

# **ETL CASESTUDY – Redshift setup**

## Contents

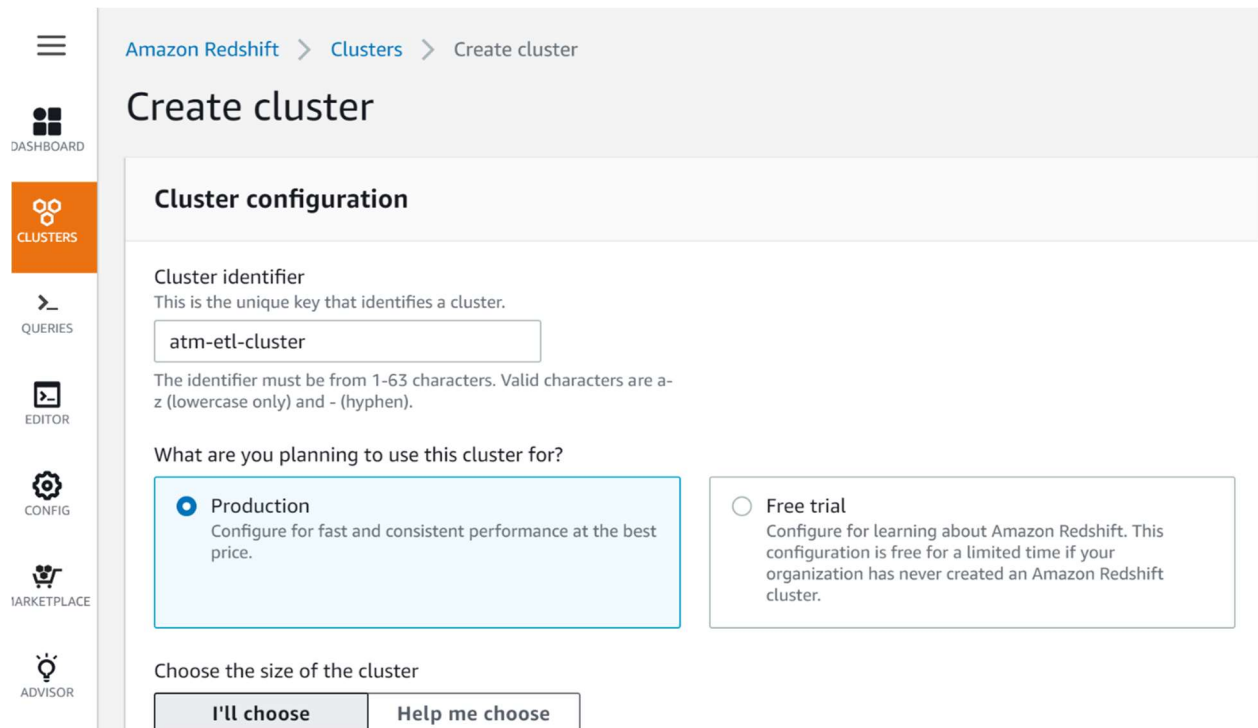
ETL CASESTUDY – Redshift setup .....	1
<b>Amazon redshift cluster setup .....</b>	<b>3</b>
Step 1 – Cluster identifier .....	3
Step 2- Node type .....	3
Step 3 – Database configuration.....	4
Step 4 - VPC.....	5
Step 5 – Availability zone .....	5
Step 6 – Create cluster .....	6
Step 7- IAM role .....	6
Step 8 – Associate IAM role .....	6
Step 9 – Redshift cluster .....	7
<b>Create Tables .....</b>	<b>8</b>
Connect to database. ....	8
Create schema. ....	8
Creating location dimension .....	8
Creating ATM dimension table .....	9
Creating date dimension.....	10
Creating card type dimension .....	11
Creating Fact table .....	12
<b>Load Data .....</b>	<b>13</b>
Load data to Location dimension.....	13
Load data to ATM dimension.....	14
Load data into date dimension .....	15
Load data into card type dimension .....	16
Load data into atm transaction fact table .....	16

## Amazon redshift cluster setup

As a part of the ETL casestudy data is loaded into redshift warehouse to run analytical queries and derive insights.

### Step 1 – Cluster identifier

Navigate to Redshift dashboard on AWS console and click on create cluster.  
Enter cluster identifier.



