

EPAM's Snowflake Hands-on Lab

Lab Overview

For those who begin to study Snowflake from scratch, it is recommended to start with “[Hands-On Lab Guide for Snowflake Free Trial](#)” that describes how to work with the main database features in the form of step-by-step guide.

This Lab (prepared by your EPAM colleagues) offers a high-level description of the practical task for self-directed learning.

The target group for the Lab are DWBI engineers with experience in building Data Warehouses using other databases (Oracle, MS SQL, Teradata, etc.).

Lab Data Set

Data set from [TPC-H benchmark](#) is proposed for the Lab. TPC-H allows you to generate data for 8 tables. The data volume (in gigabytes) is defined by scale factor (SF). For the Lab purpose, you can [download](#) prepared in advance data set (2 GB of raw data, SF=2):

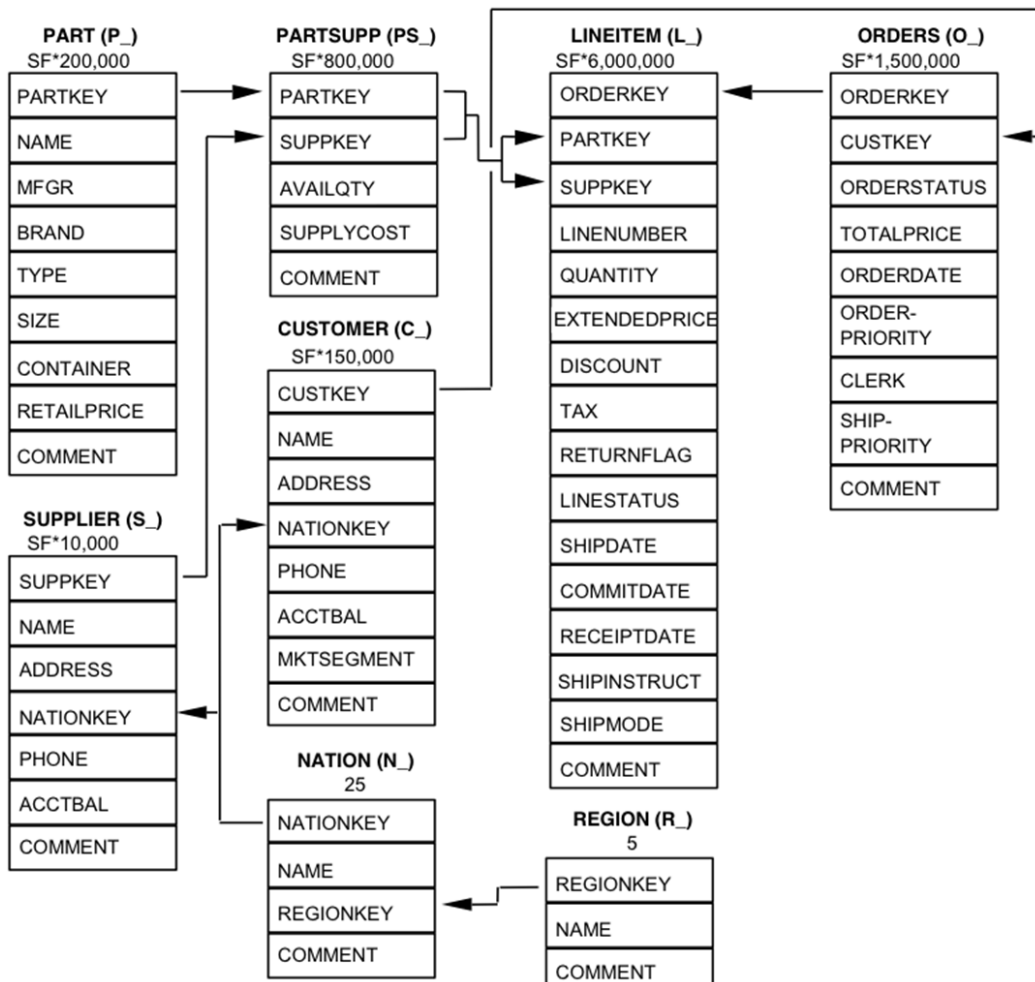


Table	Number of rows
H_LINEITEM	11 996 782
H_ORDER	3 000 000
H_PARTSUPP	1 600 000
H_PART	400 000
H_CUSTOMER	300 000
H_SUPPLIER	20 000
H_NATION	25
H_REGION	5

In the shared folder, you can also find DDL script for the tables: *tpch_ddl.sql*.

