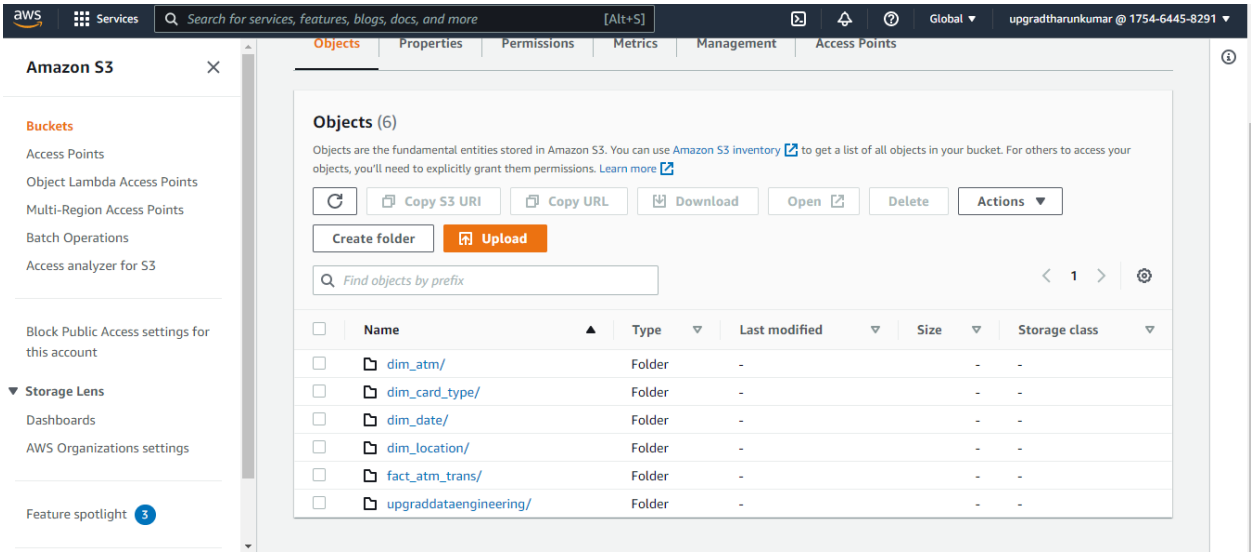
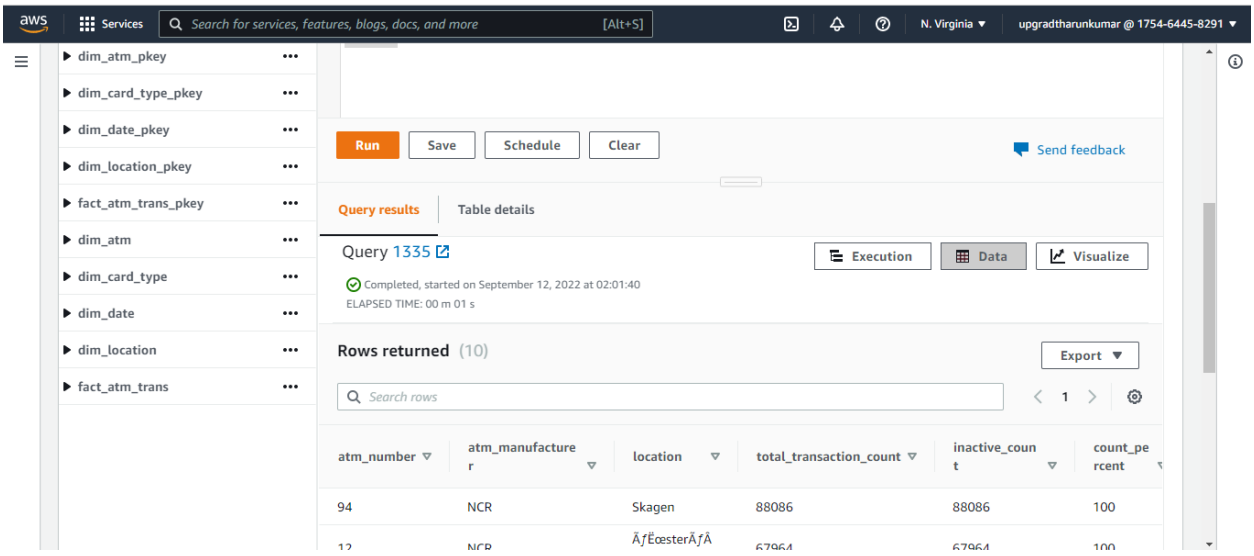
From Pyspark data from csv file made into a tabular form with a schema, then they are exported to the s3 buckets(as shown in the image). Data is stored in the form of CSV files.

