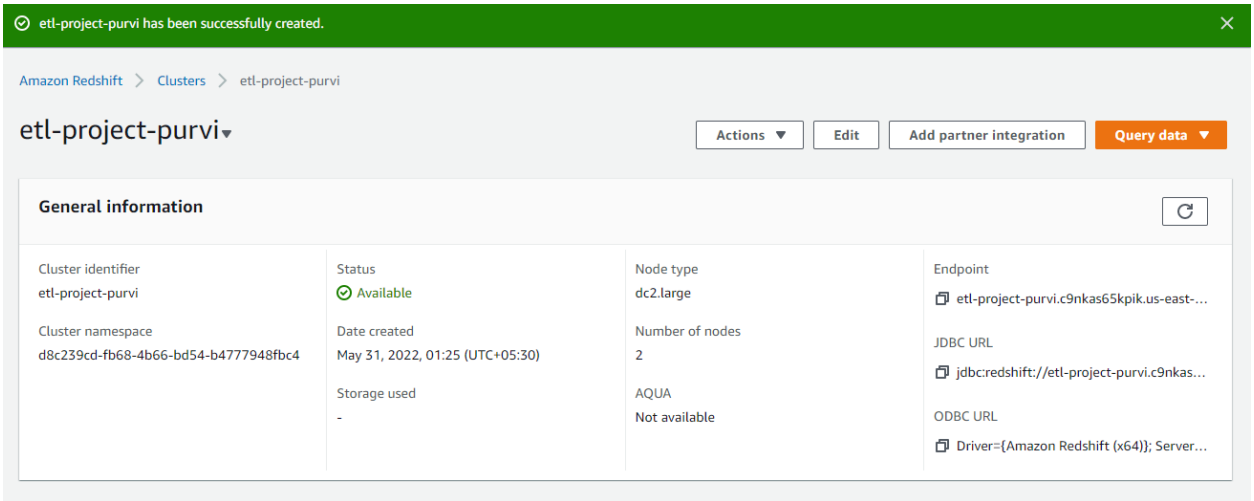
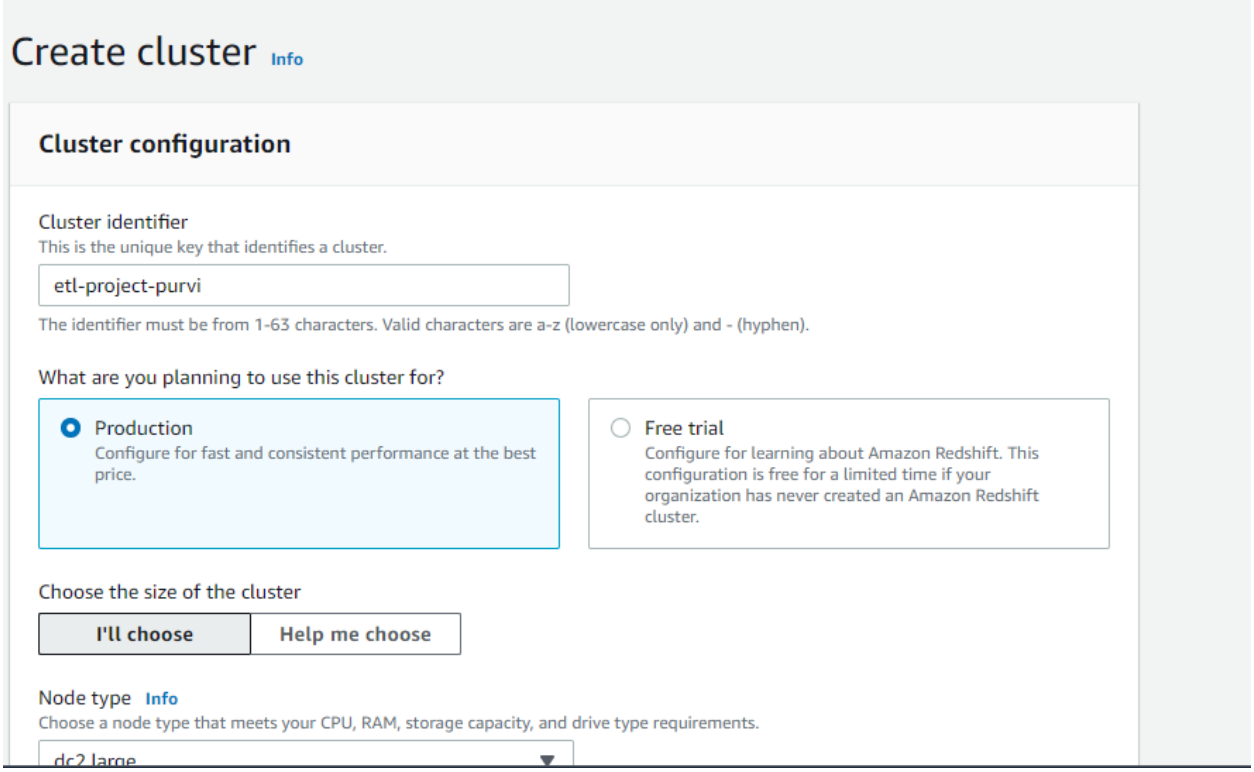
Creation of a Redshift Cluster

Screenshots of the configuration of the Redshift cluster that you have created:
The cluster created is etl-project-purvi and there are two nodes used



The configurations used are –



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

Number of 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 by purchasing reserved nodes. Learn more

320 GB

Total compressed storage

The total storage capacity for the cluster if you deploy the number of nodes that you chose.