Lab Description

Hands-on-lab is considered as completed if you score ≥ 60 points.

(Tasks 1, 8 – 5 points each, Tasks 2, 4, 5, 6 – 10 points each, Task 3 – 30 points, Task 7 – 20 points).

1. Database creation

First, you need to create a separate database EPAM_LAB in Snowflake.

The screenshot shows the Snowflake console interface. At the top, there's a navigation bar with icons for Databases, Shares, Data Marketplace, Warehouses, Worksheets, and History. Below this, the 'Databases' section is active, showing a list of databases. The list includes columns for Database name, Origin, Creation Time, Owner, and Comment. The 'EPAM_LAB' database is highlighted in blue, indicating it is the current selection. Other databases listed include SOCIAL_MEDIA_FLOODGATES, LIBRARY_CARD_CATALOG, USDA_NUTRIENT_STDREF, SNOWFLAKE_SAMPLE_DATA, DEMO_DB, and UTIL_DB.

Database	Origin	Creation Time	Owner	Comment
EPAM_LAB		8/4/2021, 3:52 PM	SYSADMIN	
SOCIAL_MEDIA_FLOODGATES		7/28/2021, 2:49 PM	SYSADMIN	There's so much data from social media - flood warning
LIBRARY_CARD_CATALOG		7/28/2021, 10:45 AM	SYSADMIN	Essentials Lesson 9
USDA_NUTRIENT_STDREF		7/27/2021, 2:30 PM	SYSADMIN	Snowflake Lab
SNOWFLAKE_SAMPLE_DATA	SFC_SAMPLES.SA...	7/27/2021, 10:29 AM	ACCOUNTADMIN	TPC-H, OpenWeatherMap, etc
DEMO_DB		7/27/2021, 10:29 AM	SYSADMIN	demo database
UTIL_DB		7/27/2021, 10:29 AM	SYSADMIN	utility database

2. Data loading

In this step, you need to load Lab data set to internal (Snowflake) or external stage. If you have an existing account in AWS/GCP/Azure cloud, external stage would be preferable. Please note that you may need some data preparation steps before loading.

← → ↻ wpa11502.snowflakecomputing.com/console#/data/tables?databaseName=EPAM_LAB

Enjoy your free trial! Visit our [documentation](#) to learn more about using Snowflake or [contact our support](#)

Databases > EPAM_LAB

Tables Views Schemas Stages File Formats Sequences Pipes

+ Create... + Create Like... Clone... Load Data... Drop... Transfer Ownership

Table Name	Schema	Creation Time ▼	Owner	Rows	Size
ORDERS_WF	CORE_DWH	8/9/2021, 12:16:24 ...	SYSADMIN	3M	65.9MB
NATION_WF	CORE_DWH	8/9/2021, 12:16:23 ...	SYSADMIN	25	2.5KB
REGION_WF	CORE_DWH	8/9/2021, 12:16:23 ...	SYSADMIN	5	1.5KB
CUSTOMER_WF	CORE_DWH	8/9/2021, 12:14:30 ...	SYSADMIN	300K	9.9MB
PARTSUPP	TPCH	8/4/2021, 5:14:41 PM	SYSADMIN	1.6M	32.6MB
ORDERS	TPCH	8/4/2021, 5:10:13 PM	SYSADMIN	3M	85.0MB
LINEITEM	TPCH	8/4/2021, 4:58:02 P...	SYSADMIN	12.0M	348.0MB
CUSTOMER	TPCH	8/4/2021, 4:37:40 P...	SYSADMIN	300K	14.0MB
PART	TPCH	8/4/2021, 3:53:15 PM	SYSADMIN	400K	9.0MB
NATION	TPCH	8/4/2021, 3:53:14 PM	SYSADMIN	25	2.5KB
REGION	TPCH	8/4/2021, 3:53:14 PM	SYSADMIN	5	1.5KB

The data load was done using SnowSQL, in the following file you will find the code for loading through the internal stage.

