

## Pre-requisites csv Datasets created via the spark setup to be available in s3

The screenshot shows the AWS S3 console interface. The left sidebar is collapsed, and the main area displays the contents of the 'upgradijassignment' bucket. The bucket path is 'Amazon S3 > upgradijassignment > ETL/'. The 'Objects' tab is selected, showing 5 items:

Name	Type	Last modified	Size	Storage class
DIM_ATM/	Folder	-	-	-
DIM_CARD_TYPE/	Folder	-	-	-
DIM_DATE/	Folder	-	-	-
DIM_LOCATION/	Folder	-	-	-
FACT_ATM_TRANS/	Folder	-	-	-

# Creation of a Redshift Cluster

Screenshots of the configuration of the RedShift cluster that you have created:

The screenshot shows the AWS Redshift console. At the top, there's a navigation bar with tabs like 'Subscription Details | Nuvepro', 'Redshift', and a search bar. Below the navigation is a sidebar with various service icons. The main content area has a heading 'Connect to Redshift clusters'. It includes sections for 'Query data using Redshift query editor' (with a 'Query data' button), 'Work with your client tools' (with a 'Cluster' dropdown and 'Copy JDBC URL'/'Copy ODBC URL' buttons), and 'Choose your JDBC or ODBC driver' (with a 'Driver' dropdown set to 'JDBC 4.2 without AWS SDK (.jar)' and a 'Download driver' button). Below this is a table titled 'Clusters (1) Info' showing one entry: 'redshift-atm-ett-cluster-1' (Cluster namespace: dba6d788-fff0-4102-8...), Status: Available, CPU utilization: < 1%, and Storage capacity usage: 41%. There are buttons for 'Query data', 'Actions', and 'Create cluster' at the top of the table. The bottom of the page includes standard AWS footer links.

Cluster Name

The screenshot shows the AWS Redshift console under 'Configurations > Subnet groups > Subnet group'. The title is 'cluster-subnet-group-1'. The main section is 'Cluster subnet group details' showing VPC ID 'vpc-0ead4701cc0502e24', Description 'ATM ETL Assignment', and Status 'Complete'. It also lists 'Attached clusters' as 'redshift-atm-ett-cluster-1'. Below this is a table for 'Subnets (1)' with columns 'Availability Zone', 'Subnet ID', and 'CIDR block'. One row is shown: 'us-east-1f' with 'subnet-0d1f25488d0062e8f' and '10.0.0.0/24'. At the bottom is a 'Tags (0)' section with a 'Manage tags' button. The bottom of the page includes standard AWS footer links.

Subnet group and subnet

The screenshot shows the 'Amazon Redshift > Clusters > redshift-atm-etl-cluster-1' page. The cluster identifier is 'redshift-atm-etl-cluster-1'. The status is 'Available'. The node type is 'dc2.large'. The endpoint is 'redshift-atm-etl-cluster-1.cvtv73lza8p.us...'. The date created is 'November 04, 2021, 10:19(UTC+05:30)'. The number of nodes is '2'. Storage used is '0.02% (0.06 of 320 GB used)'. The JDBC URL is 'jdbc:redshift://redshift-atm-etl-cluster-1....'. The ODBC URL is 'Driver={Amazon Redshift (x64)}; Server=...'. The properties tab is selected.