Enter a login ID for the admin user of your DB instance.

awsuser

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

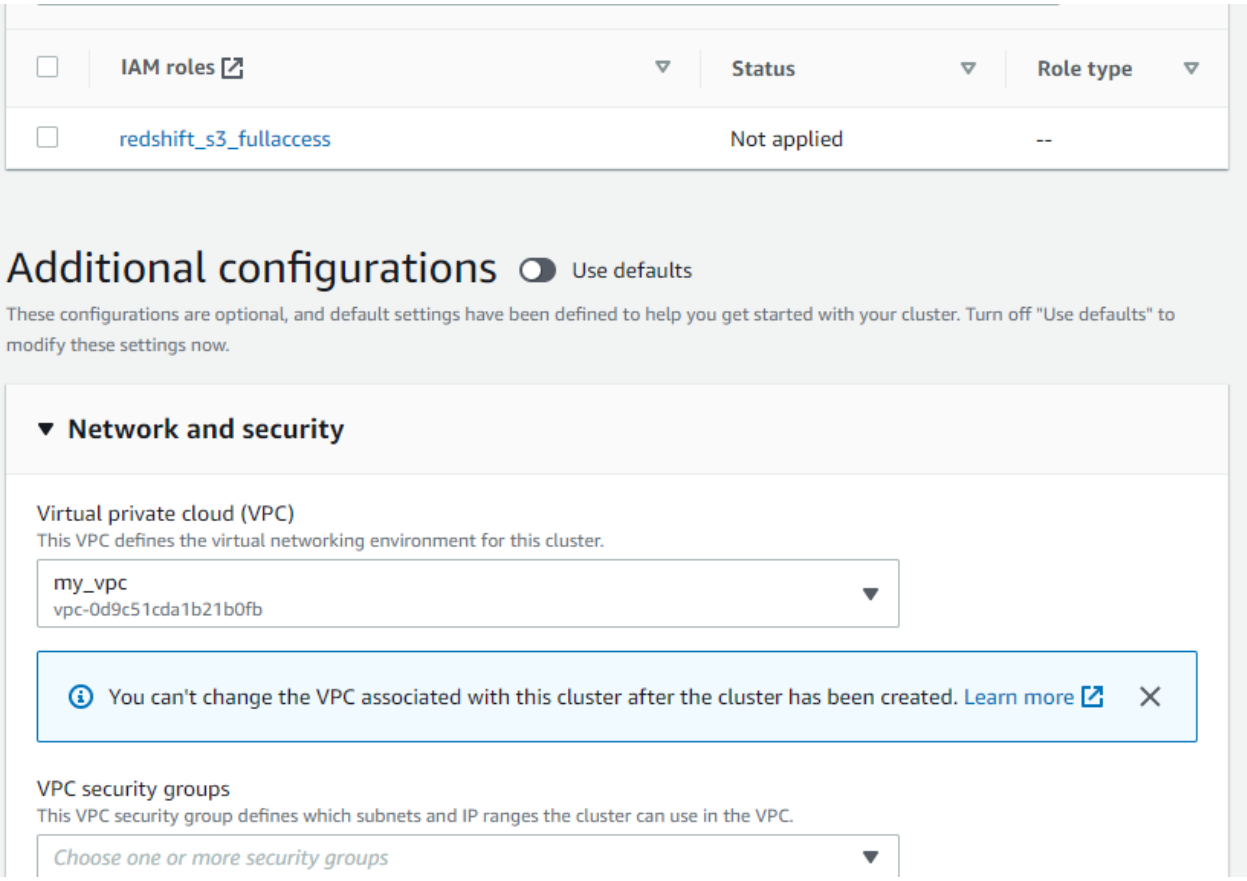
Must be 8-64 characters long. Must contain at least one uppercase letter, one lowercase letter and one number. Can be any printable ASCII character except "/", """, or "@".

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 with Amazon Redshift. The policy also grants permissions to run SELECT statements for related services, such as Amazon S3, Amazon CloudWatch logs, Amazon SageMaker, and AWS Glue.

Manage IAM roles

Create, associate, or remove an IAM role. You can associate up to 50 IAM roles. You can also choose an IAM role and set it as the default for this cluster.



VPC security groups

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

Choose one or more security groups

default
sg-0226bd8f56b333210

Cluster subnet group

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

cluster-subnet-group-1

Availability Zone

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

No preference

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.

etlproject

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.

5550

The port must be numeric (1150-65535).

Parameter groups
Defines database parameter and query queues for all the databases.

default.redshift-1.0
Default parameter group for redshift-1.0

Encryption
Encrypt all data on your cluster.

- (•) Disabled
- ( ) Use AWS Key Management Service (AWS KMS)
- ( ) Use a hardware security module (HSM)

Automated snapshot retention period
Specify how many days to retain automated snapshots.

1

The retention period must be 0-35 days. Choose zero (0) to not create automated snapshots.

Manual snapshot retention period
Specify how long do you want to retain your snapshot.

Indefinitely

The retention period must be 1-3653 days.

