Date

Snowflake as query enigine

# Goals

In this assignment, we will learn how to use snowflake as a query engine. We store our data in aws s3 and we will learn various methods to query it from snowflake.

## Query data in s3 from snowflake.

## Create view over data in aws s3.

## Disadvantages and advantages of this approach.

# Preparation

Before we start, let’s upload some sample data from snowflake to s3. Then we will try to query data in s3 from snowflake.

Create table,

CREATE OR REPLACE TRANSIENT TABLE DEMO\_DB.PUBLIC.CUSTOMER\_TEST

AS

SELECT \* FROM

"SNOWFLAKE\_SAMPLE\_DATA"."TPCDS\_SF100TCL"."CUSTOMER"

Execute below copy command to copy data to s3,

COPY INTO @DEMO\_DB.PUBLIC.MY\_S3\_STAGE/Customer\_data/

from

DEMO\_DB.PUBLIC.CUSTOMER\_TEST

![Graphical user interface, application, Teams

Description automatically generated]()

# Query data in s3 from snowflake.

Now data got uploaded to s3. We have 100 Million records uploaded and data size is 4.5 GB. Uploaded files will be csv compressed files.

Let’s try to query this data in s3 from snowflake.

SELECT $1 C\_CUSTOMER\_SK,

$2 C\_CUSTOMER\_ID ,

$3 C\_CURRENT\_CDEMO\_SK ,

$4 C\_CURRENT\_HDEMO\_SK ,

$5 C\_CURRENT\_ADDR\_SK ,

$6 C\_FIRST\_SHIPTO\_DATE\_SK ,

$7 C\_FIRST\_SALES\_DATE\_SK ,

$8 C\_SALUTATION ,

$9 C\_FIRST\_NAME ,

$10 C\_LAST\_NAME,

$11 C\_PREFERRED\_CUST\_FLAG ,

$12 C\_BIRTH\_DAY ,

$13 C\_BIRTH\_MONTH ,

$14 C\_BIRTH\_YEAR,

$16 C\_LOGIN ,

$17 C\_EMAIL\_ADDRESS ,

$18 C\_LAST\_REVIEW\_DATE

FROM @DEMO\_DB.PUBLIC.MY\_S3\_STAGE/Customer\_data/. ---replace it with new stage

(file\_format => DEMO\_DB.PUBLIC.MY\_CSV\_FORMAT)

**Filter data directly from s3,**

SELECT $1 C\_CUSTOMER\_SK,

$2 C\_CUSTOMER\_ID ,

$3 C\_CURRENT\_CDEMO\_SK ,

$4 C\_CURRENT\_HDEMO\_SK ,

$5 C\_CURRENT\_ADDR\_SK ,

$6 C\_FIRST\_SHIPTO\_DATE\_SK ,

$7 C\_FIRST\_SALES\_DATE\_SK ,

$8 C\_SALUTATION ,

$9 C\_FIRST\_NAME ,

$10 C\_LAST\_NAME,

$11 C\_PREFERRED\_CUST\_FLAG ,

$12 C\_BIRTH\_DAY ,

$13 C\_BIRTH\_MONTH ,

$14 C\_BIRTH\_YEAR,

$16 C\_LOGIN ,

$17 C\_EMAIL\_ADDRESS ,

$18 C\_LAST\_REVIEW\_DATE

FROM @DEMO\_DB.PUBLIC.MY\_S3\_STAGE/Customer\_data/

(file\_format => DEMO\_DB.PUBLIC.MY\_CSV\_FORMAT)

WHERE C\_CUSTOMER\_SK ='64596949'

**Execute group by,**

SELECT $9 C\_FIRST\_NAME,$10 C\_LAST\_NAME,COUNT(\*)

FROM @DEMO\_DB.PUBLIC.MY\_S3\_STAGE/Customer\_data/

(file\_format => DEMO\_DB.PUBLIC.MY\_CSV\_FORMAT)

GROUP BY $9,$10

# Create view over s3 data

CREATE OR REPLACE VIEW CUSTOMER\_DATA

AS

SELECT $1 C\_CUSTOMER\_SK,

$2 C\_CUSTOMER\_ID ,

$3 C\_CURRENT\_CDEMO\_SK ,

$4 C\_CURRENT\_HDEMO\_SK ,

$5 C\_CURRENT\_ADDR\_SK ,

$6 C\_FIRST\_SHIPTO\_DATE\_SK ,

$7 C\_FIRST\_SALES\_DATE\_SK ,

$8 C\_SALUTATION ,

$9 C\_FIRST\_NAME ,

$10 C\_LAST\_NAME,

$11 C\_PREFERRED\_CUST\_FLAG ,

$12 C\_BIRTH\_DAY ,

$13 C\_BIRTH\_MONTH ,

$14 C\_BIRTH\_YEAR,

$16 C\_LOGIN ,

$17 C\_EMAIL\_ADDRESS ,

$18 C\_LAST\_REVIEW\_DATE

FROM @DEMO\_DB.PUBLIC.MY\_S3\_STAGE/Customer\_data/

(file\_format => DEMO\_DB.PUBLIC.MY\_CSV\_FORMAT)

**Query data directly on view,**

SELECT \* FROM CUSTOMER\_DATA;



Now we can directly query data from s3 through view. What is the disadvantage of using this approach ? Can you see partitions being scanned in the backend ?

Now let’s try to Join the view we created with a table on snowflake,

Create a sample snowflake table as below,

Create or replace transient table CUSTOMER\_SNOWFLAKE\_TABLE

AS

SELECT \* FROM CUSTOMER\_TEST limit 10000

Join this with the view we created earlier,

SELECT B.\*

FROM CUSTOMER\_SNOWFLAKE\_TABLE B

LEFT OUTER JOIN

CUSTOMER\_DATA A

ON

A.C\_CUSTOMER\_SK = B.C\_CUSTOMER\_SK

Now we successfully joined data in s3 with snowflake table. It may look simple but this approach has lot of potential. Can you mention few below,

In the above example you joined a snowflake table with data on s3. Please check the profile page and observe the execution plan.

How many partitions got scanned from snowflake table :

# Unload data back to s3

This approach leverages micro partitions in snowflake for lookup table still giving us the freedom to query data which we have stored in s3.

Once we are done looking up we can copy data back to s3 with new derived lookup column.

COPY INTO @DEMO\_DB.PUBLIC.MY\_S3\_STAGE/Customer\_joined\_data/

from(

SELECT B.\*

FROM CUSTOMER\_SNOWFLAKE\_TABLE B

LEFT OUTER JOIN

CUSTOMER\_DATA A

ON

A.C\_CUSTOMER\_SK = B.C\_CUSTOMER\_SK

)

# Advantages and disadvantages

Write your views below,