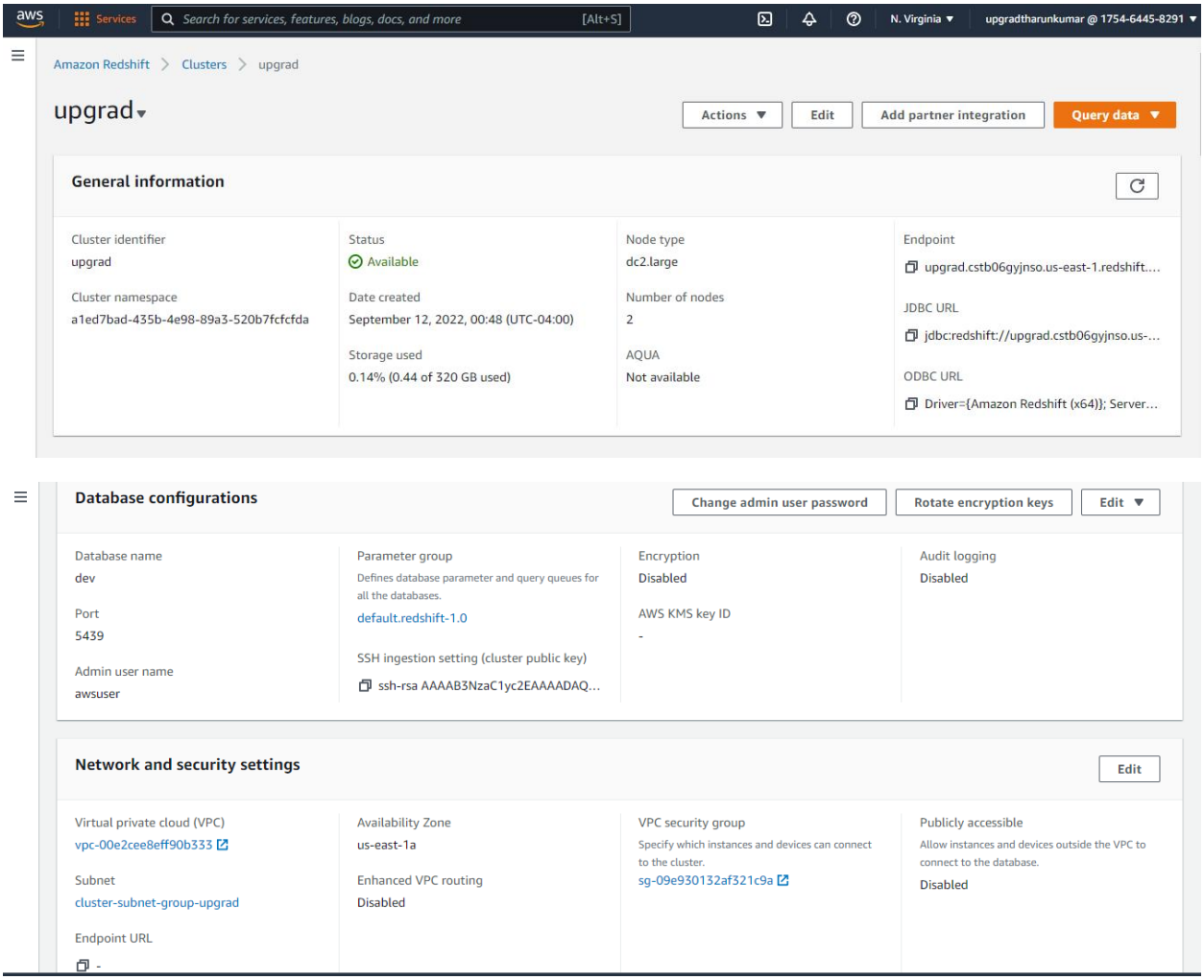


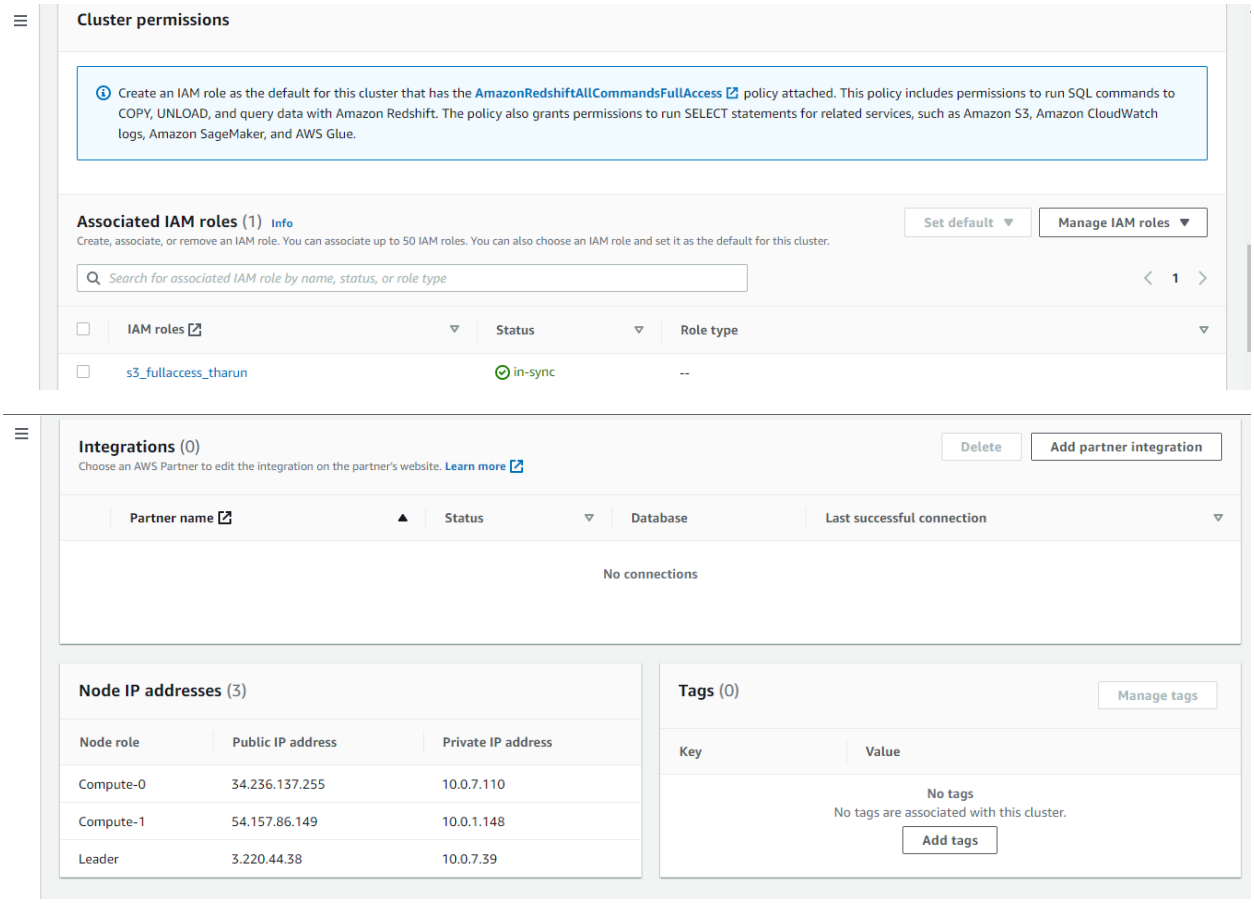

From S3 buckets data is transferred to amazon redshift data base called dev which is configured while launching amazon redshift.



