SanKir Technologies https://www.sankir.com

email - info@sankir.com

❄

POC on retail data using Snowflake on AWS, SnowSQL

POC Objective

- Loading data from AWS S3 files into Snowflake tables
- Automate SQL actions like DDL & DML using SnowSQL
- Creation of Snowflake Objects like Database, Warehouse and Table
- Load data from S3 to Snowflake Tables using COPY INTO command

---

About retail data set

This is a transactional data set which contains all the transactions for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Data set path in S3

s3://retail-sankir/data/retail_data/q1 There are 75 data files.

The schema for the data is defined in sql script file.

SnowSQL installer

The SnowSQL installer on Windows ver 1.2.23 ver can be downloaded from the Snowflake Client Repository.

https://sfc-repo.snowflakecomputing.com/snowsql/bootstrap/1.2/windows_x86_64/index.html

Here is the snowsql command to get the interactive shell.

```
snowsql -a wh98438.us-east-2.aws -u sankir --variable SNOWSQL_PWD=%SNOWSQL_PWD%
```

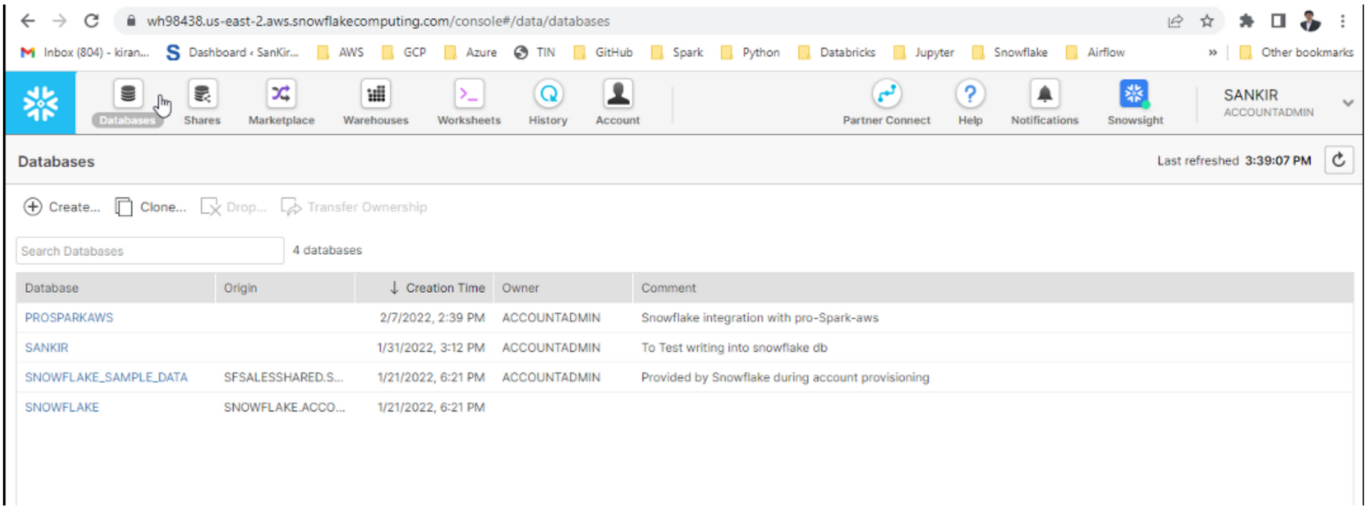
This is part of your snowflake account URL for AWS.

snowflake account password is to be provided as authentication option.

Add SNOWSQL_PWD as system env variable and set the password.

Snowflake Objects in AWS account before running the sql script

- The usecase is Loading data from files in AWS S3 into Snowflake tables.
- Initially, there is no database or schema in our snowflake account related to this usecase.
- We will create database, schema, table, Stage and File format through sql script file.
- Stages in Snowflake are locations used to store data. Here it is called as External Stage as data is stored in AWS S3.



wh98438.us-east-2.aws.snowflakecomputing.com/console#/data/databases

Databases

Last refreshed 3:39:07 PM

4 databases

Database	Origin	↓ Creation Time	Owner	Comment
PROSPARKAWS		2/7/2022, 2:39 PM	ACCOUNTADMIN	Snowflake integration with pro-Spark-aws
SANKIR		1/31/2022, 3:12 PM	ACCOUNTADMIN	To Test writing into snowflake db
SNOWFLAKE_SAMPLE_DATA	SFSALESSHARED.S...	1/21/2022, 6:21 PM	ACCOUNTADMIN	Provided by Snowflake during account provisioning
SNOWFLAKE	SNOWFLAKE.ACCO...	1/21/2022, 6:21 PM		