Configure cross-region snapshot

- (•) Disabled
- ( ) Enabled

Cluster relocation
Enable the ability to relocate your cluster in another Availability Zone. After you enable relocation, you use the VPC endpoint of the cluster to determine the cluster IP address, instead of the leader node IP address. You can find the VPC endpoint in the Network and security section of the cluster details page. Learn more

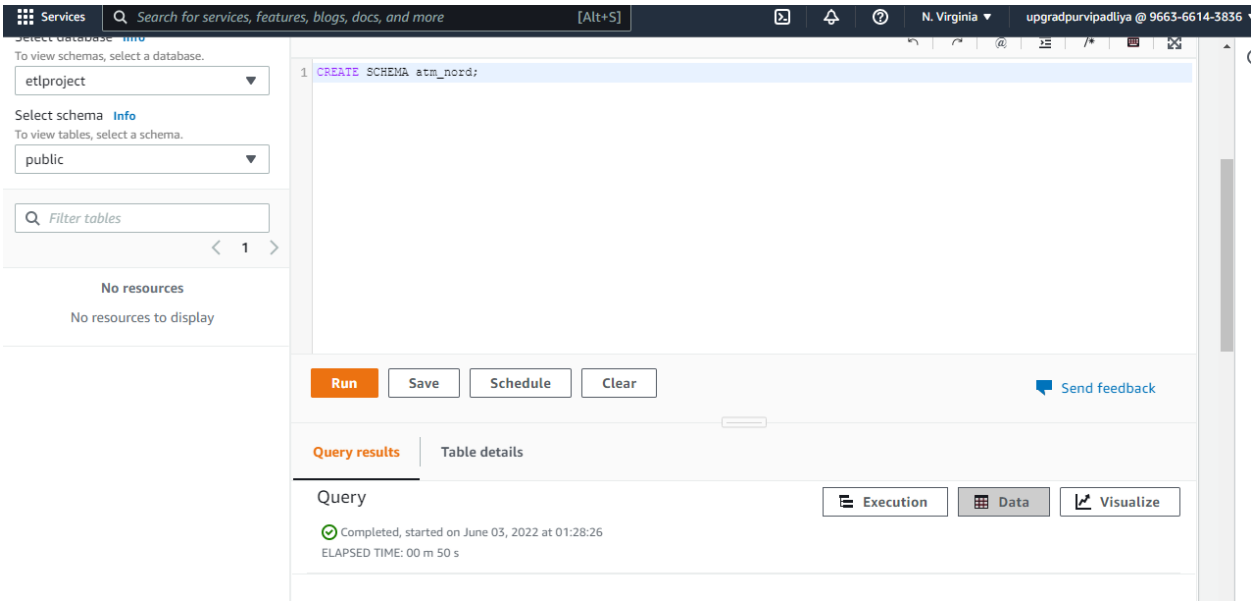
- (•) No
- ( ) Enable

Cancel | Create cluster

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

Query to create the database's schema:

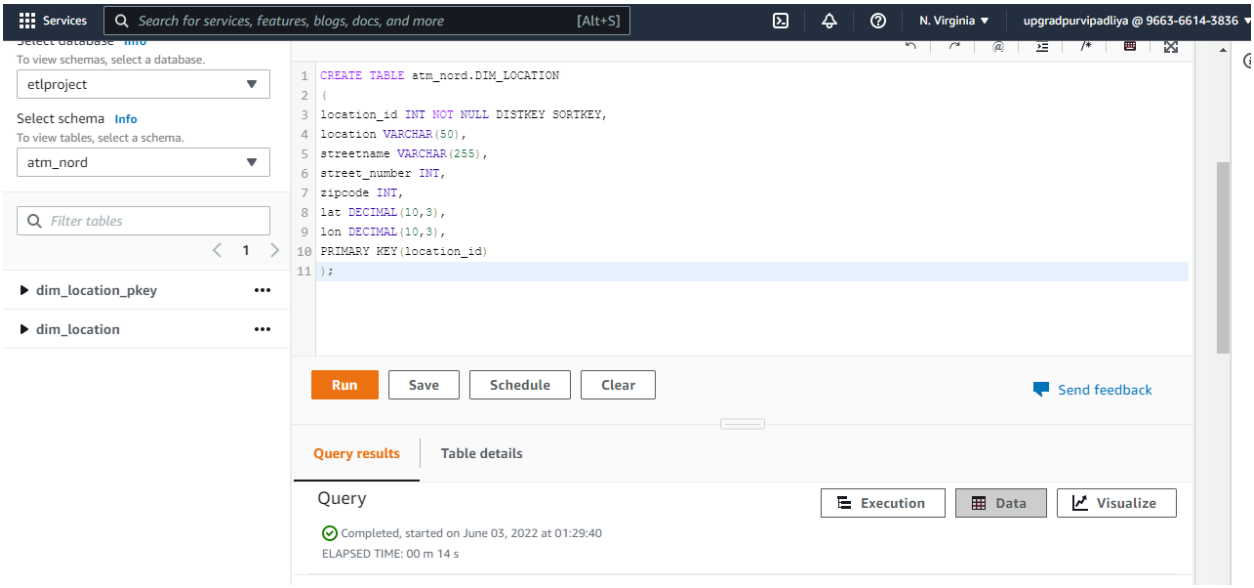
CREATE SCHEMA atm_nord;