Amazon Redshift > Clusters > Create cluster

## Create cluster

### Cluster configuration

**Cluster identifier**  
This is the unique key that identifies a cluster.

atm-etl-cluster

The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

**What are you planning to use this cluster for?**

☒ **Production**  
Configure for fast and consistent performance at the best price.

☐ **Free trial**  
Configure for learning about Amazon Redshift. This configuration is free for a limited time if your organization has never created an Amazon Redshift cluster.

**Choose the size of the cluster**

**I'll choose** **Help me choose**

### Step 2- Node type

Choose the data node type as DC2.large with 2 nodes.

Choose the size of the cluster

I'll choose

Help me choose

#### Node type

Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

dc2.large

#### Nodes

Enter the number of nodes that you need.

2

Range (1-32)

### Configuration summary

dc2.large | 2 nodes

\$360.00/month

Estimated on-demand compute

320 GB

Total compressed storage

## Step 3 – Database configuration

Provide database details. Changed the default port number for security purposes.

### Database configurations

#### Database name (optional)

Specify a database name to create an additional database.

atmtransdb

The name must be 1-64 alphanumeric characters (lowercase only), and it can't be a **reserved word**.

#### Database port (optional)

Port number where the database accepts inbound connections. You can't change the port after the cluster has been created.

5300

The port must be numeric (1150-65535).

#### Master user name

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

awsuser

The name must be 1-128 alphanumeric characters, and it can't be a **reserved word**.

#### Master user password

.....

☐ Show password

- The master password must be 8 - 64 characters.
- The value must contain at least one uppercase letter.
- The value must contain at least one lowercase letter.
- The value must contain at least one number.

## Step 4 - VPC

In the additional configurations section provide VPC and security group settings.

### Additional configurations ☐ Use defaults

These configurations are optional, and default settings have been defined to help you get started with your cluster. Turn off "Use defaults" to modify these settings now.

▼ Network and security

Virtual private cloud (VPC)

This VPC defines the virtual networking environment for this cluster.

my\_vpc  
vpc-01d6e4490999c4463 ▼

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 ▼

cloudera  
sg-08bd70e0f697ddfb1 ✕

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.

us-east-1d ▼

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

## Step 5 – Availability zone

Further provide the availability zone and subnet group.

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.

us-east-1d ▼

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

---

## Step 6 – Create cluster

Click on create cluster.



☒ Disable  
☐ Enable

► Database configurations

► Maintenance

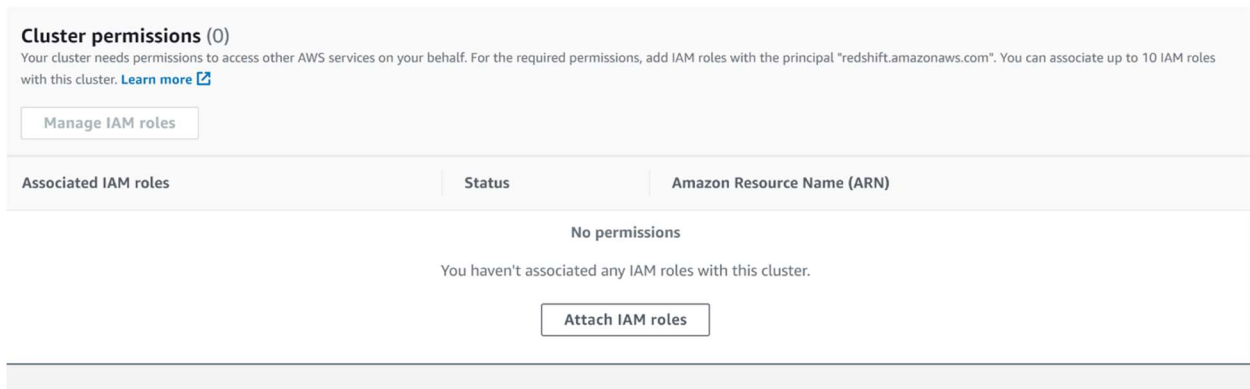
► Monitoring

► Backup

Cancel Create cluster

## Step 7- IAM role

After the cluster is successfully created, associate the IAM role giving other AWS services permission to access redshift. Navigate to cluster properties tab to find Cluster permissions section, then click on Attach IAM roles.