Commands in SnowSQL script

SnowSQl script contains Commands to :

- Create database, schema and warehouse
- Create Table with schema
- Create File format and Set the Delimiters and set skip_header =1 as we have header row in data file
- Create Satge. Since we access data in AWS S3, access credentials namely , AWS_KEY_ID & AWS_SECRET_KEY are set
- Load the content of files in AWS S3 into snowflake table using COPY INTO

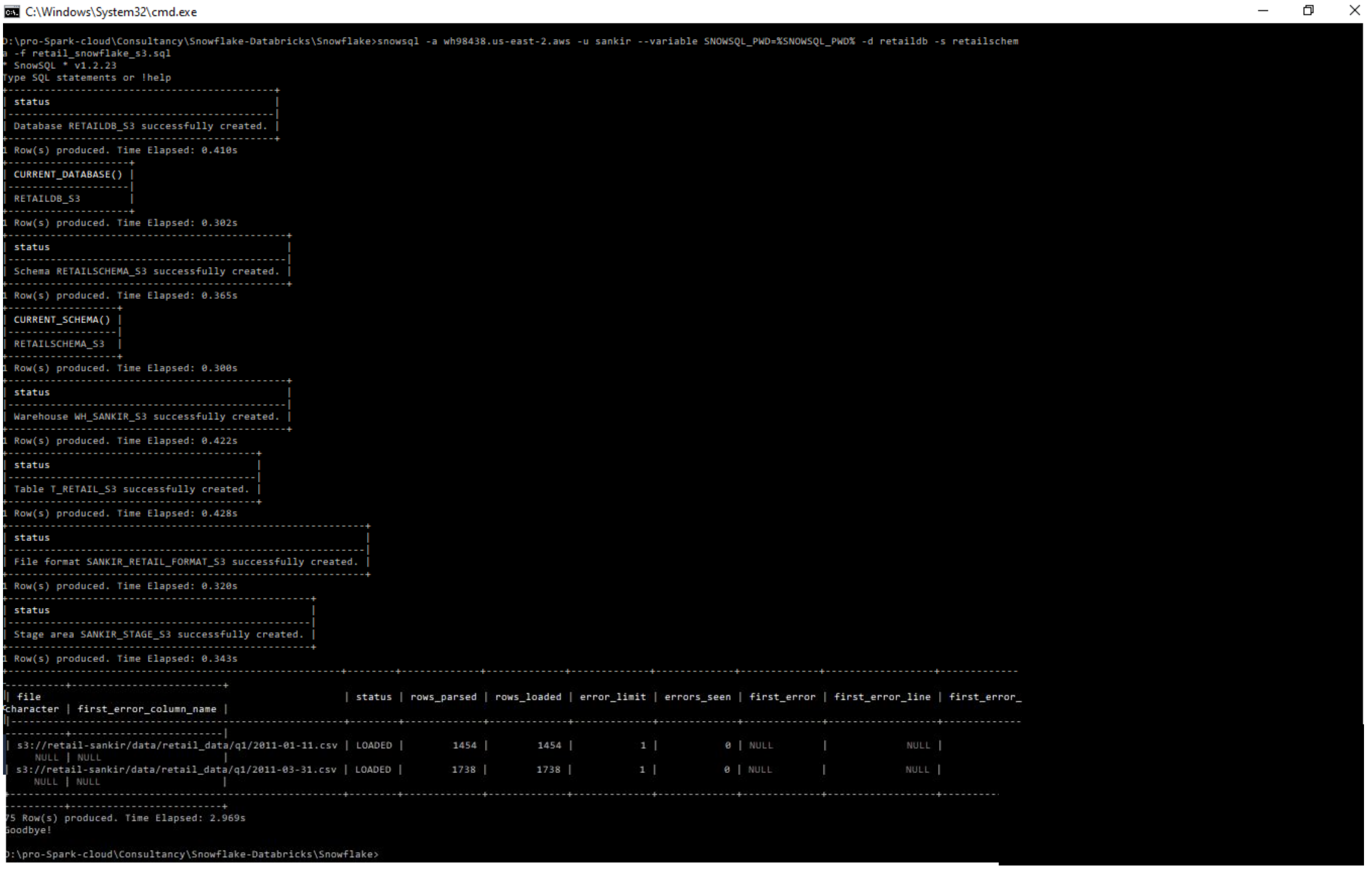
SnowSQL script

```
CREATE OR REPLACE DATABASE retaildb_s3;
SELECT CURRENT_DATABASE();

CREATE OR REPLACE SCHEMA retailschema_s3;
SELECT CURRENT_SCHEMA();

CREATE OR REPLACE WAREHOUSE wh_sankir_s3 with warehouse_size = 'X-SMALL'
auto_suspend = 180 auto_resume=true initially_suspended = true;

CREATE OR REPLACE TABLE retailschema_s3.t_retail_s3 (
    InvoiceNo varchar(255),
    StockCode varchar(255),
    Description varchar(255),
    Quantity number(10),
    InvoiceDate date,
    UnitPrice number(10),
    CustomerID number(10),
    Country varchar(255)
    );

CREATE OR REPLACE FILE FORMAT SANKIR_RETAIL_FORMAT_S3
FIELD_DELIMITER = ','
RECORD_DELIMITER = '\n'
SKIP_HEADER = 1
FIELD_OPTIONALLY_ENCLOSED_BY = '"';

CREATE OR REPLACE STAGE "RETAILDB_S3"."RETAILSCHEMA_S3".sankir_stage_s3
file_format = sankir_retail_format_s3
URL = 's3://retail-sankir/data/retail_data/q1'
CREDENTIALS = (AWS_KEY_ID = '***' AWS_SECRET_KEY = '***');

COPY INTO retaildb_s3.retailschema_s3.t_retail_s3
  from @sankir_stage_s3
  file_format = (format_name = sankir_retail_format_s3)
  pattern = '.*.csv'
  on_error = 'skip_file';
```

