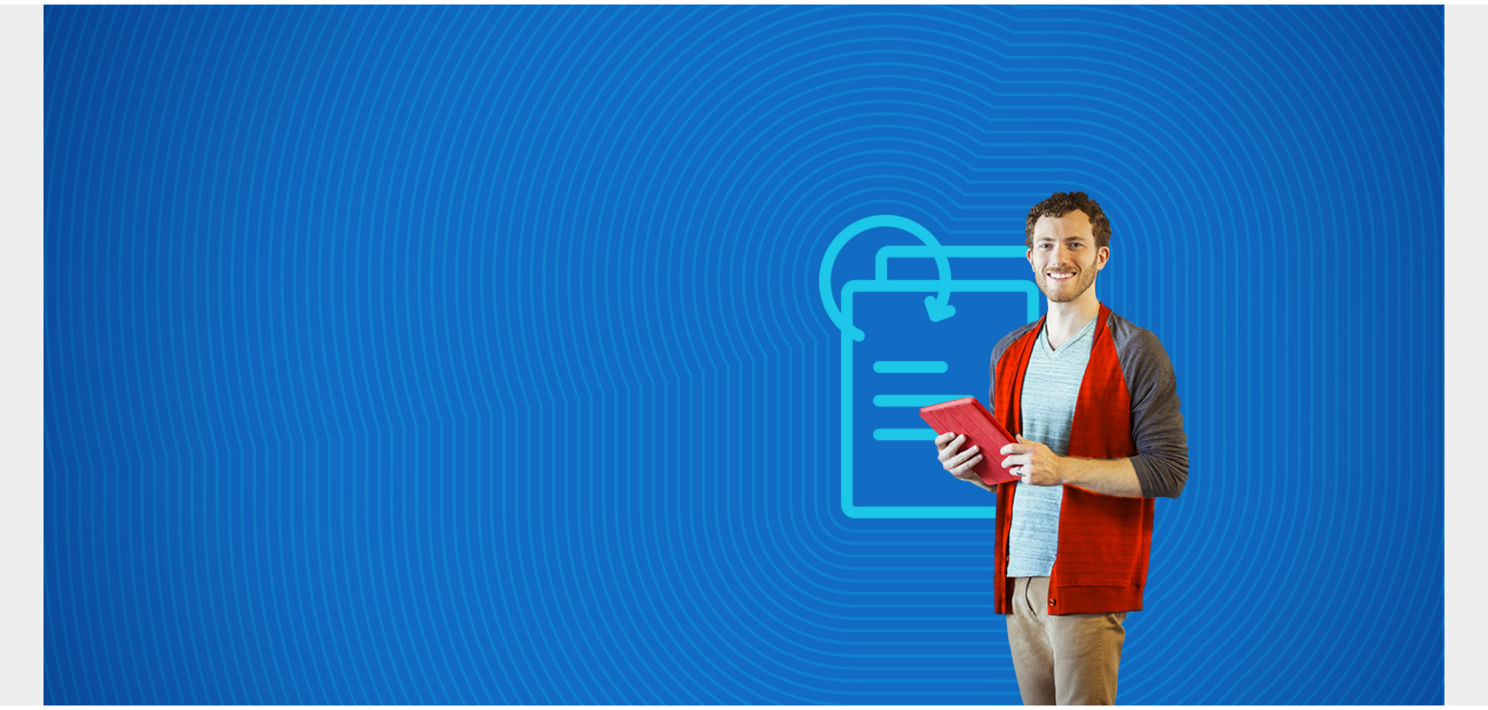
HOW TO IMPORT AMAZON S3 DATA TO SNOWFLAKE



In this article, we'll show how to load JSON data into Snowflake from Amazon S3.

Snowflake and JSON files

Snowflake is a data warehouse on AWS. The Snowflake **COPY** command lets you copy JSON, XML, CSV, Avro, Parquet, and XML format data files.

But to say that Snowflake supports JSON files is a little misleading—it does not parse these data files, as we showed in an example with Amazon Redshift. Instead, Snowflake copies the entirety of the data into one Snowflake column of type **variant**. Then you run JSON SQL queries against that.

