HOW TO LOAD DATA TO AMAZON REDSHIFT FROM S3

There are several ways to load data into Amazon Redshift. In this tutorial, we'll show you one method: how to copy JSON data from S3 to Amazon Redshift, where it will be converted to SQL format.

What is Amazon Redshift?

Amazon Redshift is a data warehouse that is known for its incredible speed. Redshift can handle large volumes of data as well as database migrations.

(Infamously, Amazon came up with the name Redshift in response to Oracle's database dominance. Oracle is informally known as "Big Red".)

Other methods for loading data to Redshift

Here are other methods for data loading into Redshift:

- Write a program and use a JDBC or ODBC driver.
- Paste SQL into Redshift.
- Write data to Redshift from Amazon Glue.
- Use EMR.
- Copy JSON, CSV, or other data from S3 to Redshift.

Now, onto the tutorial.

Getting started

We will upload two JSON files to S3. Download them from here:

- Customers
- Orders

Note the format of these files:

- JSON
- There is no comma between records.
- It is not a JSON array. Just JSON records one after another.

The **orders** JSON file looks like this. It only has two records. Notice that there is no comma between records.

```
{
        "customernumber": "d5d5b72c-edd7-11ea-ab7a-0ec120e133fc",
        "ordernumber": "d5d5b72d-edd7-11ea-ab7a-0ec120e133fc",
        "comments": "syjizruunqxuaevyiaqx",
        "orderdate": "2020-09-03",
        "ordertype": "sale",
        "shipdate": "2020-09-16",
        "discount": 0.1965497953690316,
        "quantity": 29,
        "productnumber": "d5d5b72e-edd7-11ea-ab7a-0ec120e133fc"
} {
        "customernumber": "d5d5b72f-edd7-11ea-ab7a-0ec120e133fc",
        "ordernumber": "d5d5b730-edd7-11ea-ab7a-0ec120e133fc",
        "comments": "uixjbivlhdtmaelfjlrn",
        "orderdate": "2020-09-03",
        "ordertype": "sale",
        "shipdate": "2020-09-16",
        "discount": 0.6820749537170963,
        "quantity": 42,
        "productnumber": "d5d5b731-edd7-11ea-ab7a-0ec120e133fc"
}
```

