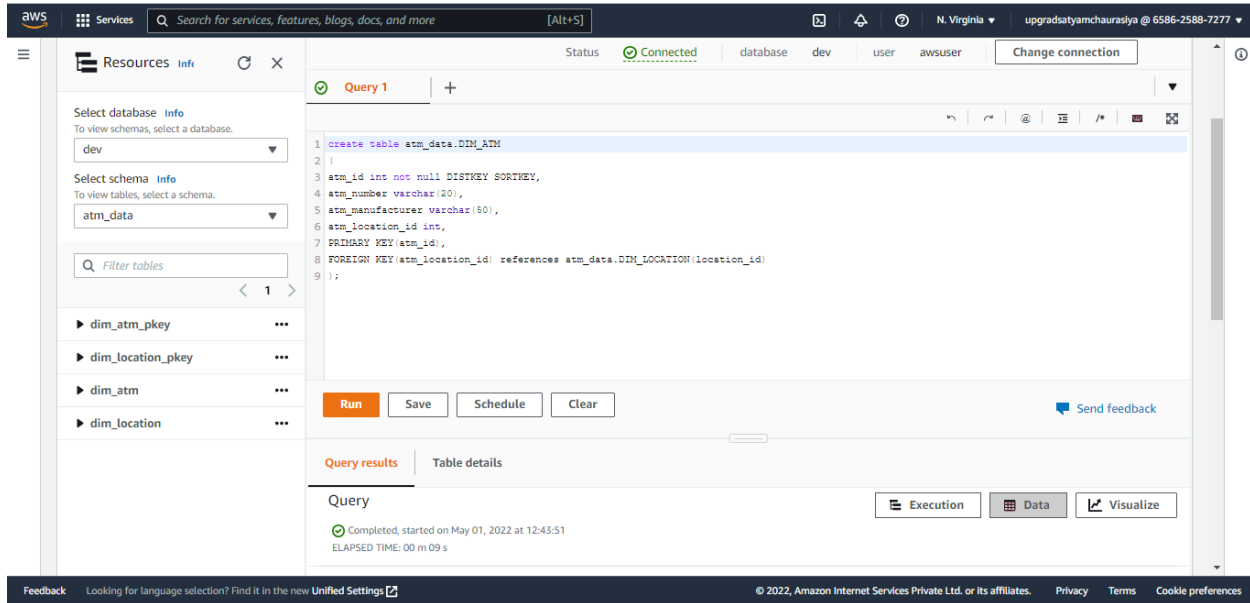
```
create table atm_data.DIM_LOCATION
(
  location_id int not null DISTKEY SORTKEY,
  location varchar(50),
  streetname varchar(255),
  street_number int,
  zipcode int,
  lat decimal(10,3),
  lon decimal(10,3),
  PRIMARY KEY(location_id)
);
```



- Creating atm dimension table

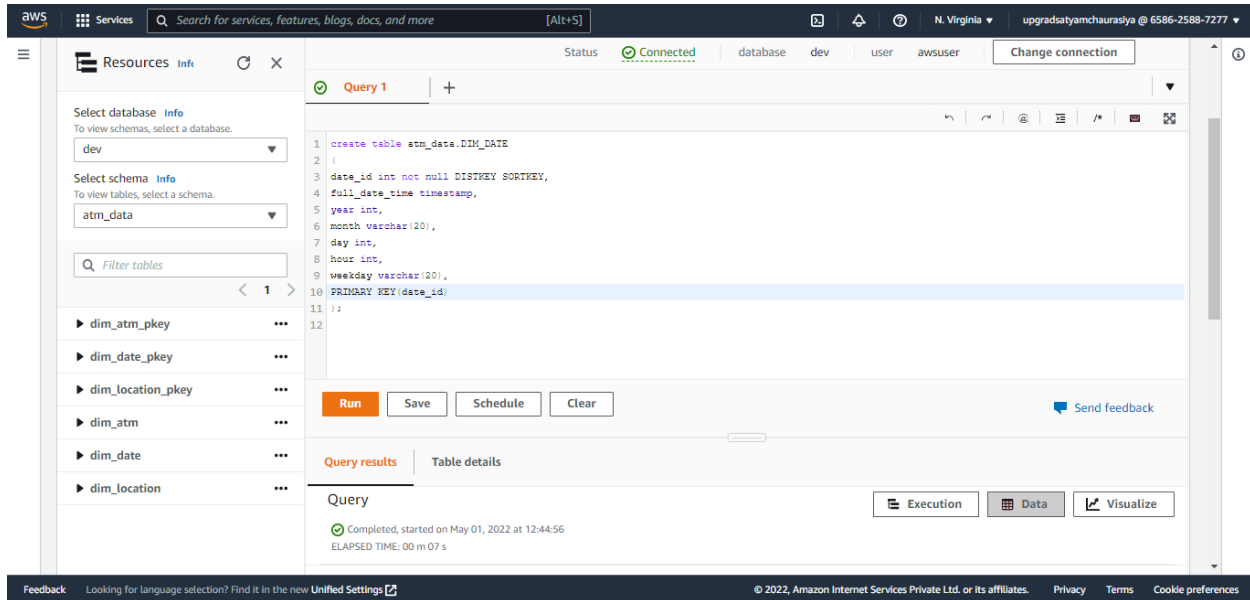
```
create table atm_data.DIM_ATM
(
  atm_id int not null DISTKEY SORTKEY,
  atm_number varchar(20),
  atm_manufacturer varchar(50),
  atm_location_id int,
```