Run the SnowSQL script

The script runs succesfully and it shows creation of database, schema, file foramt and stage in Snowflake aacount.

Snowflake Table is created and loaded with data.



```
D:\pro-Spark-cloud\Consultancy\Snowflake-Databricks\Snowflake>snowsql -a wh98438.us-east-2.aws -u sankir --variable SNOWSQL_PWD=%SNOWSQL_PWD% -d retaildb -s retailschem
a -f retail_snowflake_s3.sql
* SnowSQL * v1.2.23
Type SQL statements or !help
+---------------------------------------------+
| status                                      |
|---------------------------------------------|
| Database RETAILDB_S3 successfully created.  |
+---------------------------------------------+
1 Row(s) produced. Time Elapsed: 0.410s
+--------------------+
| CURRENT_DATABASE() |
|--------------------|
| RETAILDB_S3        |
+--------------------+
1 Row(s) produced. Time Elapsed: 0.302s
+-----------------------------------------------+
| status                                        |
|-----------------------------------------------|
| Schema RETAILSCHEMA_S3 successfully created.  |
+-----------------------------------------------+
1 Row(s) produced. Time Elapsed: 0.365s
+------------------+
| CURRENT_SCHEMA() |
|------------------|
| RETAILSCHEMA_S3  |
+------------------+
1 Row(s) produced. Time Elapsed: 0.300s
+-----------------------------------------------+
| status                                        |
|-----------------------------------------------|
| Warehouse WH_SANKIR_S3 successfully created.  |
+-----------------------------------------------+
1 Row(s) produced. Time Elapsed: 0.422s
+----------------------------------------+
| status                                 |
|----------------------------------------|
| Table T_RETAIL_S3 successfully created. |
+----------------------------------------+
1 Row(s) produced. Time Elapsed: 0.428s
+-------------------------------------------------------------+
| status                                                      |
|-------------------------------------------------------------|
| File format SANKIR_RETAIL_FORMAT_S3 successfully created.   |
+-------------------------------------------------------------+
1 Row(s) produced. Time Elapsed: 0.320s
+----------------------------------------------------+
| status                                             |
|----------------------------------------------------|
| Stage area SANKIR_STAGE_S3 successfully created.   |
+----------------------------------------------------+
1 Row(s) produced. Time Elapsed: 0.343s
+---------------------------------------------------------+--------+-------------+------------+-------------+-------------+-------------+-----------------+-------------
-----------+------------------------+
|| file                                                    | status | rows_parsed | rows_loaded | error_limit | errors_seen | first_error | first_error_line | first_error_
character | first_error_column_name |
||---------------------------------------------------------+--------+-------------+------------+-------------+-------------+-------------+-----------------+-------------
-----------+------------------------|
| s3://retail-sankir/data/retail_data/q1/2011-01-11.csv | LOADED |        1454 |        1454 |           1 |           0 | NULL        |            NULL |
    NULL | NULL                   |
| s3://retail-sankir/data/retail_data/q1/2011-03-31.csv | LOADED |        1738 |        1738 |           1 |           0 | NULL        |            NULL |
    NULL | NULL                   |
+---------------------------------------------------------+--------+-------------+------------+-------------+-------------+-------------+-----------------+---------
-----------+------------------------+
75 Row(s) produced. Time Elapsed: 2.969s
Goodbye!

D:\pro-Spark-cloud\Consultancy\Snowflake-Databricks\Snowflake>
```