We will load two data files, which you can download from here:

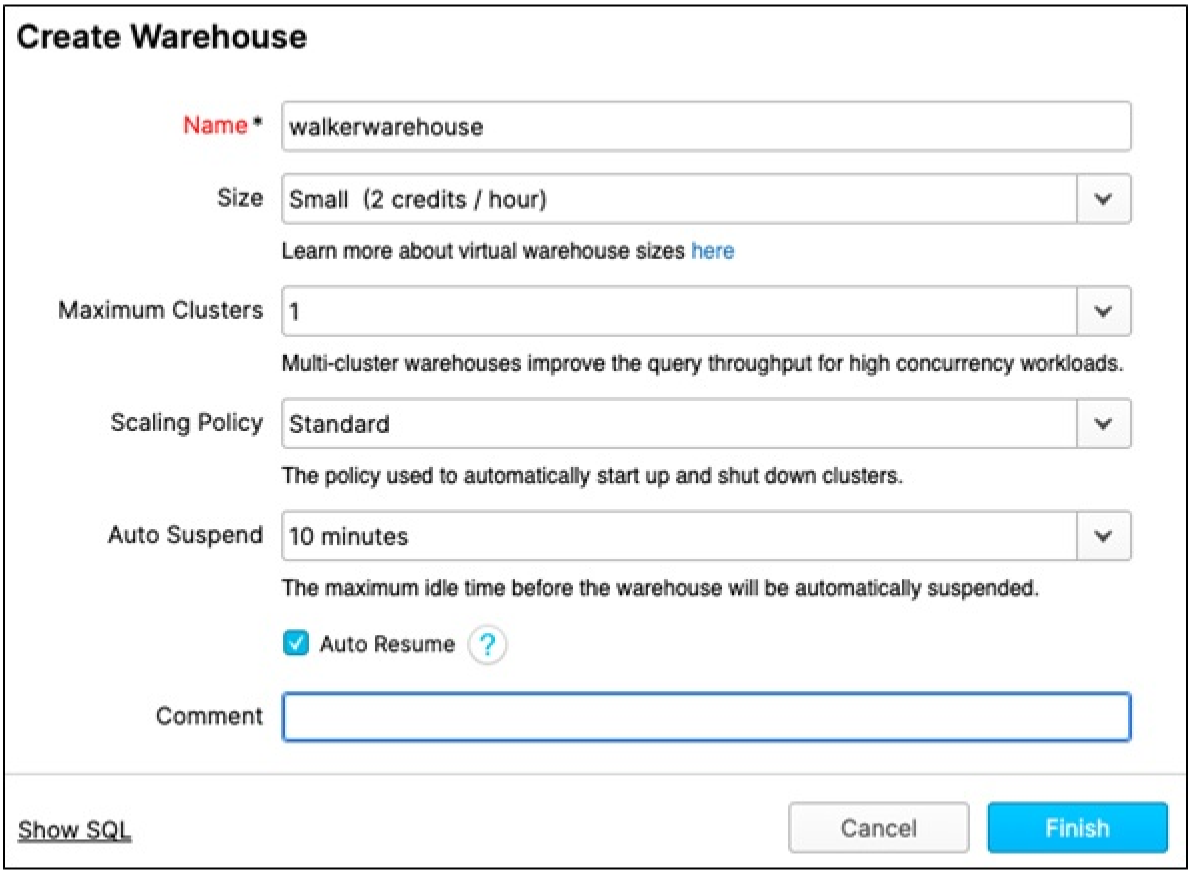
- Customers
- Orders

Note: This is in **NDJSON** format. That means the entire file is not a valid JSON file. Instead it is composed of individual JSON records.

Create database and warehouse

We are running our Snowflake cluster on Amazon AWS. (It is not listed as a service on the Amazon AWS Console. Instead you sign up for it on the Snowflake site, then it launches an instance on Amazon, Microsoft, or Google clouds.)

You create a warehouse like this. Here is where you pick the machine site and number of servers, thus picking the computing power and cost.



Create Warehouse

Name * walkerwarehouse

Size: Small (2 credits / hour)

Learn more about virtual warehouse sizes here

Maximum Clusters: 1

Multi-cluster warehouses improve the query throughput for high concurrency workloads.

Scaling Policy: Standard

The policy used to automatically start up and shut down clusters.