## Cluster Details

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

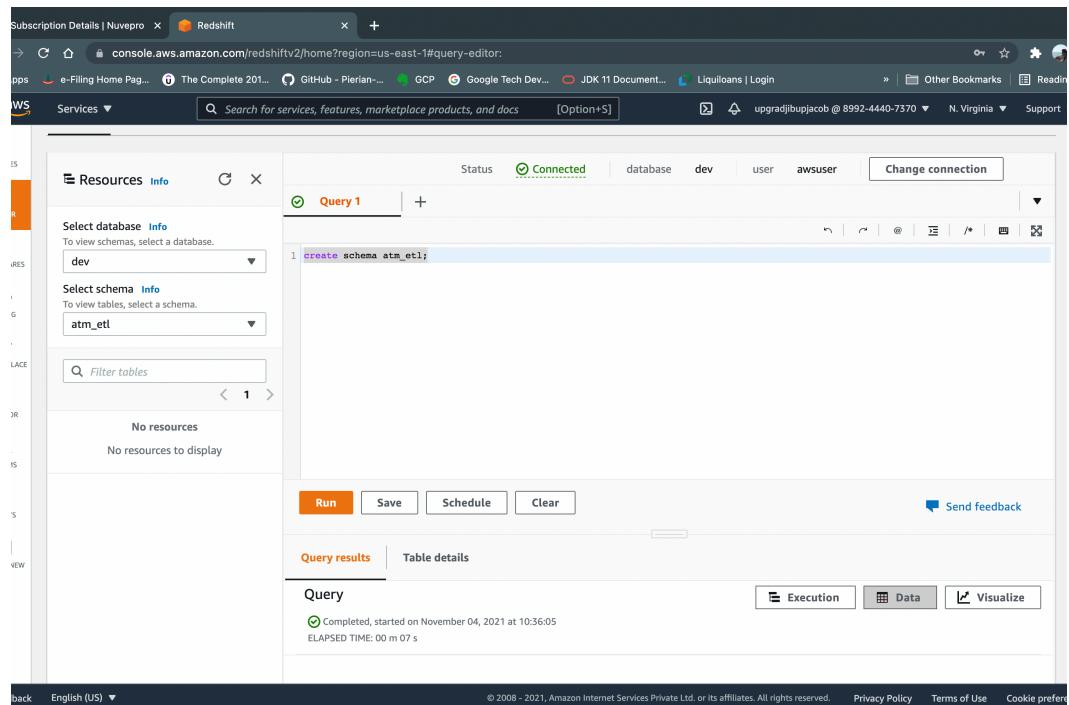
**Queries to create the various dimension and fact tables with appropriate primary and foreign keys:**

### 1. Schema Creation:

```
create schema atm_etl;
```

## 2. DIM\_LOCATION dimension table creation

```
create table atm_etl.dim_location(
    location_id integer not null,
    location varchar(50),
    streetname varchar(255),
    street_number integer,
    zipcode integer,
    lat numeric(10,3),
    lon numeric(10,3),
    primary key(location_id));
```



### 3. DIM\_ATM dimension table creation

```
create table atm_etl.dim_atm(
atm_id integer not null,
atm_number varchar(20),
atm_manufacturer varchar(50),
atm_location_id integer,
primary key(atm_id),
foreign key (atm_location_id) references atm_etl.dim_location(location_id));
```

The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar with navigation links like 'Subscription Details | Nuvepro', 'Redshift', and 'Services'. The main area has tabs for 'Resources' (selected), 'Info', 'Status' (Connected), 'database', 'dev', 'user', and 'awsuser'. A 'Change connection' button is also present. The central part of the screen is the 'Query 1' editor, which contains the SQL code for creating the DIM\_ATM table. Below the editor are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. To the right of the editor is a 'Send feedback' link. At the bottom, there are tabs for 'Query results' (selected) and 'Table details', along with buttons for 'Execution', 'Data', and 'Visualize'. The status bar at the bottom indicates the query completed successfully on November 04, 2021, at 10:44:27, with an elapsed time of 01 m 10 s.

```
create table atm_etl.dim_atm(
atm_id integer not null,
atm_number varchar(20),
atm_manufacturer varchar(50),
atm_location_id integer,
primary key(atm_id),
foreign key (atm_location_id) references atm_etl.dim_location(location_id));
```

#### 4.DIM\_DATE dimension table creation

```
create table atm_etl.dim_date(
    date_id integer not null,
    full_date_time timestamp,
    year integer,
    month varchar(20),
    day integer,
    hour integer,
    weekday varchar(20),
    primary key(date_id));
```