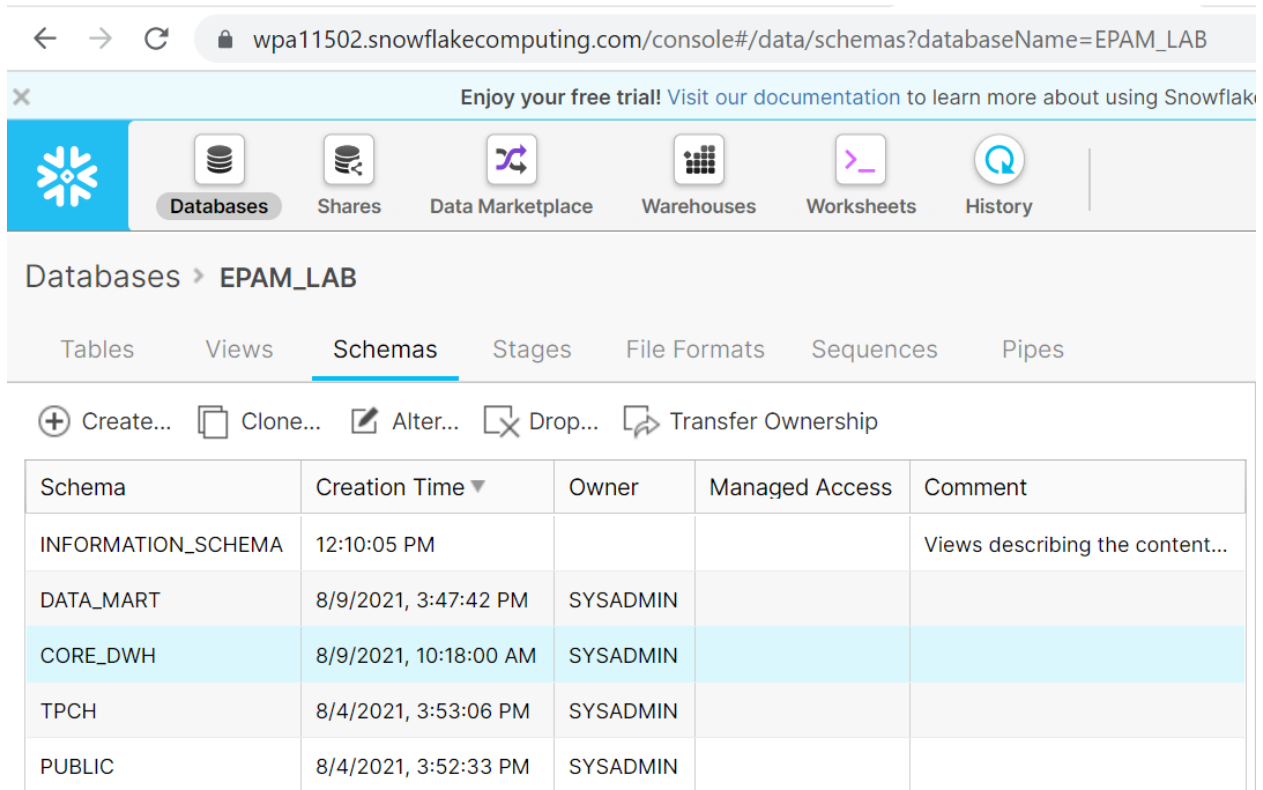


Data Load Point 2.txt

3. ELT Data Workflow

Create two schemas in the DB you created before:

- CORE_DWH
- DATA_MART



Schema	Creation Time ▼	Owner	Managed Access	Comment
INFORMATION_SCHEMA	12:10:05 PM			Views describing the content...
DATA_MART	8/9/2021, 3:47:42 PM	SYSADMIN		
CORE_DWH	8/9/2021, 10:18:00 AM	SYSADMIN		
TPCH	8/4/2021, 3:53:06 PM	SYSADMIN		
PUBLIC	8/4/2021, 3:52:33 PM	SYSADMIN		

Develop the following automated data workflow:

Stage -> CORE_DWH -> DATA_MART

Data in CORE_DWH should be modeled according to 3NF (as is - no transformation). Star Schema is a target data model for DATA_MART (data should be transformed accordingly).

The following Snowflake features should be used:

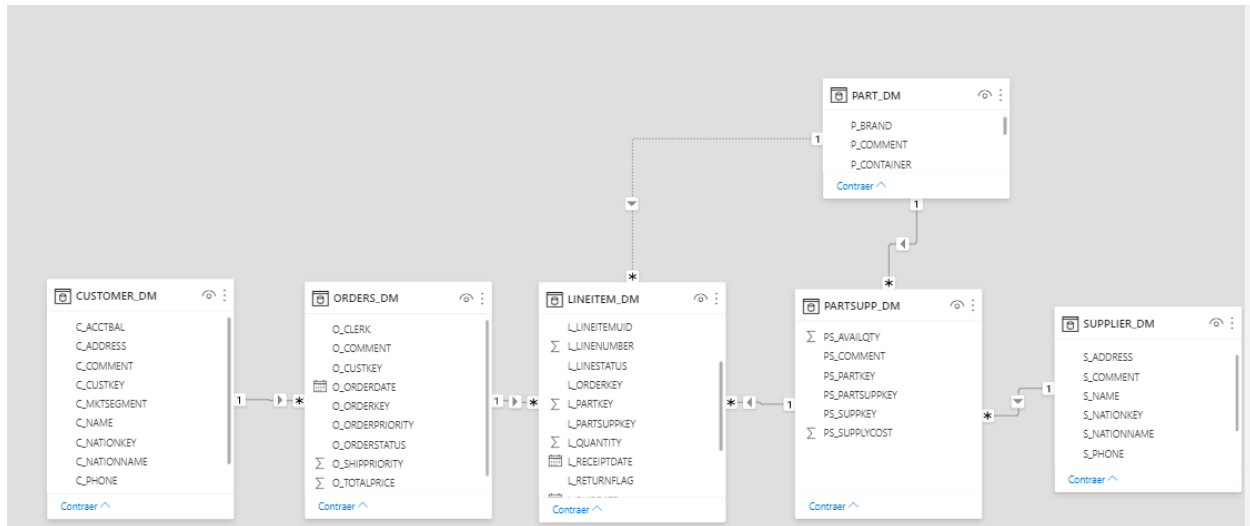
- Orchestration Tasks
- Stored Procedures
- Tables Streams

The dataflow was created, I was able to create a simple stored procedure due to the fact that my knowledge in JavaScript is fairly limited. Here you can see the code used:



Data Load - WF.txt

And here's the final Star Schema created, the image is taken from the Power BI service:



4. Snowflake & 3rd party tools

When the data is loaded to DATA_MART schema, connect Snowflake as a data source from any BI tool (Tableau, PowerBI, Qlik Sense, etc.) and create a simple dashboard.

Also, try connecting to Snowflake from any SQL editor (e.g. [DBeaver](#)).