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

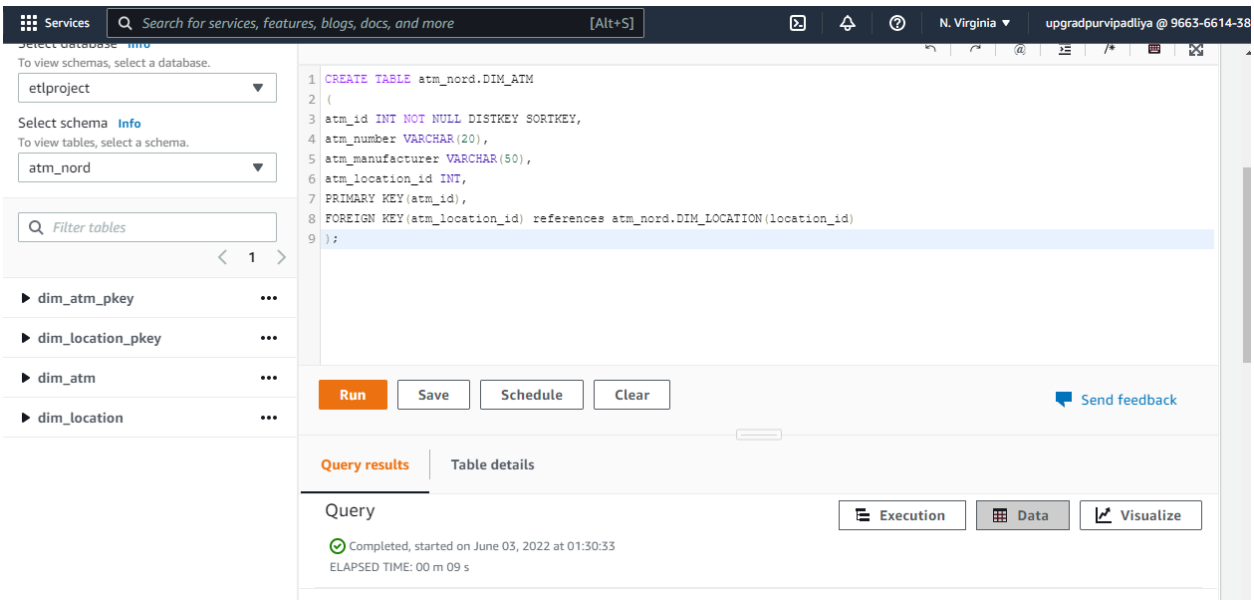
Query to create the dimension table DIM_LOCATION:

```
CREATE TABLE atm_nord.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)
);
```



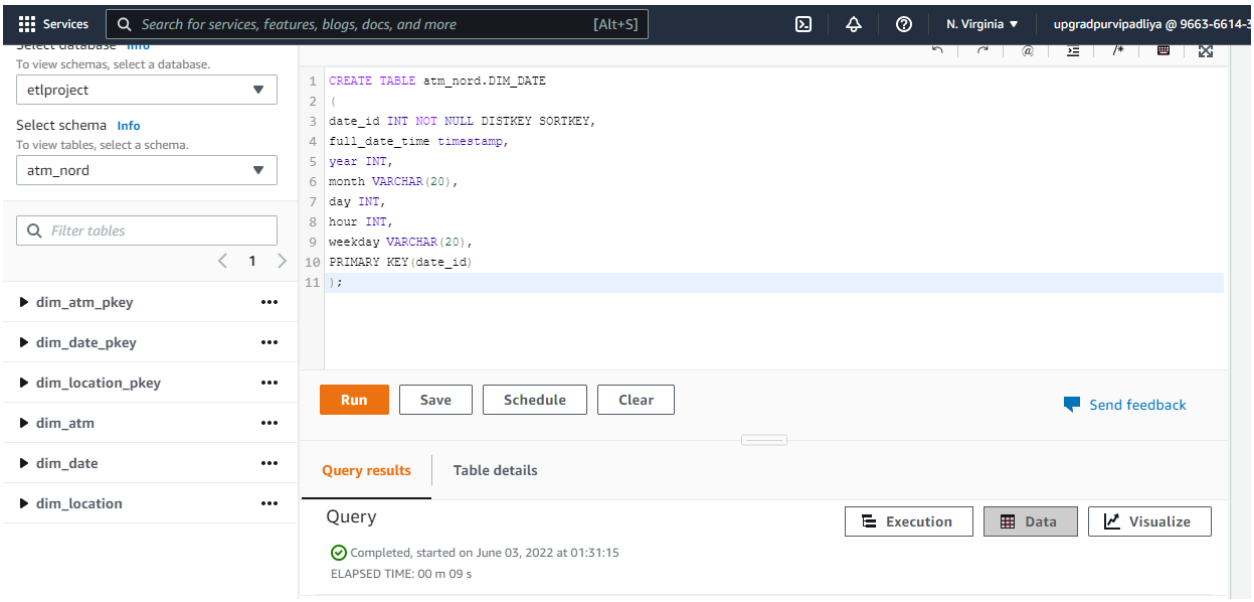
Query to create the dimension table DIM_ATM:

```
CREATE TABLE atm_nord.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_nord.DIM_LOCATION(location_id)
);
```



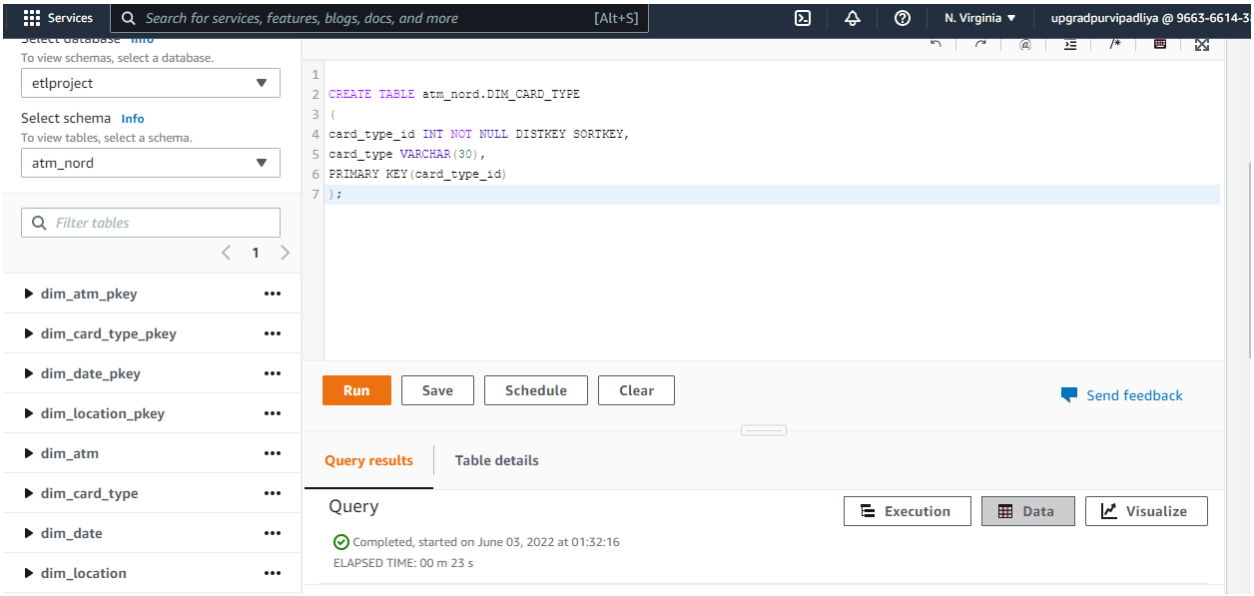
Query to create the dimension table DIM_DATE:

```
CREATE TABLE atm_nord.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)
);
```