The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar titled 'Resources' with dropdown menus for 'Select database' (set to 'dev') and 'Select schema' (set to 'public'). Below these are buttons for 'Filter tables' and 'No resources'. The main area is titled 'Query 1' and contains the SQL code for creating the 'dim\_date' table. At the bottom of the query editor, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. To the right of the query editor, there's a 'Query results' section which shows a green checkmark indicating the query completed successfully on November 04, 2021, at 10:48:11, with an elapsed time of 00 m 19 s. There are also tabs for 'Table details', 'Execution', 'Data', and 'Visualize'.

```
create table atm_etl.dim_date(
    date_id integer not null,
    full_date_time timestamp,
    year integer,
    month varchar(20),
    day integer,
    hour integer,
    weekday varchar(20),
    primary key(date_id));
```

## 5. DIM\_CARD\_TYPE dimension table creation

```
create table atm_etl.dim_card_type(
    card_type_id integer not null,
    card_type varchar(20),
    primary key(card_type_id));
```

The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar with various service icons. The main area has tabs for 'Resources' and 'Info'. Under 'Info', it says 'Select database: dev' and 'Select schema: atm\_etl'. A search bar at the top says 'Search for services, features, marketplace products, and docs'. Below the search bar, there's a connection status indicator 'Connected' and a dropdown for 'awsuser'. The central part of the screen is a code editor titled 'Query 1' containing the SQL code for creating the table. At the bottom of the code editor are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. To the right of the code editor is a 'Query results' section which shows a green success icon and the message 'Completed, started on November 04, 2021 at 10:50:05 ELAPSED TIME: 00 m 35 s'. There are also tabs for 'Table details', 'Execution', 'Data', and 'Visualize'.

Altered the size of the card type to 23 to accommodate the data

```
alter table atm_etl.dim_card_type alter column card_type type varchar(23);
```

## 6. FACT\_ATM\_TRANS fact table creation

```
create table atm_etl.fact_atm_trans(
trans_id integer not null,
atm_id integer,
weather_loc_id integer,
date_id integer,
card_type_id integer,
atm_status varchar(20),
currency varchar(10),
service varchar(20),
transaction_amount integer,
message_code varchar(255),
message_text varchar(255),
rain_3h numeric(10,3),
clouds_all integer,
weather_id integer,
weather_main varchar(50),
weather_description varchar(255),
primary key(trans_id),
foreign key (weather_loc_id) references atm_etl.dim_location(location_id),
foreign key (atm_id) references atm_etl.dim_atm(atm_id),
foreign key (date_id) references atm_etl.dim_date(date_id),
foreign key (card_type_id) references atm_etl.dim_card_type(card_type_id));
```

The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar with 'Resources' and 'Info' sections, showing databases like 'dev' and schemas like 'atm\_etl'. The main area has a 'Query 1' tab open with the SQL code for creating the table. Below the query editor are tabs for 'Query results' and 'Table details'. At the bottom, there are execution metrics: 'Completed, started on November 04, 2021 at 10:57:07' and 'ELAPSED TIME: 00 m 49 s'.