Creation of a Redshift Cluster

These following images are the screenshots of the configurations made while we launch the amazon redshift cluster named as upgrad.





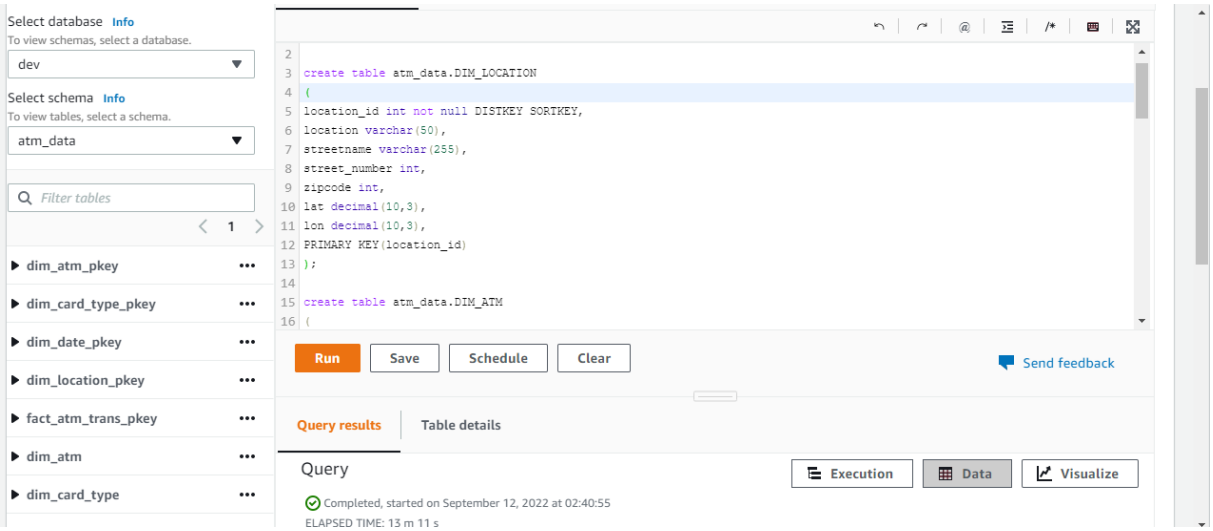




## Setting up a database in the Redshift cluster and running queries to create the dimension and fact tables

I created a atm_data schema and then I made the table using the following sql queries

create schema atm_data;




```
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)
);
```

```
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)
);
```

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)

);

```
15 create table atm_data.DIM_ATM
16 (
17 atm_id int not null DISTKEY SORTKEY,
18 atm_number varchar(20),
19 atm_manufacturer varchar(50),
20 atm_location_id int,
21 PRIMARY KEY(atm_id),
22 FOREIGN KEY(atm_location_id) references atm_data.DIM_LOCATION(location_id)
23 );
```

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)

);

```
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)
);

```

create table atm_data.DIM_CARD_TYPE

(

card_type_id int not null DISTKEY SORTKEY,

card_type varchar(30),

PRIMARY KEY(card_type_id)

);