The following files include a .pbix file and a .pdf:



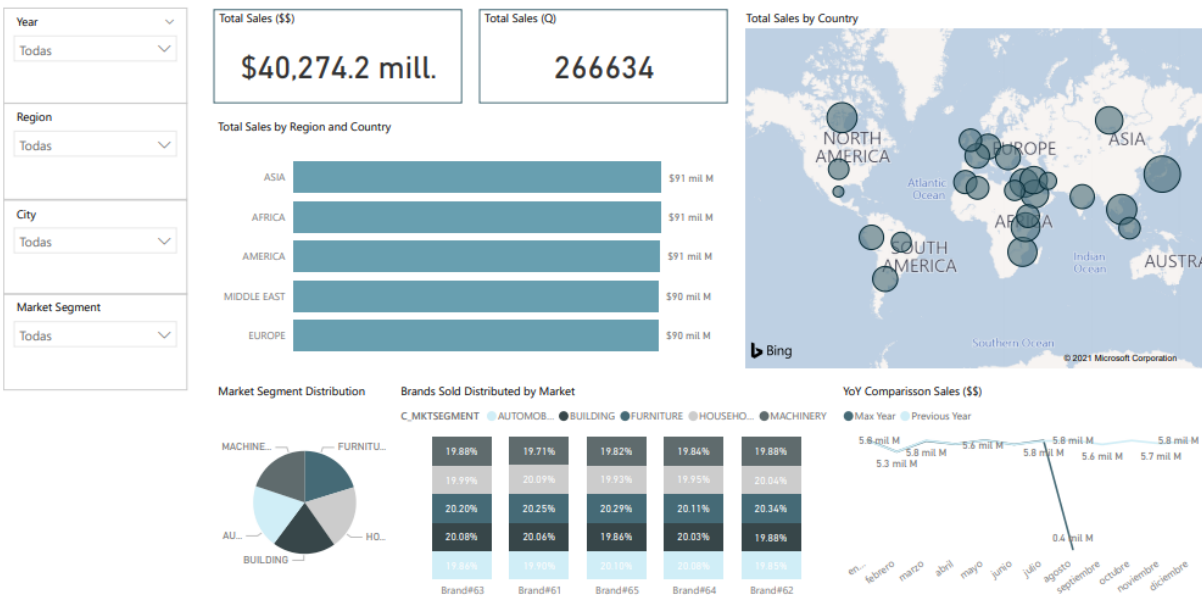
EPAM_LAB_DASHBOARD.pdf



EPAM_LAB_DASHBOARD.pbix



EPAM Snowflake Lab



5. Snowflake SQL

From the shared folder you can also [download](#) the file with 22 TPC-H benchmarking queries (tpch_benchmark_queries.sql). Please note that the queries were modified to execute in AWS RedShift database, so some of them may require modifications for Snowflake. Use the queries to test how Snowflake works:

- Create several warehouses of different sizes and compare their performance;
- Test how Snowflake leverages different types of caches;
- Rewrite a couple of queries to execute on Star Schema data model and compare performance (3NF vs Star Schema);
- Execute queries using SnowSQL (CLI Client).

Data load code using SnowSQL.



Data Load Point 2.txt

Screen shot of the SnowSQL client connected

```
Command Prompt - snowsql -a WPA11502 -u julianadiaz

--disable-request-pooling    if there is no activity from the user..
                             Disable request pooling. This can help speed
                             up connection failover
--token TEXT                 The token to be used with oauth
                             authentication method
-?, --help                   Show this message and exit.

C:\Users\Juliana_Diaz>snowsql -a WPA wpa11502 -u julianadiaz
Got unexpected extra argument (wpa11502)
Try "snowsql --help" for more information.

C:\Users\Juliana_Diaz>snowsql -a WPA11502 -u julianadiaz
Failed to initialize log. No logging is enabled: [Errno 13] Permission denied: 'C:\\Users\\snowsql_rt.log'
Password:
250001 (08001): Failed to connect to DB: WPA11502.snowflakecomputing.com:443. Incorrect username or password was specified.
If the error message is unclear, enable logging using -o log_level=DEBUG and see the log to find out the cause. Contact support for further help.
Goodbye!

C:\Users\Juliana_Diaz>snowsql -a WPA11502 -u julianadiaz
Failed to initialize log. No logging is enabled: [Errno 13] Permission denied: 'C:\\Users\\snowsql_rt.log'
Password:
* SnowSQL * v1.2.9
Type SQL statements or !help
julianadiaz#COMPUTE_WH@no database).(no schema)>use DATABASE
                                     EPAM_LAB;

+-----+
| status |
+-----+
| Statement executed successfully. |
+-----+
1 Row(s) produced. Time Elapsed: 0.247s
julianadiaz#COMPUTE_WH@EPAM_LAB.PUBLIC>
```