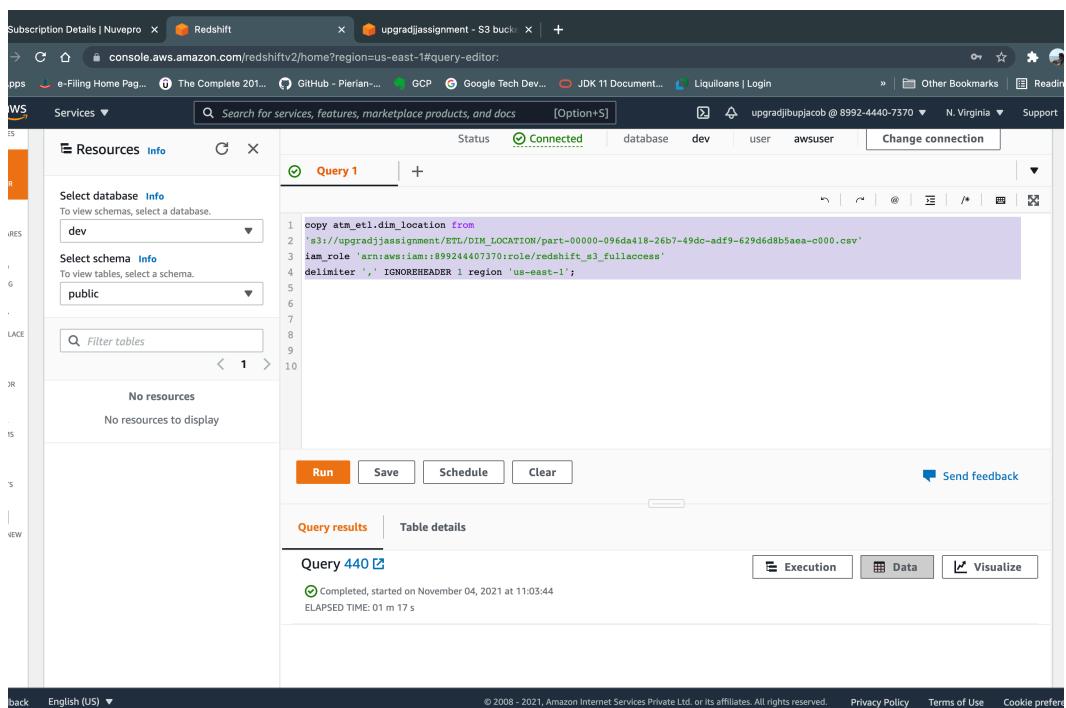
```
create table atm_etl.fact_atm_trans(
1 trans_id integer not null,
2 atm_id integer,
3 weather_loc_id integer,
4 date_id integer,
5 card_type_id integer,
6 atm_status varchar(20),
7 currency varchar(10),
8 service varchar(20),
9 transaction_amount integer,
10 message_code varchar(255),
11 message_text varchar(255),
12 rain_3h numeric(10,3),
13 clouds_all integer,
14 weather_id integer,
15 weather_main varchar(50),
16 weather_description varchar(255),
17 primary key(trans_id),
18 foreign key (weather_loc_id) references atm_etl.dim_location(location_id),
19 foreign key (atm_id) references atm_etl.dim_atm(atm_id),
20 foreign key (date_id) references atm_etl.dim_date(date_id),
21 foreign key (card_type_id) references atm_etl.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**

## 1. Copy DIM\_LOCATION data from S3 to Redshift

```
copy atm_etl.dim_location from
's3://upgradjjassignment/ETL/DIM_LOCATION/part-00000-096da418-26b7-49dc-
adf9-629d6d8b5aea-c000.csv'
iam_role 'arn:aws:iam::899244407370:role/redshift_s3_fullaccess'
delimiter ',' IGNOREHEADER 1 region 'us-east-1';
```



## 2. Copy DIM\_ATM data from S3 to Redshift