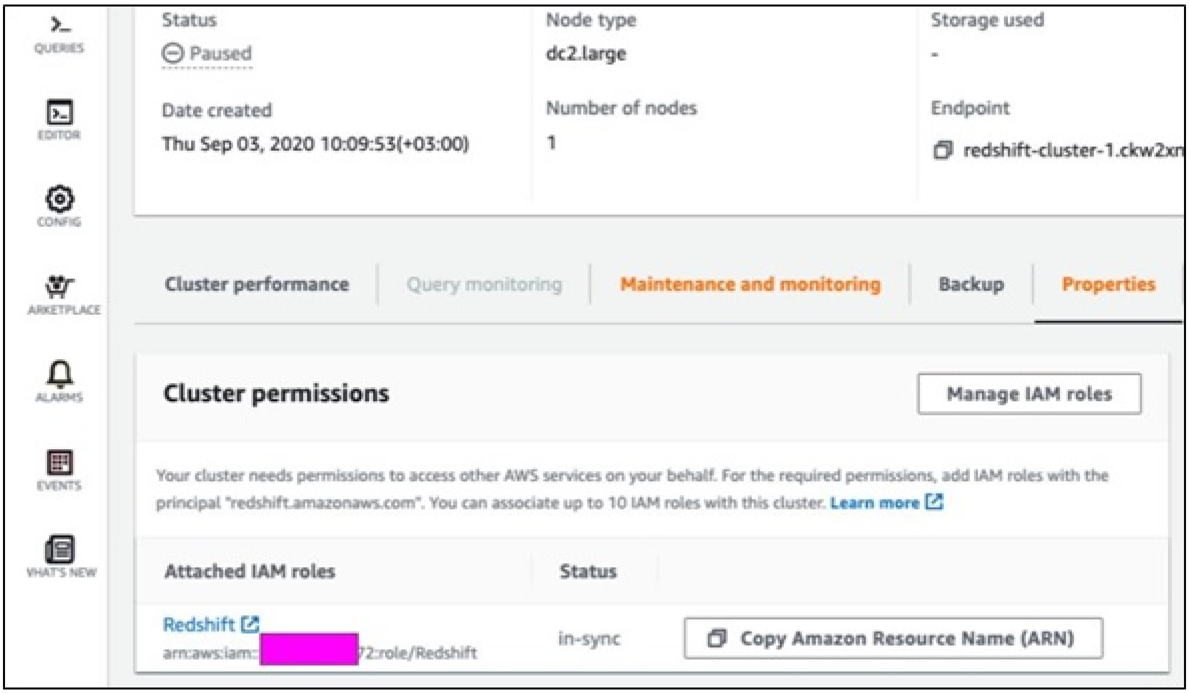
IAM role

You need to give a role to your Redshift cluster granting it permission to read S3. You don't give it to an IAM user (that is, an Identity and Access Management user).

Attach it to a cluster—a Redshift cluster in a virtual machine where Amazon installs and starts Redshift for you.

Create the role in IAM and give it some name. I used **Redshift**. Give it the permission **AmazonS3ReadOnlyAccess**. and then paste the ARN into the cluster. It will look like this:

```
arn:aws:iam::xxxxxxxxx:role/Redshift
```



Status	Node type	Storage used
Paused	dc2.large	-
Date created	Number of nodes	Endpoint
Thu Sep 03, 2020 10:09:53(+03:00)	1	redshift-cluster-1.ckw2xn