Queries:

```

juliandiaz@COMPUTE_WH@EPAM_LAB.TPCH>Select * from nation;

```

N_NATIONKEY	N_NAME	N_REGIONKEY	N_COMMENT
0	ALGERIA	0	packages sleep .
1	ARGENTINA	1	quickly final instructions wake alongside of .
2	BRAZIL	1	carefully ironic ideas after affix quickly above .
3	CANADA	1	packages cajole carefully furiously even pinto beans .
4	EGYPT	4	blithely ironic pinto beans along haggle carefully ruthlessly special Ti
5	ETHIOPIA	0	regular, ironic deposits across wake after .
6	FRANCE	3	quickly even platelets among sleep about .
7	GERMANY	3	packages about use blithely furiously regular ideas .
8	INDIA	2	blithely express pinto beans along use blithely packages .
9	INDONESIA	2	final, ironic deposits poach ruthlessly across :
10	IRAN	4	express, silent deposits cajole carefully ironic pinto beans .
11	IRAQ	4	blithely final theodolites haggle carefully against .
12	JAPAN	2	quickly express platelets integrate quickly .
13	JORDAN	4	packages sleep about .
14	KENYA	0	regular, bold deposits sleep .
15	MOROCCO	0	fluffily bold dolphins haggle carefully quickly regular instructions .
16	MOZAMBIQUE	0	silent accounts use blithely according to .
17	PERU	1	quickly bold instructions sleep alongside of .
18	CHINA	2	blithely furious theodolites cajole quickly bold instructions .
19	ROMANIA	3	evenly bold pains sleep special, ironic deposits .
20	SAUDI ARABIA	4	even accounts could cajole .
21	VIETNAM	2	blithely even instructions use blithely .
22	RUSSIA	3	blithely silent pinto beans nag blithely .
23	UNITED KINGDOM	3	blithely regular theodolites mold slowly :
24	UNITED STATES	1	regular accounts was quickly even, express deposits .

```

juliandiaz@COMPUTE_WH@EPAM_LAB.TPCH> SELECT * FROM ORDERS LIMIT 4;

```

O_ORDERKEY	O_CUSTKEY	O_ORDERSTATUS	O_ORDERPRIORITY	O_CLERK	O_SHIPPRIORITY	O_COMMENT	O_TOTALPRICE	O_ORDERDATE
8236871	143534	F	3-MEDIUM	Clerk#000001814	0	carefully ironic foxes haggle carefully after .	40297.3235	1995-05-1
8236903	81907	P	3-MEDIUM	Clerk#000000951	0	pending requests above cajole furiously bold pint	164358.9028	1995-05-1
8236930	217568	P	4-NOT SPECIFIED	Clerk#000000399	0	carefully brave ideas sleep .	290492.3663	1995-05-1
8236997	284752	P	4-NOT SPECIFIED	Clerk#000001273	0	quickly final dependencies wake bold accounts .	90221.7012	1995-05-2

4 Row(s) produced. Time Elapsed: 2.475s

```

juliandiaz@COMPUTE_WH@EPAM_LAB.TPCH>SELECT SUM (O_TOTALPRICE) AS PRICEPERPRIORITY,
O_ORDERPRIORITY
FROM ORDERS
GROUP BY O_ORDERPRIORITY;

```

PRICEPERPRIORITY	O_ORDERPRIORITY
90825667127.2229	3-MEDIUM
90713197066.928	1-URGENT
90526540370.2413	2-HIGH
90640568676.5932	5-LOW
90683397435.5099	4-NOT SPECIFIED