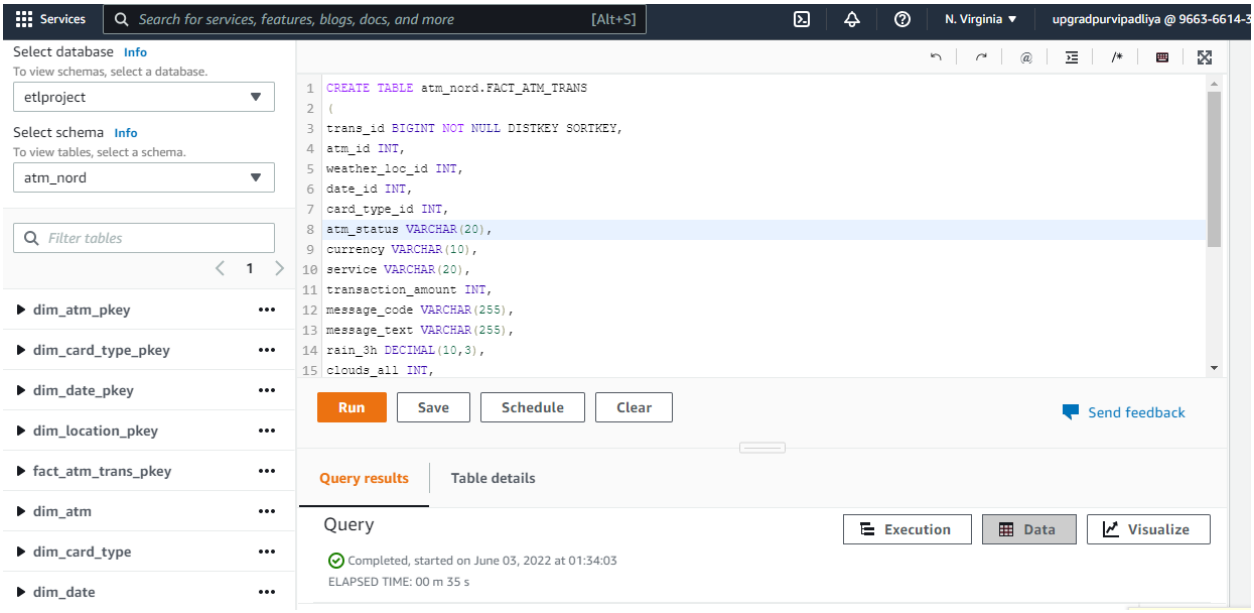
Query to create the dimension table DIM_CARD_TYPE:

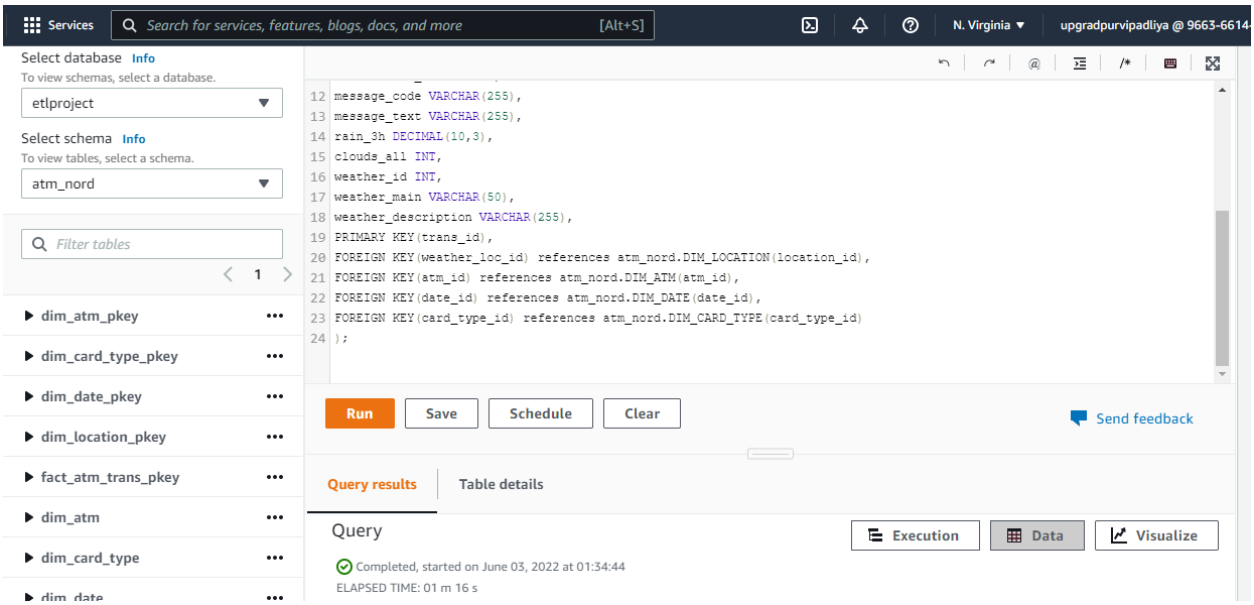
```
CREATE TABLE atm_nord.DIM_CARD_TYPE
(
card_type_id INT NOT NULL DISTKEY SORTKEY,
card_type VARCHAR(30),
PRIMARY KEY(card_type_id)
);
```