Auto Suspend: 10 minutes

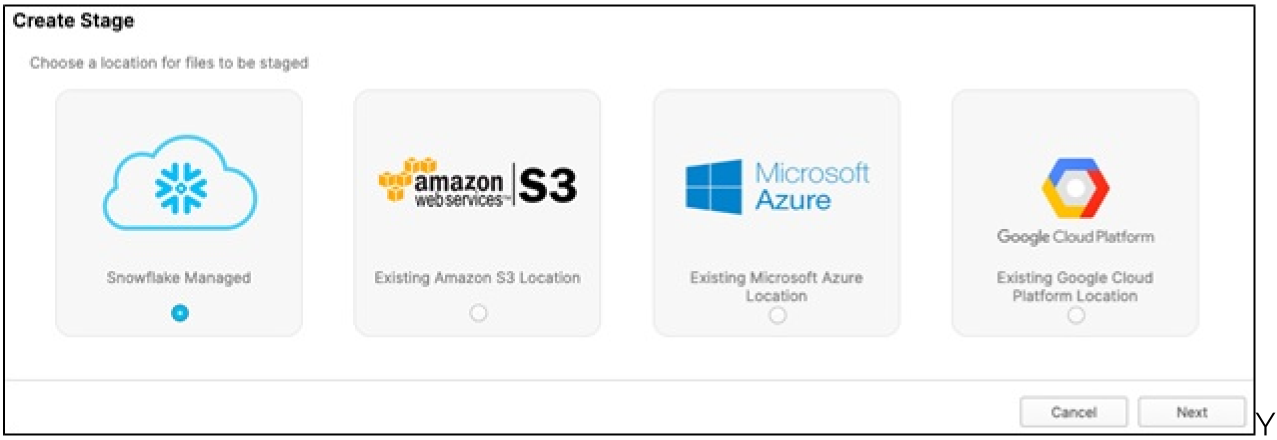
The maximum idle time before the warehouse will be automatically suspended.

☑ Auto Resume ?

Comment:

Create external stage

You can copy data directly from Amazon S3, but Snowflake recommends that you use their **external stage** area. They give no reason for this. But, doing so means you can store your credentials and thus simplify the copy syntax plus use wildcard patterns to select files when you copy them.



Create Stage

Choose a location for files to be staged

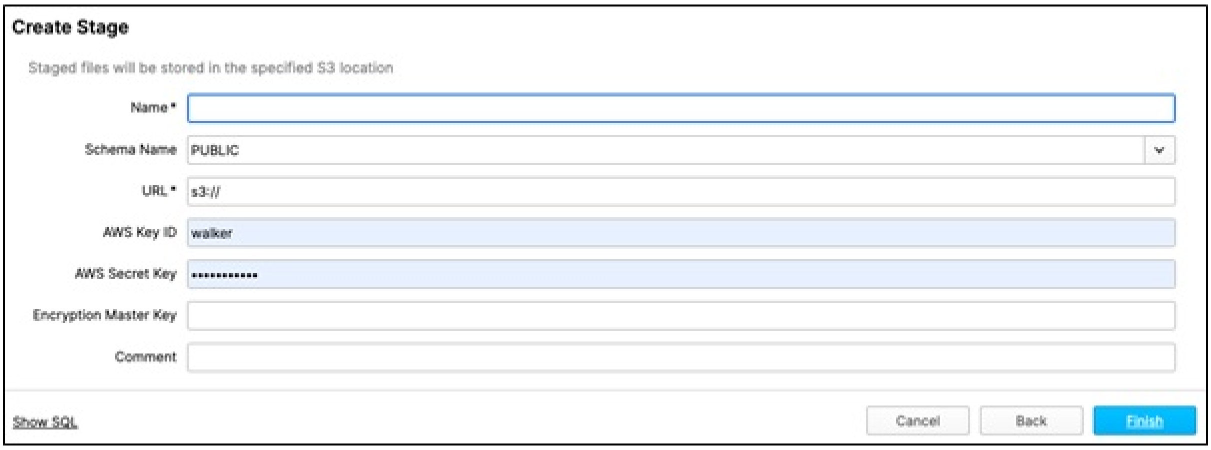
● Snowflake Managed

○ Existing Amazon S3 Location

○ Existing Microsoft Azure Location

○ Existing Google Cloud Platform Location

You give it a name and point it to an S3 bucket.