5 Row(s) produced. Time Elapsed: 1.690s

```

juliandiaz@COMPUTE_WH@EPAM_LAB.TPCH>

```

6. Other Snowflake features

Learn and test other interesting Snowflake features:

- Object Cloning;

```

164 CREATE TABLE lineitem_dm CLONE "EPAM_LAB"."CORE_DWH"."LINEITEM_WF";
165 |
166 ALTER TABLE "EPAM_LAB"."DATA_MART"."LINEITEM_DM"
167 ADD L_EXTENDEDPRICE1 FLOAT8,
168     L_DISCOUNT1 FLOAT8,
169     L_TAX1 FLOAT8,
170     L_SHIPDATE1 DATE,
171     L_COMMITDATE1 DATE,
172     L_RECEIPTDATE1 DATE,
173     L_PARTSUPPKEY VARCHAR,
174     L_LINEITEMUID VARCHAR;

```

- Time Travel;

```
select * from "EPAM_LAB"."DATA_MART"."PARTSUPP_DM" at(offset => -60*5);
select * from "EPAM_LAB"."DATA_MART"."PARTSUPP_DM" at(timestamp => 'Mon, 16 August 2021 16:20:00 -0500'::timestamp_tz);
```

ts Data Preview Open History

Query ID SQL 4.58s 1,600,000 rows

result... Download Copy Columns

Row	PS_PARTKEY	PS_SUPPKEY	PS_AVAILQTY	PS_COMMENT	PS_SUPPLYCOST	PS_PARTSUPPKEY
1	191111	6139	506	quickly final instructi...	51.49	191111-6139
2	191113	11114	1473	packages are about .	148.16	191113-11114

- Data Sharing - share your DATA_MART schema with a colleague who helps you with this Lab.

7. Snowpipe

Automated incremental data loading using Snowpipe. Split lineitem & order files into several parts and simulate their sequential loading to stage buckets.

```
12 //STAGE CREATION
13
14 CREATE OR REPLACE STAGE Snowpipe_Epam
15 url='s3://snowpipe-epamlab'
16 credentials=(aws_key_id='AKIASDGNXNS2OH2BXMKP' aws_secret_key='pVzjY6LhzW5Xw8D38R1sx4pKEbQzYaZvK8aQW/P')
17 file_format=DSV;
18
19 SHOW STAGES;
20
21 //PIPE CREATION
22 CREATE OR REPLACE PIPE epam_lab_pipe auto_ingest=true AS
23 COPY INTO snowpipe.nation
24 FROM @Snowpipe_Epam
25 file_format=DSV;
26
27 SHOW PIPES;
28
29 list @snowpipe_Epam;
30
31 alter pipe epam_lab_pipe refresh;
32
33 SELECT COUNT (*) FROM nation;
```

results Data Preview Open History

Query ID SQL 44ms 1 rows

filter result... Download Copy Columns

Row	COUNT (*)
1	25

8. Additional tasks

Connect your Snowflake account with partner applications available for a free trial (e.g. Fivetran, Periscope Data, Matillion in Partner Connect menu). Explore how selected tools work.

My account is currently connected to SnapLogic, in the following JSON snap we can see the table Region loaded through Snaps:

The screenshot shows the 'Snowflake - Bulk Load output0' window. The top bar includes 'Designer', 'Manager', and 'Dashboard' tabs. The left sidebar contains a search bar and a list of active directories. The main content area displays a JSON tree view of the bulk load output.

Preview Type	Indent Level	Expand Level
JSON	2	1+

```

{
  "table": "'PUBLIC','REGIONSL'",
  "input_records": 0,
  "results": {
    {
      "file": "s3://snowpipe-epamlab/h_region.csv",
      "status": "LOAD_SKIPPED",
      "rows_parsed": "0",
      "rows_loaded": "0",
      "error_limit": null,
      "errors_seen": "1",
      "first_error": "File was loaded before.",
      "first_error_line": null,
      "first_error_character": null,
      "first_error_column_name": null
    }
  }
}
  
```