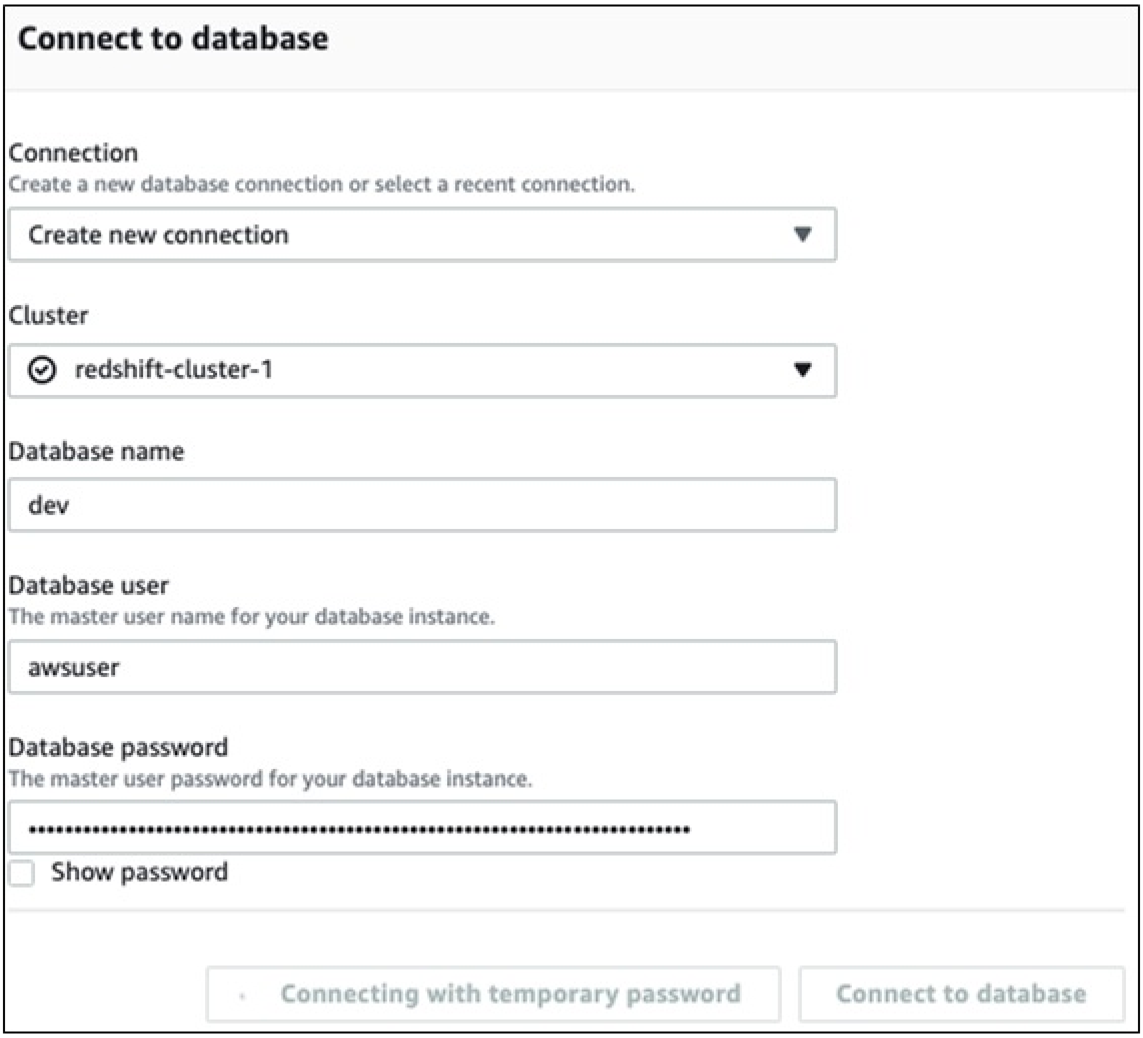
Cluster permissions

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

Attached IAM roles	Status
Redshift arn:aws:iam::[redacted]72:role/Redshift	in-sync

Create connection to a database

After you start a Redshift cluster and you want to open the editor to enter SQL commands, you login as the **awsuser** user. The default database is **dev**. Use the option **connect with temporary password**.



Connect to database

Connection
Create a new database connection or select a recent connection.
Create new connection ▼

Cluster
redshift-cluster-1 ▼

Database name
dev

Database user
The master user name for your database instance.
awsuser

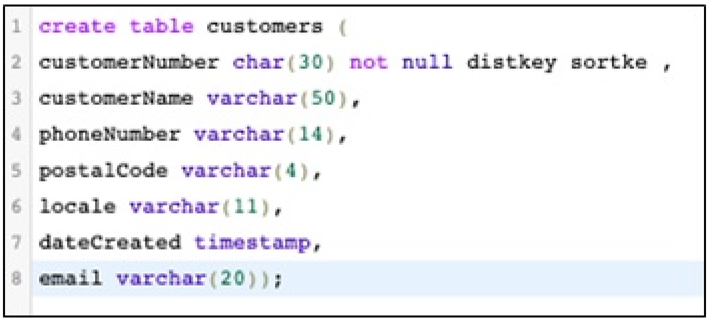
Database password
The master user password for your database instance.
••

□ Show password

Create tables

Paste in these two SQL commands to create the customers and orders table in Redshift.

```
create table customers (
customerNumber char(40) not null distkey sortkey ,
customerName varchar(50),
phoneNumber varchar(14),
postalCode varchar(4),
locale varchar(11),
dateCreated timestamp,
email varchar(20));
```



```
create table customers (
customerNumber char(30) not null distkey sortke ,
customerName varchar(50),
phoneNumber varchar(14),
postalCode varchar(4),
locale varchar(11),
dateCreated timestamp,
email varchar(20));
```

```
create table orders (
    customerNumber char(40)  not null distkey sortkey,
    orderNumber char(40) not null,
    comments varchar(200),
    orderDate timestamp,
    orderType varchar(20),
    shipDate timestamp,
    discount real,
    quantity integer,
    productNumber varchar(50));
```

Upload JSON data to S3

Create an S3 bucket if you don't already have one. If you have installed the AWS client and run **aws configure** you can do that with **aws s3 mkdir**. Then copy the JSON files to S3 like this:

```
aws s3 cp customers.json s3:/(bucket name)

 aws s3 cp orders.json s3://(bucket name)
```

Copy S3 data into Redshift

Use these SQL commands to load the data into Redshift. Some items to note:

- Use the arn string copied from IAM with the credentials aws_iam_role.
- You don't need to put the region unless your Glue instance is in a different Amazon region than your S3 buckets.
- JSON auto means that Redshift will determine the SQL column names from the JSON. Otherwise you would have to create a JSON-to-SQL mapping file.