Snowflake Objects in AWS account after running the sql script

After The script is run, following Snowflake obejcts are acreated :

- RETAILDB_S3 database
- RETAILSCHEMA_S3 schema
- T_RETAIL_S3 table

The table content are shown here.

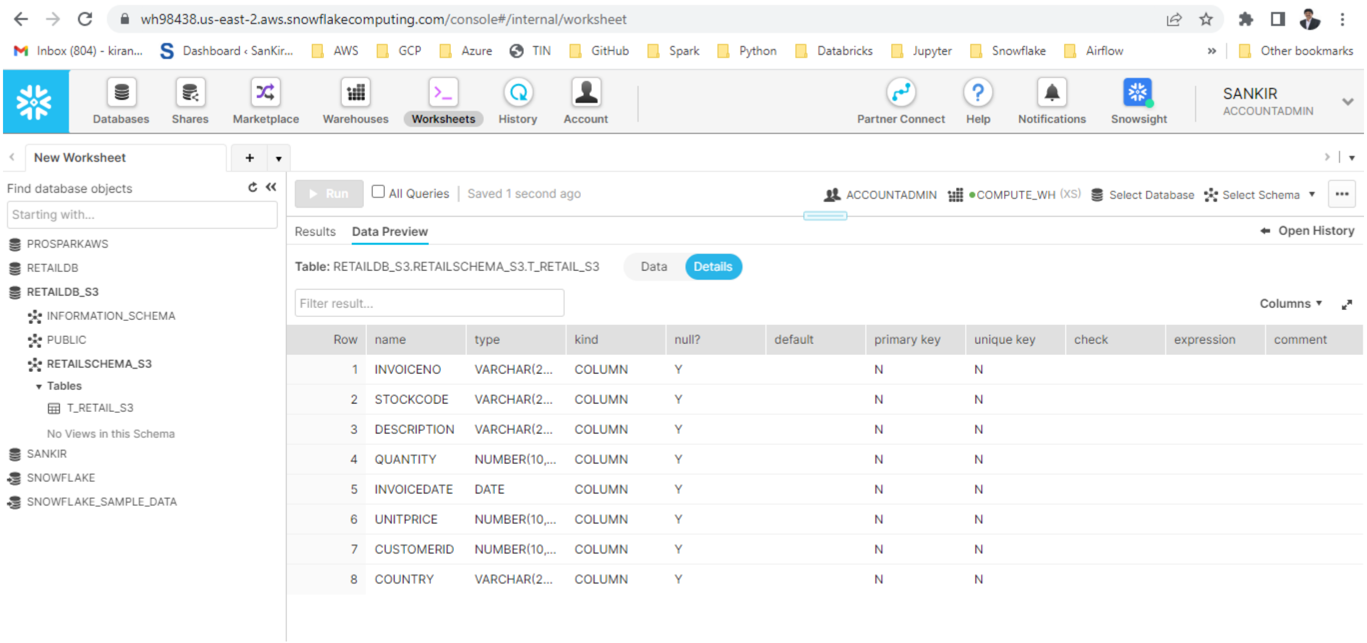
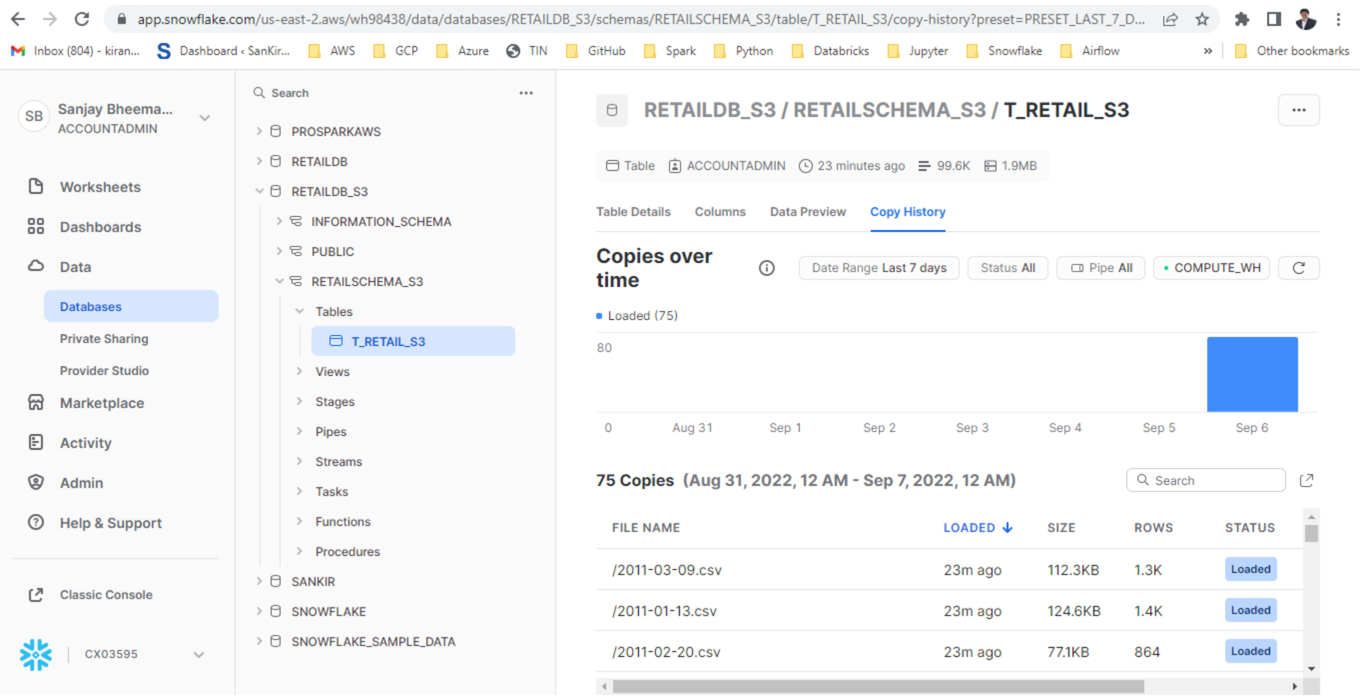