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)
);



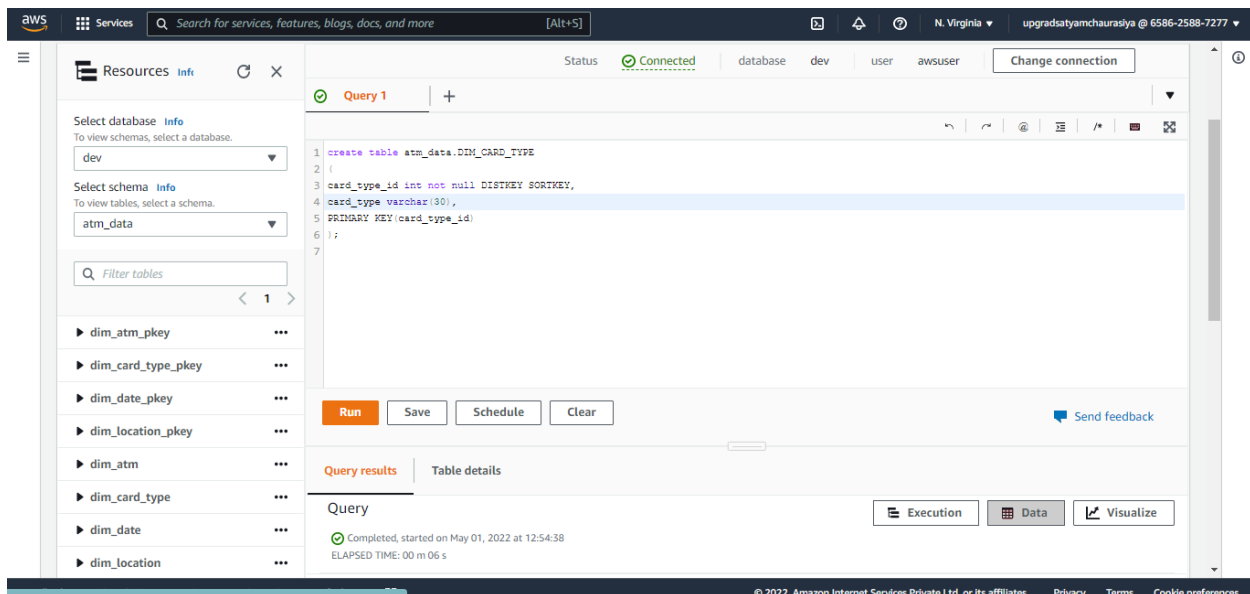
The screenshot shows the AWS Glue console interface. On the left, the 'Resources' panel is open, showing the 'dev' database and the 'atm_data' schema. The 'Filter tables' search bar is empty. Below the search bar, a list of tables is displayed, including 'dim_atm_pkey', 'dim_date_pkey', 'dim_location_pkey', 'dim_atm', 'dim_date', and 'dim_location'. The main panel shows the 'Query 1' editor with 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 'Run' button is highlighted. To the right of the buttons is a 'Send feedback' link. Below the buttons, the 'Query results' tab is selected, showing the 'Query' details. The status indicates 'Completed, started on May 01, 2022 at 12:44:56' and 'ELAPSED TIME: 00 m 07 s'. At the bottom, there are links 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)
);
```



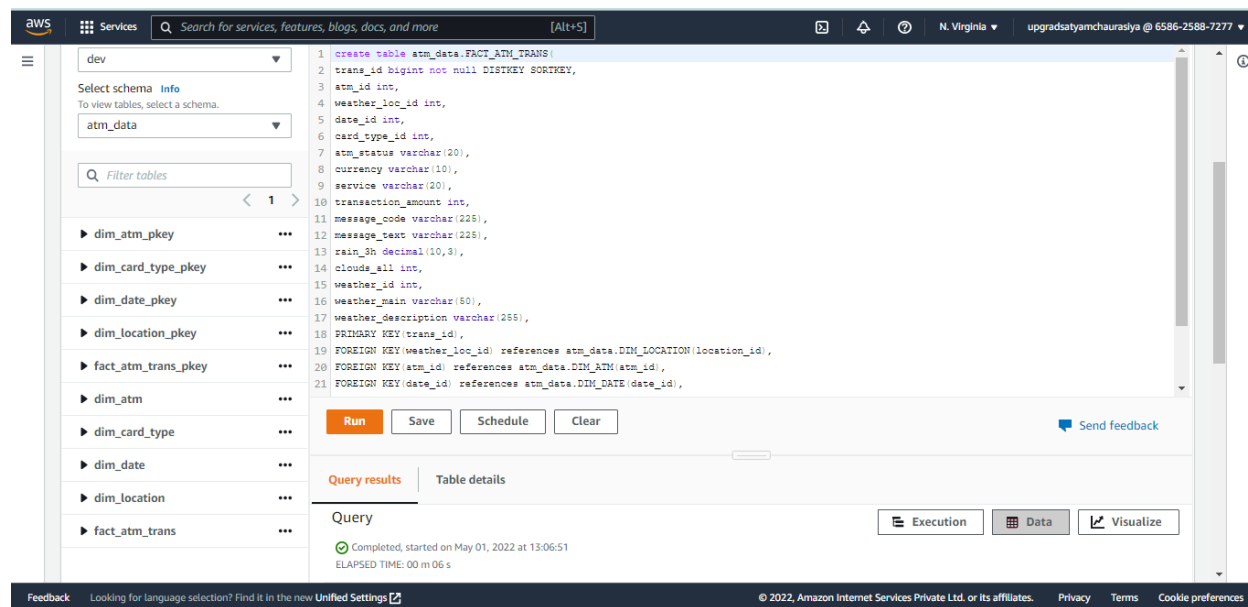
The screenshot shows the AWS Glue console interface. On the left, the 'Resources' panel is open, showing the 'dev' database and the 'atm_data' schema. The 'Filter tables' search bar is empty. Below the search bar, a list of tables is displayed, including 'dim_atm_pkey', 'dim_card_type_pkey', 'dim_date_pkey', 'dim_location_pkey', 'dim_atm', 'dim_card_type', 'dim_date', and 'dim_location'. The main panel shows the 'Query 1' editor with 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 );
7
```