Create Stage

Staged files will be stored in the specified S3 location

Name *

Schema Name PUBLIC

URL * s3://

AWS Key ID walker

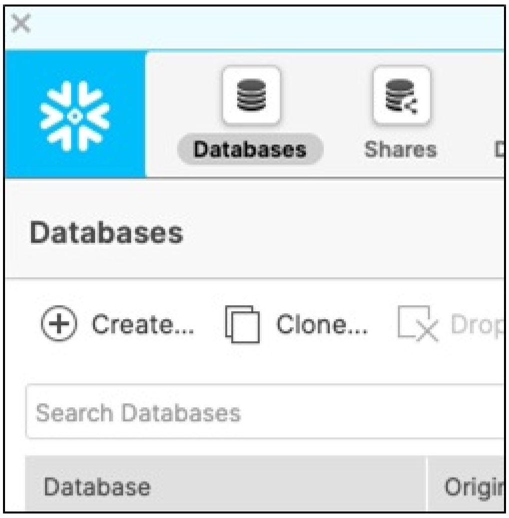
AWS Secret Key •••••••••••

Encryption Master Key

Comment

Create a database

As with other databases, a **database** is a collection of tables. So, it's just a name in this simple example. Create it from the Snowflake console like this:



Create tables

Next, open the worksheet editor and paste in this SQL to create the **customers** and **orders** tables. Note that both have one column of type **variant**.

```
use database inventory;

create table customers (customer variant);

create table orders(orders variant);
```

Upload JSON data to S3

Copy the data to S3 using the Amazon S3 console or AWS CLI command line:

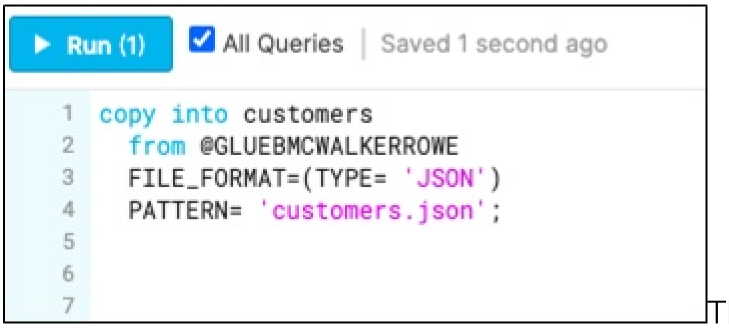
```
aws s3 cp customers.json s3:/(bucket name)

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

Bulk load the JSON data into Snowflake

Copy the customers and orders data into Snowflake like this. Since that S3 bucket contains both files we give the name using PATTERN. That can be any regular expression.

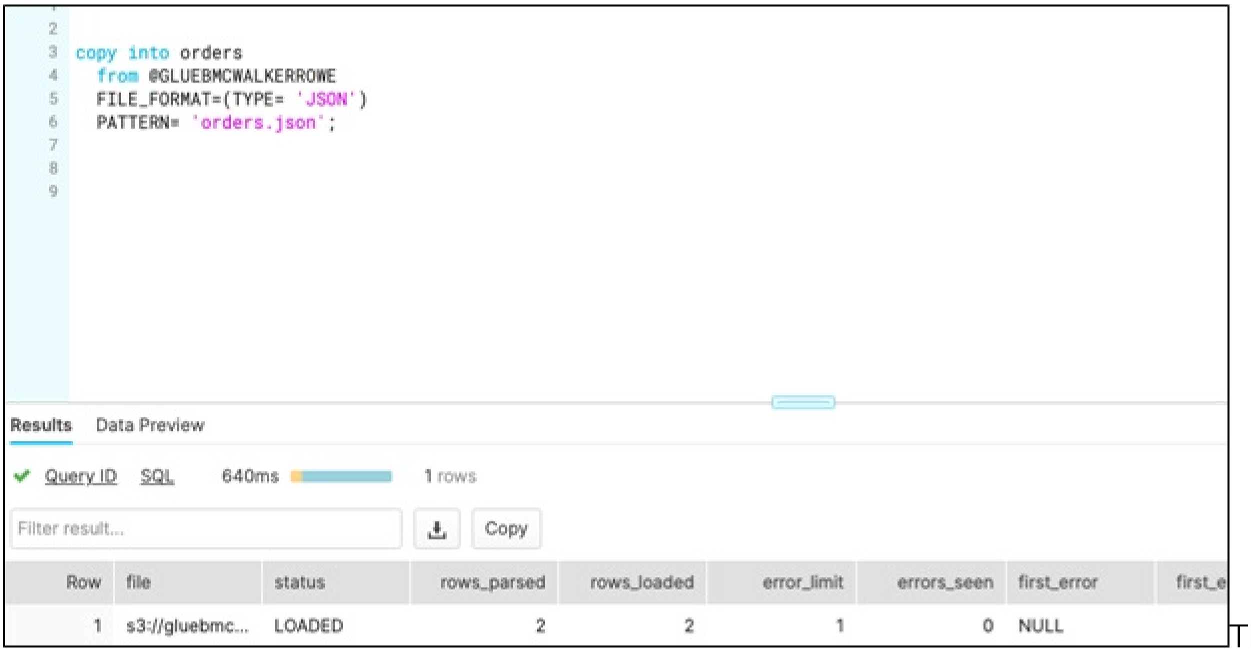
```
copy into customers
  from @GLUEBMCWALKERROWE
  FILE_FORMAT=(TYPE= 'JSON')
  PATTERN= 'customers.json';
```

One annoying feature is that you have to select the **All Queries** checkbox in order to enable the **run** button:



```
copy into customers
  from @GLUEBMCWALKERROWE
  FILE_FORMAT=(TYPE= 'JSON')
  PATTERN= 'customers.json';
```

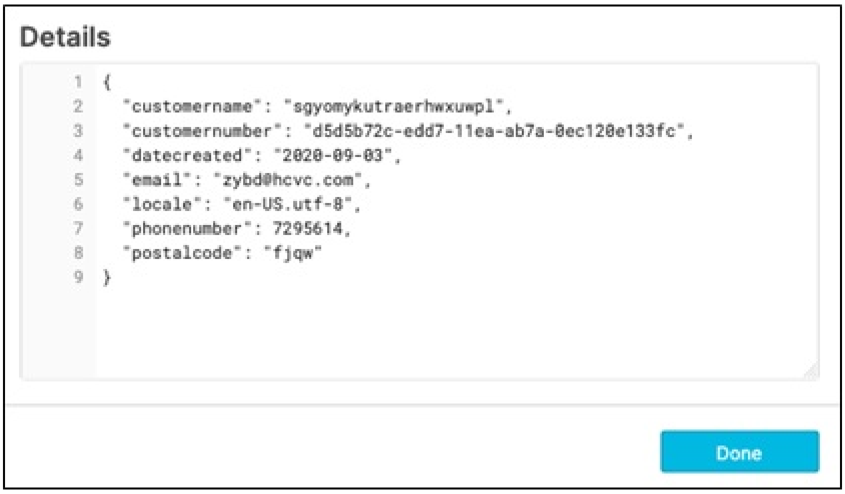
The worksheet looks like this when you execute the SQL.



```
copy into orders
  from @GLUEBMCWALKERROWE
  FILE_FORMAT=(TYPE= 'JSON')
  PATTERN= 'orders.json';
```

Row	file	status	rows_parsed	rows_loaded	error_limit	errors_seen	first_error	first_e
1	s3://gluebmc...	LOADED	2	2	1	0	NULL	

Then click on the **file** and see the data that it loaded.



```
{
  "customername": "sgyomykutraerhwxuwpl",
  "customernumber": "d5d5b72c-edd7-11ea-ab7a-0ec120e133fc",
  "datecreated": "2020-09-03",
  "email": "zybd@hcvc.com",
  "locale": "en-US.utf-8",
  "phonenumber": 7295614,
  "postalcode": "fjqw"
}
```

BMC, Control-M support Snowflake

BMC is a member of the Snowflake Technology Alliance Partner program. Snowflake's cloud data platform helps customers to accelerate the data-driven enterprise with Snowflake's market-leading, built-for-cloud data warehouse and Control-M, our market-leading enterprise application workflow orchestration platform.

(Learn how to integrate Snowflake with Control-M.)

Additional resources

For more tutorials like this, explore these resources:

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