Query to create the fact table FACT_ATM_TRANS:

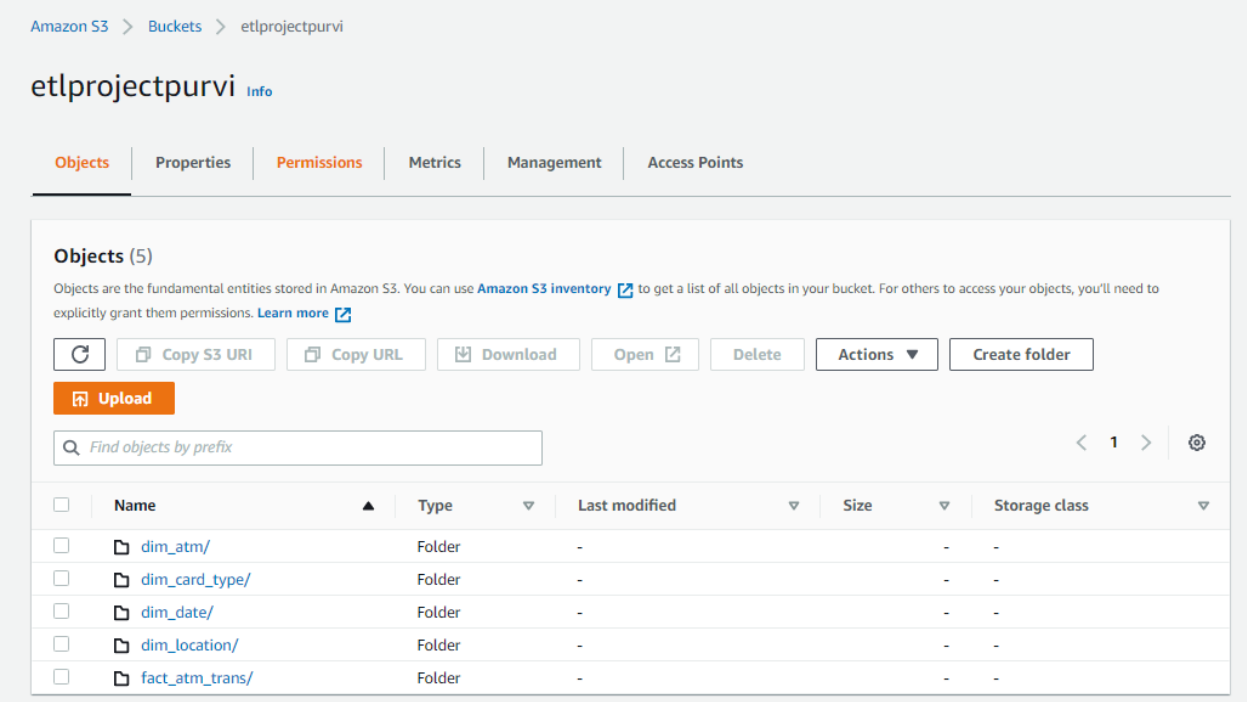
```
CREATE TABLE atm_nord.FACT_ATM_TRANS
(
trans_id BIGINT NOT NULL DISTKEY SORTKEY,
atm_id INT,
weather_loc_id INT,
date_id INT,
card_type_id INT,
atm_status VARCHAR(20),
currency VARCHAR(10),
service VARCHAR(20),
transaction_amount INT,
message_code VARCHAR(255),
message_text VARCHAR(255),
rain_3h 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 atm_nord.DIM_LOCATION(location_id),
FOREIGN KEY(atm_id) references atm_nord.DIM_ATM(atm_id),
FOREIGN KEY(date_id) references atm_nord.DIM_DATE(date_id),
FOREIGN KEY(card_type_id) references atm_nord.DIM_CARD_TYPE(card_type_id)
);
```