Below the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Run' button is highlighted. To the right of the buttons is a 'Send feedback' link. Below the buttons, the 'Query results' tab is selected, showing the 'Query' details. The status indicates 'Completed, started on May 01, 2022 at 12:54:38' and 'ELAPSED TIME: 00 m 06 s'. At the bottom, there are links 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_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, the 'dev' environment is selected, and the 'atm_data' schema is chosen. A list of tables is visible, 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 the SQL query for creating the 'fact_atm_trans' table. The query includes columns for transaction ID, ATM ID, location ID, date ID, card type ID, status, currency, service, transaction amount, message code, message text, rain, clouds, weather ID, weather main, and weather description. It also defines a primary key on 'trans_id' and several foreign key constraints. Below the query, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' tab is active, showing a status of 'Completed, started on May 01, 2022 at 13:06:51' and an elapsed time of '00 m 06 s'. The bottom of the console shows a footer with 'Feedback', a language selection prompt, and copyright information for 2022.

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

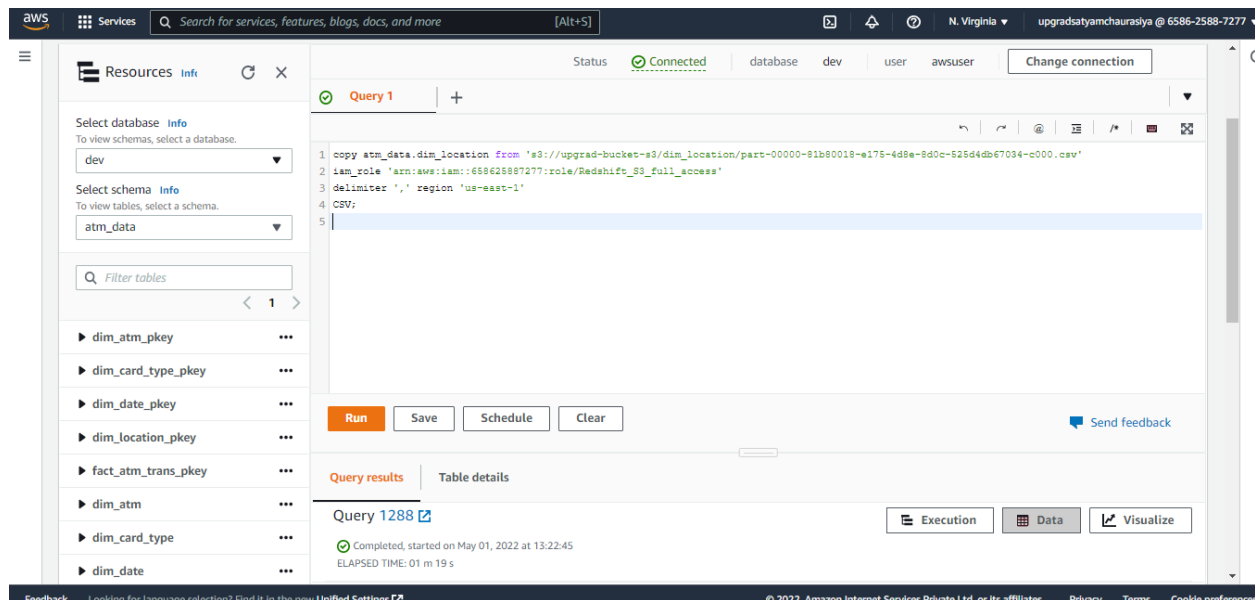
• Copying the data to dim_location table

copy atm_data.dim_location from 's3://upgrad-bucket-s3/dim_location/part-00000-81b80018-e175-4d8e-8d0c-525d4db67034-c000.csv'

iam_role 'arn:aws:iam::658625887277:role/Redshift_S3_full_access'

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

CSV;



The screenshot shows the AWS Redshift console interface. On the left, there's a sidebar with 'Resources' and a list of tables in 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', and 'dim_date'. The main area displays a SQL query for 'Query 1':

```
1 copy atm_data.dim_location from 's3://upgrad-bucket-s3/dim_location/part-00000-81b80018-e175-4d8e-8d0c-525d4db67034-c000.csv'
2 iam_role 'arn:aws:iam::658625887277:role/Redshift_S3_full_access'
3 delimiter ',' region 'us-east-1'
4 CSV;
5
```

Below the query, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Run' button is highlighted. Below these buttons, the 'Query results' section shows 'Query 1288' with a status of 'Completed, started on May 01, 2022 at 13:22:45' and an 'ELAPSED TIME: 01 m 19 s'. There are also buttons for 'Execution', 'Data', and 'Visualize'.