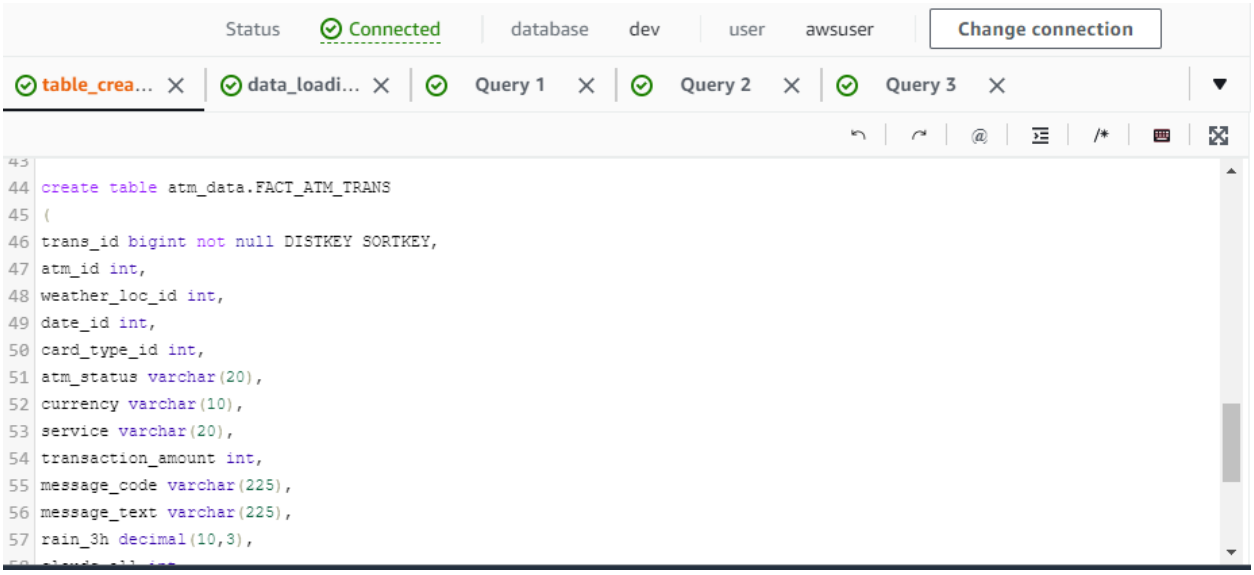
```
create table atm_data.DIM_CARD_TYPE
(
card_type_id int not null DISTKEY SORTKEY,
card_type varchar(30),
PRIMARY KEY(card_type_id)
);

```

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**

These are the queries which I made to load the data into tables that I created as per above code.

- To copy data from S3 to dim_location table

copy atm_data.dim_location from 's3://dataengineeringtharun/dim_location/part-00000-1aee43a6-4f2e-45ef-b2d3-c17ea16d6dd0-c000.csv'

iam_role 'arn:aws:iam::175464458291:role/s3_fullaccess_tharun'

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

CSV;

- To copy data from S3 to dim_atm table

copy atm_data.dim_atm from 's3://dataengineeringtharun/dim_atm/part-00000-9885cf6d-75c6-45cc-acfc-a2e770487f9f-c000.csv'

iam_role 'arn:aws:iam::175464458291:role/s3_fullaccess_tharun'

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

CSV;

- To copy data from S3 to dim_date table

copy atm_data.dim_date from 's3://dataengineeringtharun/dim_date/part-00000-ee8df625-8c88-46ed-b3d0-038d80353f57-c000.csv'

iam_role 'arn:aws:iam::175464458291:role/s3_fullaccess_tharun'

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

CSV

TIMEFORMAT 'auto';

- To copy data from S3 to dim_card table

copy atm_data.dim_card_type from 's3://dataengineeringtharun/dim_card_type/part-00000-a527b9e5-3253-4561-9d04-8689045fde03-c000.csv'

iam_role 'arn:aws:iam::175464458291:role/s3_fullaccess_tharun'

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

CSV;

- To copy data from S3 to fact table

copy atm_data.fact_atm_trans from 's3://dataengineeringtharun/fact_atm_trans/part-00000-f51a20a7-c4bc-408e-9f7f-961d8ebea103-c000.csv'

iam_role 'arn:aws:iam::175464458291:role/s3_fullaccess_tharun'

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

CSV;