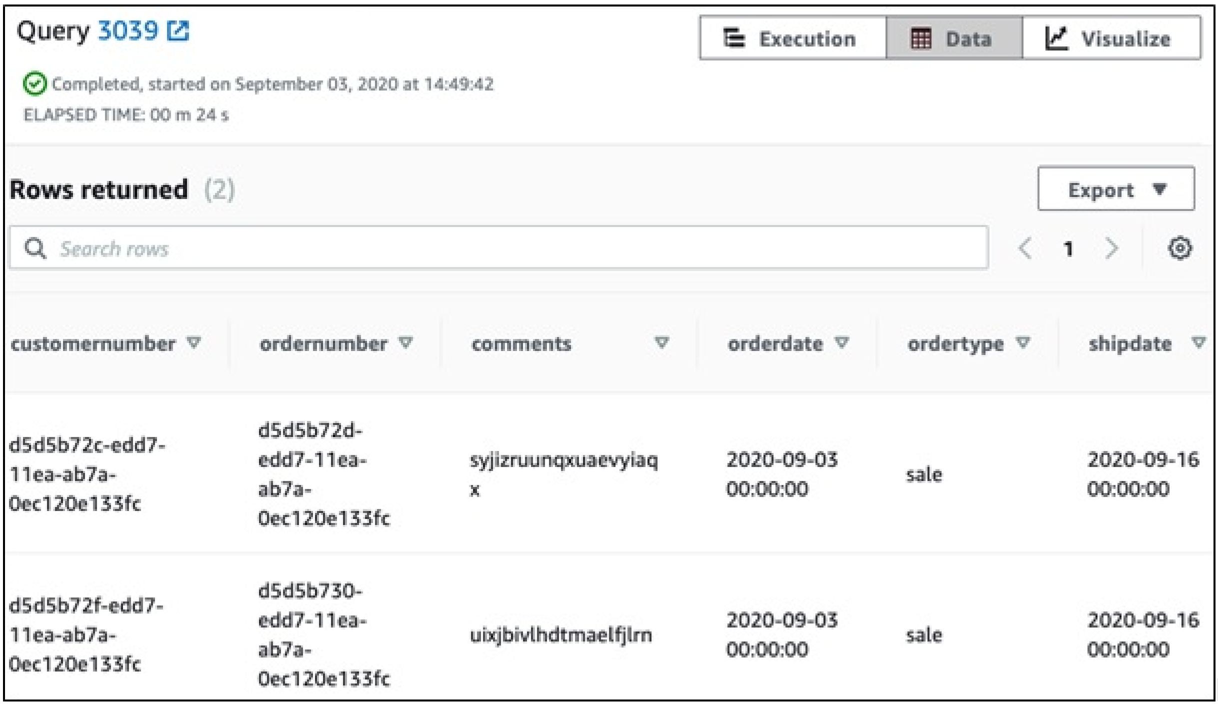
```
copy customers
from 's3://gluebmcwalkerrowe/customers.json'
credentials 'aws_iam_role=arn:aws:iam::xxxxxxx:role/Redshift'
region 'eu-west-3'
json 'auto';

copy orders
from 's3://gluebmcwalkerrowe/orders.json'
credentials 'aws_iam_role=arn:aws:iam::xxxx:role/Redshift'
region 'eu-west-3'
json 'auto';
```

Now you can run this query:

```
select * from orders;
```

And it will produce this output.



Query 3039

Completed, started on September 03, 2020 at 14:49:42
ELAPSED TIME: 00 m 24 s

Rows returned (2)

customernumber	ordernumber	comments	orderdate	ordertype	shipdate
d5d5b72c-edd7-11ea-ab7a-0ec120e133fc	d5d5b72d-edd7-11ea-ab7a-0ec120e133fc	syjizruunqxuaevyiaqx	2020-09-03 00:00:00	sale	2020-09-16 00:00:00
d5d5b72f-edd7-11ea-ab7a-0ec120e133fc	d5d5b730-edd7-11ea-ab7a-0ec120e133fc	uixjbivlhdtmaelfjlrn	2020-09-03 00:00:00	sale	2020-09-16 00:00:00

Repeat for customer data as well.

Additional resources

For more on this topic, explore these resources:

- BMC Machine Learning & Big Data Blog
- AWS Guide, with 15+ articles and tutorials
- Amazon Braket Quantum Computing: How To Get Started