**Cluster permissions (0)**  
Your cluster needs permissions to access other AWS services on your behalf. For the required permissions, add IAM roles with the principal "redshift.amazonaws.com". You can associate up to 10 IAM roles with this cluster. [Learn more](#)

Manage IAM roles

Associated IAM roles	Status	Amazon Resource Name (ARN)
No permissions		
You haven't associated any IAM roles with this cluster.		

Attach IAM roles

## Step 8 – Associate IAM role

Select the IAM role and click on associate IAM role, then click on save changes.

## Manage IAM roles: atm-etl-cluster

### IAM roles

Your cluster needs permissions to access other AWS services on your behalf. For the required permissions, add IAM roles with the principal "redshift.amazonaws.com". You can associate up to 10 IAM roles with this cluster. [Learn more](#)

Available IAM roles

redshift\_s3\_full\_access



Associate IAM role

No associated IAM roles

Cancel

Save changes

### Cluster permissions (1)

Your cluster needs permissions to access other AWS services on your behalf. For the required permissions, add IAM roles with the principal "redshift.amazonaws.com". You can associate up to 10 IAM roles with this cluster. [Learn more](#)

Manage IAM roles

Associated IAM roles	Status	Amazon Resource Name (ARN)
<a href="#">redshift_s3_full_access</a>	adding	arn:aws:iam::006945116289:role/redshift_s3_full_access

## Step 9 – Redshift cluster

The redshift cluster is successfully created.

Clusters (1)							
					Query cluster	Actions ▼	Create cluster
<input type="text" value="Filter clusters by property or value"/>							
<div>&lt; 1 &gt; </div>							
<input type="checkbox"/>	Cluster ▲	Cluster namespace ▼	Status ▼	Storage capacity us... ▼	CPU utilization ▼	Snapshots ▼	
<input type="checkbox"/>	atm-etl-cluster dc2.large   2 nodes   320 GB	974b9d98-afb5-45ac-...	Available	< 1%	2%	-	

## Create Tables

The data now resides in S3 bucket. We create a schema in redshift cluster to load all the dimension and fact tables.

Connect to database.

We need to connect to the database before we load any data.


**Connect to database** >


**Connection**  
Select a recent database connection or create a new database connection.

☐ Use a recent connection

☒ Create a new connection

**Authentication**

☒ Temporary credentials  
Use the GetClusterCredentials IAM permission and your database user to generate temporary access credentials. [Learn more](#) 

☐ AWS Secrets Manager  
Use a stored secret to authenticate access. [Learn more](#) 

**Cluster**

atm-etl-cluster (Available) ▼

**Database name**

atmtransdb

**Database user**  
User name authorized to access your database.

awsuser

Cancel **Connect**

Once connected to database we start by creating the schema.

Create schema.

*create schema atm\_trans.*

Creating location dimension

```
create table atm_trans.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 a SQL query editor interface. On the left, there's a sidebar with a dropdown menu set to 'atm\_trans' and a search bar labeled 'Filter tables'. Below the search bar, a table named 'dim\_location' is listed. The main area displays a SQL query to create a table named 'atm\_trans.DIM\_LOCATION'. The query is as follows:

```

4 create table atm_trans.DIM_LOCATION(
5     location_id INT not null distkey sortkey,
6     location VARCHAR(50),
7     streetname VARCHAR(255),
8     street_number INT,
9     zipcode INT,
10    lat DECIMAL(10,3),
11    lon DECIMAL(10,3)
12 );

```

Below the query, there are buttons for 'Run', 'Save', 'Schedule', and 'Clear'. To the right of these buttons is a 'Send feedback' link. Below the buttons, there are two tabs: 'Query results' (which is active) and 'Table details'. Under the 'Query results' tab, there's a 'Query' section showing a green checkmark icon and the text 'Completed, started on March 31, 2021 at 13:17:30' and 'ELAPSED TIME: 00 m 36 s'. To the right of this section are three buttons: 'Execution', 'Data', and 'Visualize'.

## Creating ATM dimension table