```
copy atm_etl.dim_atm from
's3://upgradjjassignment/ETL/DIM_ATM/part-00000-4f8e8763-7acc-4fa3-bb74-e32648c59e37-
c000.csv'
iam_role 'arn:aws:iam::899244407370:role/redshift_s3_fullaccess'
delimiter ',' IGNOREHEADER 1 region 'us-east-1';
```

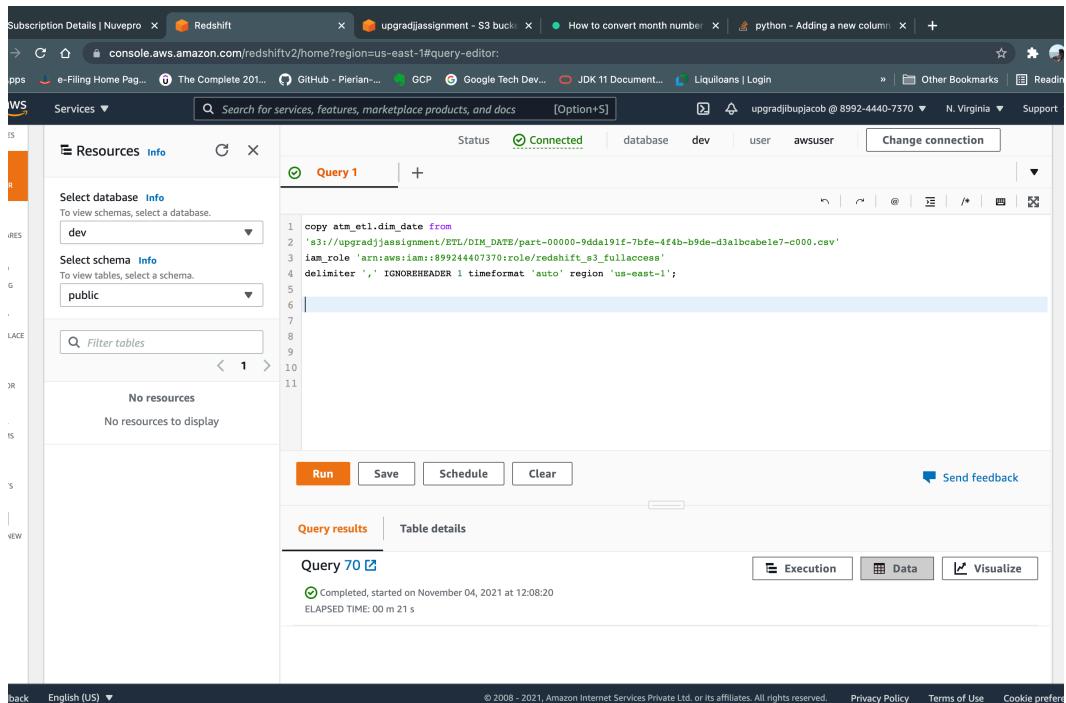
The screenshot shows the AWS Redshift Query Editor interface. On the left, there is a sidebar with various service icons. The main area has tabs for 'Status' (Connected), 'database', 'dev', 'user', and 'awsuser'. A 'Change connection' button is also present. The central part of the screen displays a query editor with a query titled 'Query 1'. The query text is:

```
copy atm_etl.dim_atm from
's3://upgradjjassignment/ETL/DIM_ATM/part-00000-4f8e8763-7acc-4fa3-bb74-e32648c59e37-
c000.csv'
iam_role 'arn:aws:iam::899244407370:role/redshift_s3_fullaccess'
delimiter ',', IGNOREHEADER 1 region 'us-east-1';
```

Below the query editor are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. To the right of these buttons is a 'Send feedback' link. At the bottom of the editor, there are tabs for 'Query results' and 'Table details'. Under 'Query results', it says 'Completed, started on November 04, 2021 at 11:07:49' and 'ELAPSED TIME: 02 m 21 s'. At the very bottom of the page, there are links for 'Privacy Policy', 'Terms of Use', and 'Cookie preferences'.

### 3. Copy DIM\_DATE data from S3 to Redshift

```
copy atm_etl.dim_date from
's3://upgradjassignment/ETL/DIM_DATE/part-00000-9dda191f-7bfe-4f4b-b9de-d3a1bcabe1e7-
c000.csv'
iam_role 'arn:aws:iam::899244407370:role/redshift_s3_fullaccess'
delimiter ',' IGNOREHEADER 1 timeformat 'auto' region 'us-east-1';
```



The screenshot shows the AWS Redshift Query Editor interface. On the left, there's a sidebar with various AWS services like Lambda, S3, and CloudWatch Metrics. The main area has tabs for 'Query 1' and 'Query results'. The query editor contains the following SQL code:

```
copy atm_etl.dim_date from
's3://upgradjassignment/ETL/DIM_DATE/part-00000-9dda191f-7bfe-4f4b-b9de-d3a1bcabe1e7-
c000.csv'
iam_role 'arn:aws:iam::899244407370:role/redshift_s3_fullaccess'
delimiter ',' IGNOREHEADER 1 timeformat 'auto' region 'us-east-1';
```

Below the code, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Query results' tab is selected, showing a status message: 'Completed, started on November 04, 2021 at 12:08:20 ELAPSED TIME: 00 m 21 s'. There are also tabs for 'Execution' and 'Data'.

Caption

#### 4. Copy DIM\_CARD\_TYPE data from S3 to Redshift

```
copy atm_etl.dim_card_type from
's3://upgradjjassignment/ETL/DIM_CARD_TYPE/part-00000-09b58326-48dc-45ff-
b105-98ea5dfca406-c000.csv'
iam_role 'arn:aws:iam::899244407370:role/redshift_s3_fullaccess'
delimiter ',' IGNOREHEADER 1 region 'us-east-1';
```

The screenshot shows the AWS Redshift Query Editor interface. On the left, there are dropdown menus for 'Select database' (set to 'dev') and 'Select schema' (set to 'public'). Below these are sections for 'Filter tables' and 'No resources'. The main area contains a code editor with the following SQL query:

```
copy atm_etl.dim_card_type from
's3://upgradjjassignment/ETL/DIM_CARD_TYPE/part-00000-09b58326-48dc-45ff-
b105-98ea5dfca406-c000.csv'
iam_role 'arn:aws:iam::899244407370:role/redshift_s3_fullaccess'
delimiter ',' IGNOREHEADER 1 region 'us-east-1';

select * from stl_load_errors;
```

Below the code editor are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. To the right of the code editor is a 'Send feedback' link. Under the code editor, there are tabs for 'Query results' and 'Table details'. The 'Query results' tab shows a summary: 'Completed, started on November 04, 2021 at 11:17:41' and 'ELAPSED TIME: 00 m 08 s'. It also displays 'Rows returned (12)' and a search bar. At the bottom of the page, there are links for 'Privacy Policy', 'Terms of Use', and 'Cookie preferences'.

## 5. Copy FACT\_ATM\_TRANS data from S3 to Redshift

```
copy atm_etl.fact_atm_trans from
's3://upgradjjassignment/ETL/FACT_ATM_TRANS/part-00000-17259abc-6328-48b2-
a7f8-69f53ada18eb-c000.csv'
iam_role 'arn:aws:iam::899244407370:role/redshift_s3_fullaccess'
delimiter ',' IGNOREHEADER 1 region 'us-east-1' TRUNCATECOLUMNS CSV;
```

The screenshot shows the AWS Redshift Query Editor interface. On the left, there is a sidebar with 'Resources' and 'Info' sections. Under 'Select database' dropdown, 'dev' is selected. Under 'Select schema' dropdown, 'public' is selected. The main area is titled 'Query 1' and contains the following SQL code:

```
1 copy atm_etl.fact_atm_trans from
2 's3://upgradjjassignment/ETL/FACT_ATM_TRANS/part-00000-17259abc-6328-48b2-a7f8-69f53ada18eb-c000.csv'
3 iam_role 'arn:aws:iam::899244407370:role/redshift_s3_fullaccess'
4 delimiter ',', IGNOREHEADER 1 region 'us-east-1' TRUNCATECOLUMNS CSV;
```

Below the code, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. The 'Run' button is highlighted in orange. To the right of the code editor, there is a 'Send feedback' link. Below the code editor, there are tabs for 'Query results' and 'Table details'. The 'Query results' tab is active, showing 'Query 350' which completed successfully on November 04, 2021 at 12:32:55. The elapsed time was 00 m 08 s. The 'Rows returned (1)' section shows one row. At the bottom of the editor, there are links for 'Execution', 'Data', and 'Visualize', along with an 'Export' button.