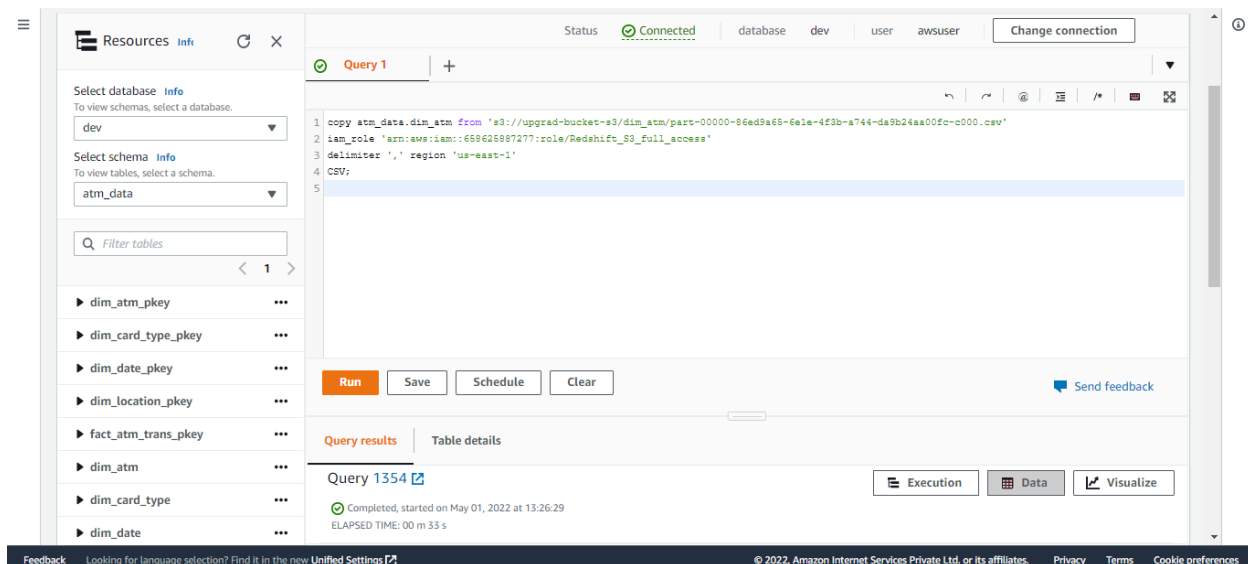
• Copying the data to dim_atm table

copy atm_data.dim_atm from 's3://upgrad-bucket-s3/dim_atm/part-00000-86ed9a65-6e1e-4f3b-a744-da9b24aa00fc-c000.csv'

iam_role 'arn:aws:iam::658625887277:role/Redshift_S3_full_access'

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

CSV;



The screenshot shows the AWS Redshift console interface. On the left, the 'Resources' sidebar is open, showing the 'dev' database and 'atm_data' schema. The main panel displays a SQL query for 'Query 1':

```
1 copy atm_data.dim_atm from 's3://upgrad-bucket-s3/dim_atm/part-00000-86ed9a65-6e1e-4f3b-a744-da9b24aa00fc-c000.csv'