```

create table atm_trans.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_trans.DIM_LOCATION(location_id)
);

```

The screenshot shows a SQL query editor interface. On the left, there is a sidebar with a search bar labeled "Filter tables" and a list of tables: "dim\_atm" and "dim\_location". The main area displays a SQL query for creating a table named "atm\_trans.DIM\_ATM". The query is as follows:

```
13  
14  
15 create table atm_trans.DIM_ATM(  
16     atm_id INT not null distkey sortkey,  
17     atm_number VARCHAR(20),  
18     atm_manufacturer VARCHAR(50),  
19     atm_location_id INT  
20 );
```

Below the query, there are buttons for "Run", "Save", "Schedule", and "Clear". To the right of these buttons is a "Send feedback" link. Below the buttons, there are two tabs: "Query results" and "Table details". The "Query results" tab is active, showing a "Query" section with a green checkmark icon and the text "Completed, started on March 31, 2021 at 13:19:55" and "ELAPSED TIME: 00 m 31 s". To the right of the "Query" section are three buttons: "Execution", "Data", and "Visualize".

### Creating date dimension

```
create table atm_trans.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)  
);
```

To view tables, select a schema.

atm\_trans

Filter tables

< 1 >

- dim\_atm ...
- dim\_date ...
- dim\_location ...

```
21
22 create table atm_trans.DIM_DATE(
23     date_id INT not null distkey sortkey,
24     full_date_time TIMESTAMP,
25     year INT,
26     month VARCHAR(20),
27     day INT,
28     hour INT,
29     weekday VARCHAR(20)
30 );
```

Run Save Schedule Clear

Send feedback

Query results Table details

Query

Execution Data Visualize

Completed, started on March 31, 2021 at 13:22:59  
ELAPSED TIME: 00 m 54 s

### Creating card type dimension

```
create table atm_trans.DIM_CARD_TYPE(
    card_type_id INT not null distkey sortkey,
    card_type VARCHAR(20),
    PRIMARY KEY(card_type_id)
);
```

Filter tables

< 1 >

- dim\_atm ...
- dim\_card\_type ...
- dim\_date ...
- dim\_location ...

```
31
32 create table atm_trans.DIM_CARD_TYPE(
33     card_type_id INT not null distkey sortkey,
34     card_type VARCHAR(20)
35 );
36
37
```

Run Save Schedule Clear

Send feedback

Query results Table details

Query

Execution Data Visualize

Completed, started on March 31, 2021 at 13:25:08  
ELAPSED TIME: 00 m 33 s

## Creating Fact table

```
create table atm_trans.FACT_ATM_TRANS(  
    trans_id BIGINT not null distkey sortkey,  
    atm_id INT not null,  
    weather_loc_id INT not null,  
    date_id INT not null,  
    card_type_id INT not null,  
    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_trans.DIM_LOCATION(location_id),  
    FOREIGN KEY(atm_id) REFERENCES atm_trans.DIM_ATM(atm_id),  
    FOREIGN KEY(date_id) REFERENCES atm_trans.DIM_DATE(date_id),  
    FOREIGN KEY(card_type_id) REFERENCES atm_trans.DIM_CARD_TYPE  
    (card_type_id)  
);
```

► dim\_location ...

► fact\_atm\_trans ...

```

42 card_type_id INT not null,
43 atm_status VARCHAR(20),
44 currency VARCHAR(10),
45 service VARCHAR(20),
46 transaction_amount INT,
47 message_code VARCHAR(255),
48 message_text VARCHAR(255),
49 rain_3h DECIMAL(10,3),
50 clouds_all INT,
51 weather_id INT,
52 weather_main VARCHAR(50),
53 weather_description VARCHAR(255)
54 );

```

Run Save Schedule Clear

Send feedback

Query results Table details

Query

Execution Data Visuali

Completed, started on March 31, 2021 at 13:29:22

ELAPSED TIME: 00 m 49 s

## Load Data

Once all the tables are created successfully, we can copy the data from S3 bucket into redshift.

Load data to Location dimension

```

copy atm_trans.DIM_LOCATION from
's3://atmtransetl/atm_trans/location/part-00000-ce21b2a1-bac6-4264-
96f5-acb2e0b66713-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' IGNOREHEADER 1;

```