Loading data into a Redshift cluster from Amazon S3 bucket

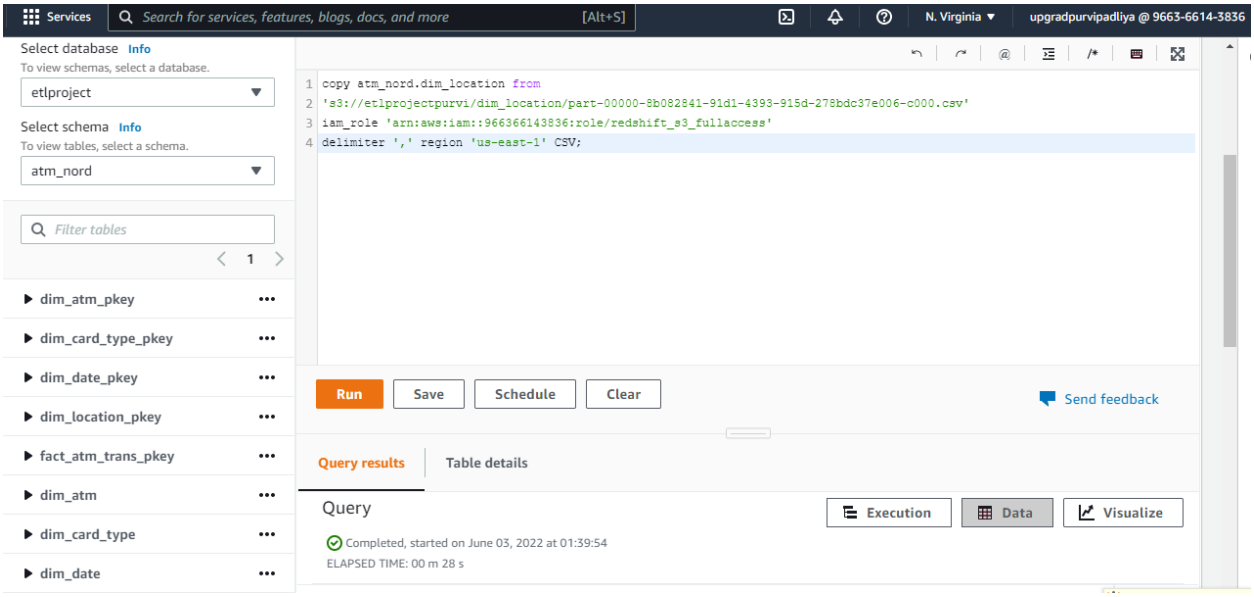
The different folders created by us for writing the data in csv format for dimension and fact table



Queries to copy the data from S3 buckets to the Redshift cluster in the appropriate tables

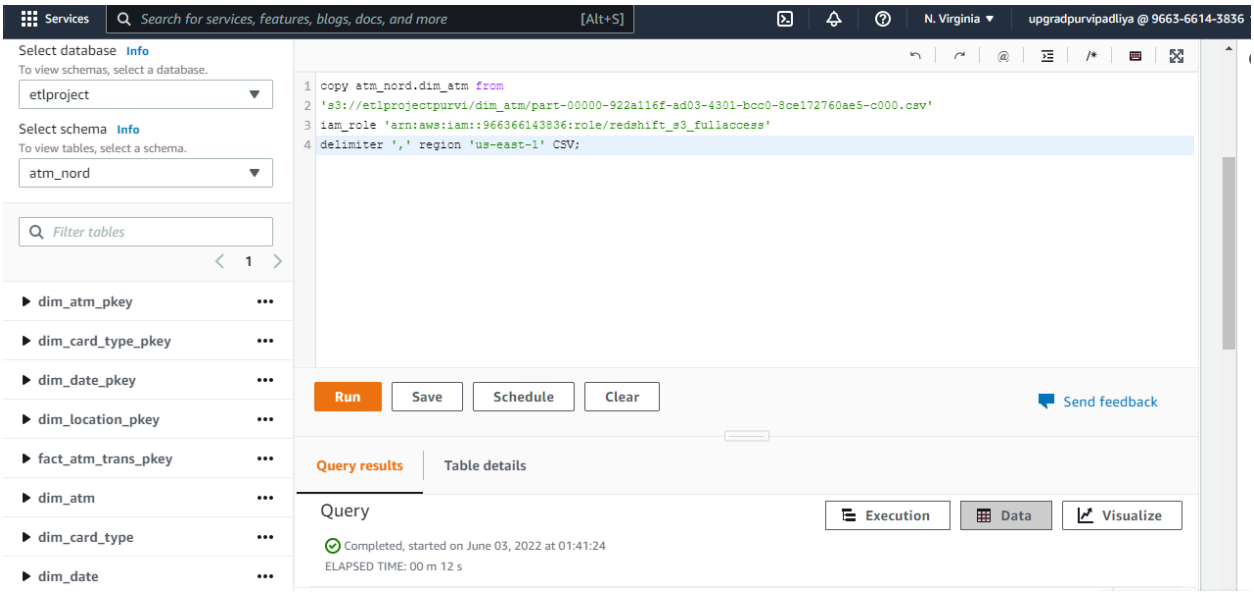
Query to copy the data from S3 bucket – etlprojectpurvi to dim_location table

```
copy atm_nord.dim_location from
's3://etlprojectpurvi/dim_location/part-00000-8b082841-91d1-4393-915d-278bdc37e006-
c000.csv'
iam_role 'arn:aws:iam::966366143836:role/redshift_s3_fullaccess'
delimiter ',' region 'us-east-1' CSV;
```