2 iam_role 'arn:aws:iam::658625887277:role/Redshift_S3_full_access'
3 delimiter ',' region 'us-east-1'
4 CSV;
5
```

Below the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' tab is active, showing 'Query 1354' with a status of 'Completed, started on May 01, 2022 at 13:26:29' and an 'ELAPSED TIME: 00 m 33 s'.

• Copying the data to dim_date table

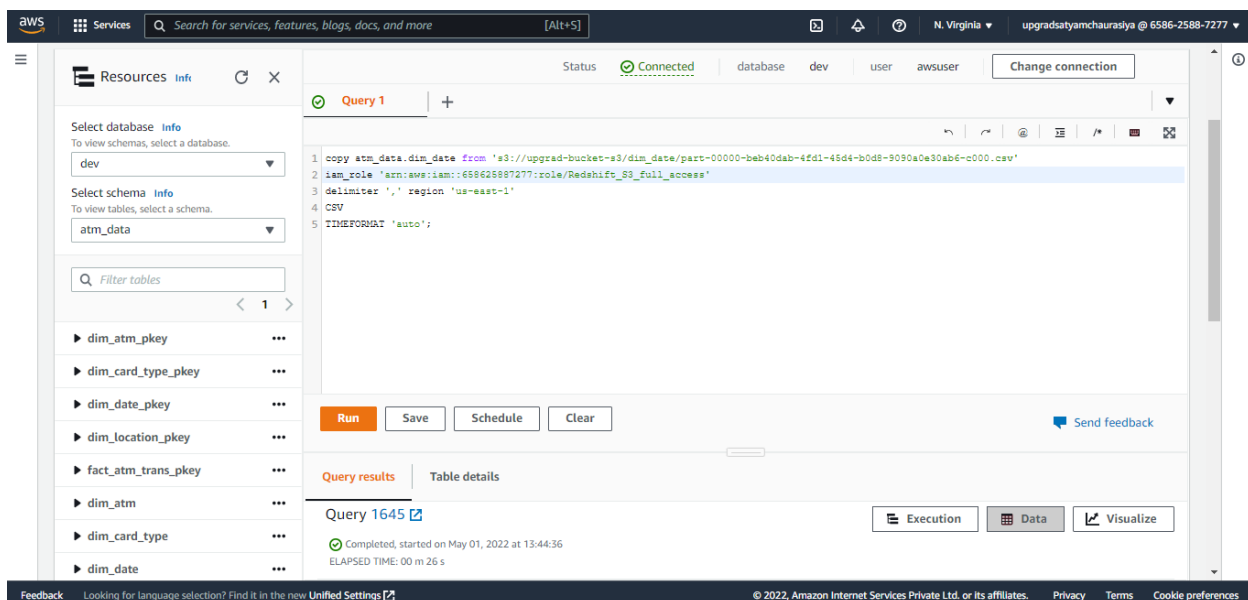
copy atm_data.dim_date from 's3://upgrad-bucket-s3/dim_date/part-00000-beb40dab-4fd1-45d4-b0d8-9090a0e30ab6-c000.csv'

iam_role 'arn:aws:iam::658625887277:role/Redshift_S3_full_access'