```
copy atm_trans.DIM_LOCATION from
's3://atmtranset1/atm_trans/location/part-00000-ce21b2a1-bac6-4264-96f5-acb2e0b66713-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' IGNOREHEADER 1;
```

Run Save Schedule Clear Send feedback

Query results Table details

Query

Execution Data Visualize

Completed, started on March 31, 2021 at 17:41:51

ELAPSED TIME: 01 m 22 s

Load data to ATM dimension

```
copy atm_trans.DIM_ATM from
's3://atmtranset1/atm_trans/atm/part-00000-ed42b98c-7830-4575-9af1-
313a5265fc8a-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' IGNOREHEADER 1;
```

```
copy atm_trans.DIM_ATM from
's3://atmtranset1/atm_trans/atm/part-00000-ed42b98c-7830-4575-9af1-313a5265fc8a-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' IGNOREHEADER 1;
```

[Run](#)[Save](#)[Schedule](#)[Clear](#)[Send feedback](#)[Query results](#)[Table details](#)

Query

[Execution](#)[Data](#)[Visualize](#)

✓ Completed, started on March 31, 2021 at 17:44:49  
ELAPSED TIME: 00 m 19 s

Load data into date dimension

```
copy atm_trans.DIM_DATE from
's3://atmtranset1/atm_trans/date/part-00000-173910bb-4eee-4d8f-8783-8898367fe347-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' timeformat 'auto' IGNOREHEADER 1;
```

```
copy atm_trans.DIM_DATE from
's3://atmtranset1/atm_trans/date/part-00000-173910bb-4eee-4d8f-8783-8898367fe347-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' timeformat 'auto' IGNOREHEADER 1;
```

[Run](#)[Save](#)[Schedule](#)[Clear](#)[Send feedback](#)[Query results](#)[Table details](#)

Query

[Execution](#)[Data](#)[Visualize](#)

✓ Completed, started on March 31, 2021 at 17:55:22  
ELAPSED TIME: 00 m 13 s

Load data into card type dimension

```
copy atm_trans.DIM_CARD_TYPE from
's3://atmtransetl/atm_trans/card/part-00000-87b4735c-ac0f-4339-b371-
5920ae84ce8a-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' IGNOREHEADER 1;
```

The screenshot displays the Amazon Redshift console interface. At the top, a code editor contains the SQL command to load data into the DIM\_CARD\_TYPE dimension. Below the editor, there are four buttons: 'Run' (highlighted in orange), 'Save', 'Schedule', and 'Clear'. To the right of these buttons is a 'Send feedback' link. Below the buttons, there are two tabs: 'Query results' (selected) and 'Table details'. Under the 'Query results' tab, the 'Query' section shows a green checkmark icon and the text 'Completed, started on March 31, 2021 at 18:02:25' and 'ELAPSED TIME: 00 m 07 s'. To the right of the query status, there are three tabs: 'Execution' (selected), 'Data', and 'Visualize'.

```
copy atm_trans.DIM_CARD_TYPE from
's3://atmtransetl/atm_trans/card/part-00000-87b4735c-ac0f-4339-b371-
5920ae84ce8a-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' IGNOREHEADER 1;
```

**Run** **Save** **Schedule** **Clear** [Send feedback](#)

**Query results** **Table details**

**Query** **Execution** **Data** **Visualize**

✓ Completed, started on March 31, 2021 at 18:02:25  
ELAPSED TIME: 00 m 07 s

Load data into atm transaction fact table

```
copy atm_trans.FACT_ATM_TRANS from
's3://atmtransetl/atm_trans/trans/part-00000-3ce97df6-9712-44ac-8fe6-
908bfd6f8753-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' IGNOREHEADER 1;
```



```
copy atm_trans.FACT_ATM_TRANS from
's3://atmtranset1/atm_trans/trans/part-00000-3ce97df6-9712-44ac-8fe6-908bfd6f8753-c000.csv'
iam_role 'arn:aws:iam::006945116289:role/redshift_s3_full_access'
delimiter ',' region 'us-east-1' IGNOREHEADER 1;
```

Run

Save

Schedule

Clear

 Send feedback

Query results

Table details

Query

 Execution

 Data

 Visualize

 Completed, started on March 31, 2021 at 18:05:06  
ELAPSED TIME: 00 m 24 s