Table: RETAILDB_S3.RETAILSCHEMA_S3.T_RETAIL_S3

Row	name	type	kind	null?	default	primary key	unique key	check	expression	comment
1	INVOICENO	VARCHAR(2...	COLUMN	Y		N	N			
2	STOCKCODE	VARCHAR(2...	COLUMN	Y		N	N			
3	DESCRIPTION	VARCHAR(2...	COLUMN	Y		N	N			
4	QUANTITY	NUMBER(10,...	COLUMN	Y		N	N			
5	INVOICEDATE	DATE	COLUMN	Y		N	N			
6	UNITPRICE	NUMBER(10,...	COLUMN	Y		N	N			
7	CUSTOMERID	NUMBER(10,...	COLUMN	Y		N	N			
8	COUNTRY	VARCHAR(2...	COLUMN	Y		N	N			

Check copy history using Snowsight

Snowsight confirms 75 copies (files) are loaded to Snowflake table.



RETAILDB_S3 / RETAILSCHEMA_S3 / T_RETAIL_S3

Copies over time

75 Copies (Aug 31, 2022, 12 AM - Sep 7, 2022, 12 AM)

FILE NAME	LOADED	SIZE	ROWS	STATUS
/2011-03-09.csv	23m ago	112.3KB	1.3K	Loaded
/2011-01-13.csv	23m ago	124.6KB	1.4K	Loaded
/2011-02-20.csv	23m ago	77.1KB	864	Loaded

Technologies leveraged in POC

- snowSQL installer v1.2.23
- AWS S3 Storage account
- Snowflake Objects like Database, Warehouse and Table
- Define File format
- Usage of COPY INTO command