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

CSV

TIMEFORMAT 'auto';



The screenshot shows the AWS Redshift console interface. On the left, the 'Resources' sidebar is open, showing the 'dev' database and 'atm_data' schema. The main panel displays a SQL query for 'Query 1':

```
1 copy atm_data.dim_date from 's3://upgrad-bucket-s3/dim_date/part-00000-beb40dab-4fd1-45d4-b0d8-9090a0e30ab6-c000.csv'
2 iam_role 'arn:aws:iam::658625887277:role/Redshift_S3_full_access'
3 delimiter ',' region 'us-east-1'
4 CSV
5 TIMEFORMAT 'auto';
```

Below the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' tab is active, showing 'Query 1645' with a status of 'Completed, started on May 01, 2022 at 13:44:36' and an 'ELAPSED TIME: 00 m 26 s'.

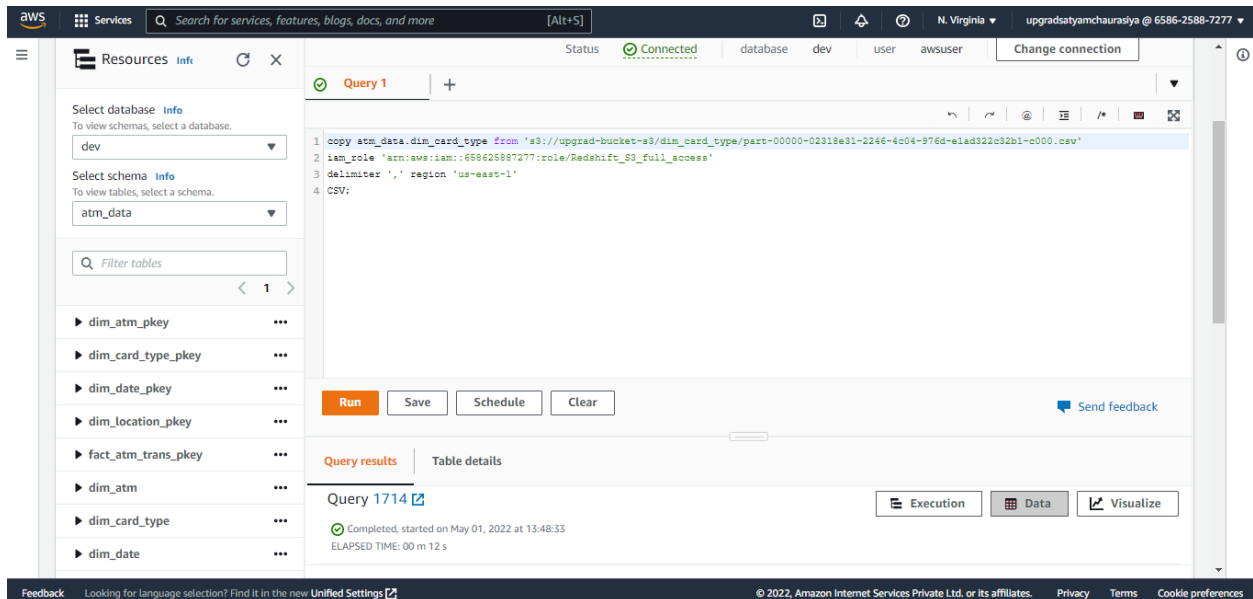
- Copying the data to dim_card_type table

copy atm_data.dim_card_type from 's3://upgrad-bucket-s3/dim_card_type/part-00000-02318e31-2246-4c04-976d-e1ad322c32b1-c000.csv'

iam_role 'arn:aws:iam::658625887277:role/Redshift_S3_full_access'

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

CSV;



- Copying the data to fact_atm_trans table

copy atm_data.fact_atm_trans from 's3://upgrad-bucket-s3/fact_atm_trans/part-00000-fb72e768-f589-4caa-bd26-1bc6fcbbdb8a-c000.csv'

iam_role 'arn:aws:iam::658625887277:role/Redshift_S3_full_access'

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

CSV;

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