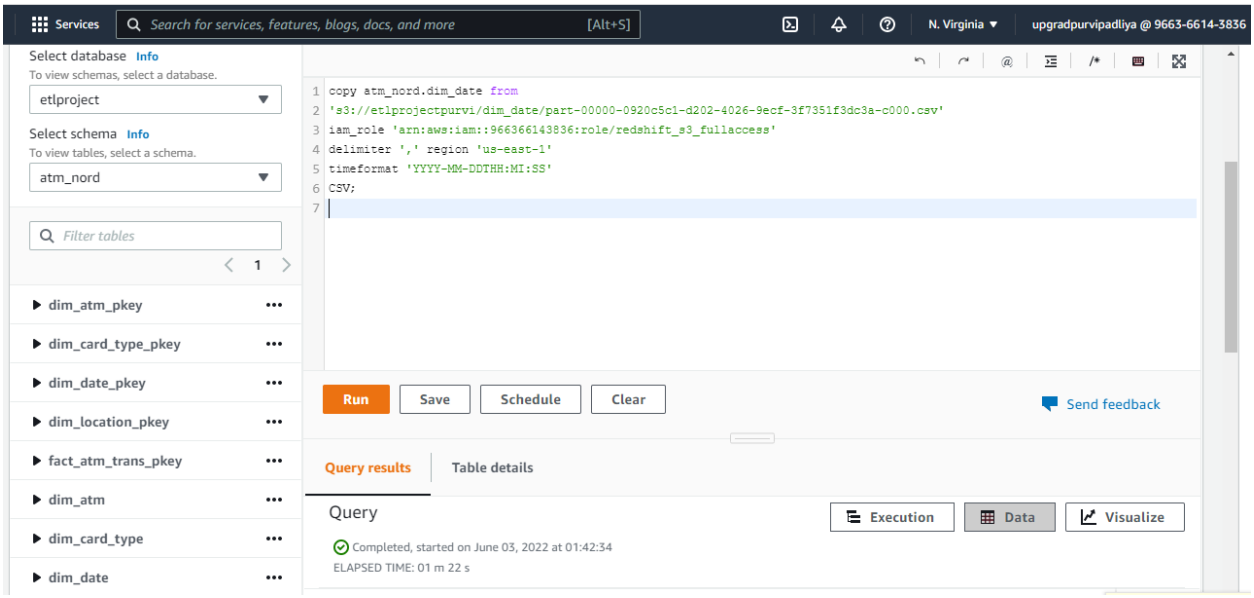
Query to copy the data from S3 bucket – etlprojectpurvi to dim_atm table

copy atm_nord.dim_atm from
's3://etlprojectpurvi/dim_atm/part-00000-922a116f-ad03-4301-bcc0-8ce172760ae5-c000.csv'
iam_role 'arn:aws:iam::966366143836:role/redshift_s3_fullaccess'
delimiter ',' region 'us-east-1' CSV;



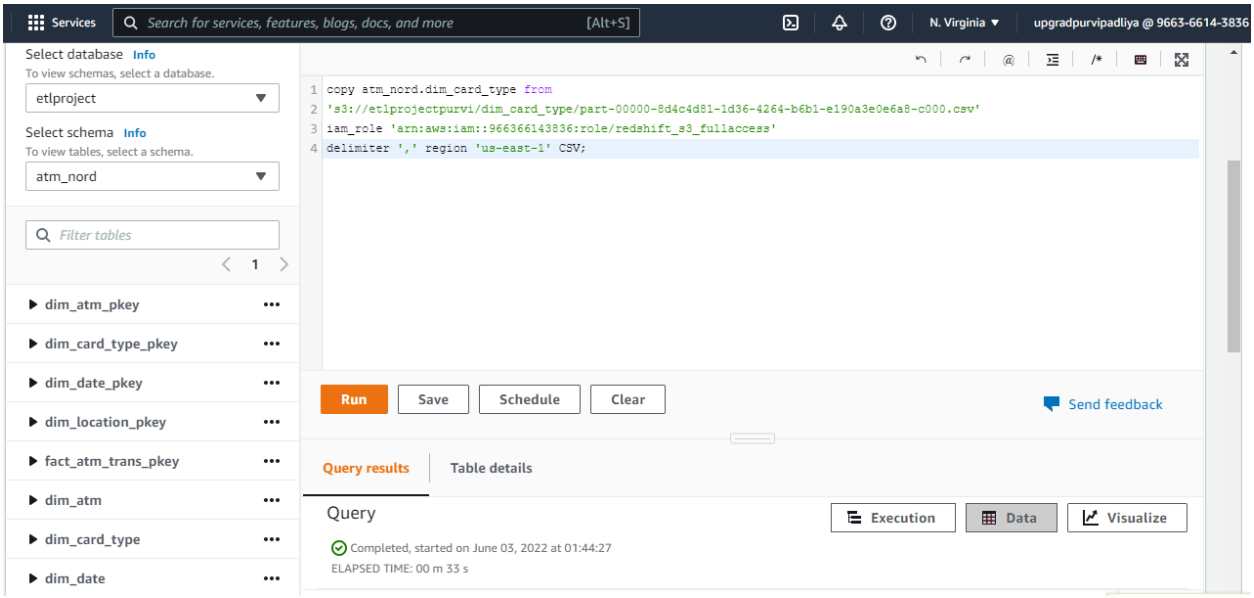
Query to copy the data from S3 bucket – etlprojectpurvi to dim_date table

```
copy atm_nord.dim_date from
's3://etlprojectpurvi/dim_date/part-00000-0920c5c1-d202-4026-9ecf-3f7351f3dc3a-c000.csv'
iam_role 'arn:aws:iam::966366143836:role/redshift_s3_fullaccess'
delimiter ',' region 'us-east-1'
timeformat 'YYYY-MM-DDTHH:MI:SS'
CSV;
```



Query to copy the data from S3 bucket – etlprojectpurvi to dim_card_type table

```
copy atm_nord.dim_card_type from
's3://etlprojectpurvi/dim_card_type/part-00000-8d4c4d81-1d36-4264-b6b1-e190a3e0e6a8-
c000.csv'
iam_role 'arn:aws:iam::966366143836:role/redshift_s3_fullaccess'
delimiter ',' region 'us-east-1' CSV;
```



Query to copy the data from S3 bucket – etlprojectpurvi to fact_atm_trans table

copy atm_nord.fact_atm_trans from
's3://etlprojectpurvi/fact_atm_trans/part-00000-8081791d-8d6f-43e8-981b-a0d717ccf77c-c000.csv'
iam_role 'arn:aws:iam::966366143836:role/redshift_s3_fullaccess'
delimiter ',' region 'us-east-1' CSV;