**SAMPLE/TABLE SAMPLE**

Taking the sample of data from the actual data to test or todo some analysis. and after testing keep the data in same position. If we have milions of data its very expensive to run and test queries on large tables which are big in size such as terabytes. So in order to reduce the compute cost a fraction of data from the large tables is sample into a smaller view or tanle to get a similar data set as large tables. In Snowflake, the SAMPLE (or TABLESAMPLE) clause is used to fetch a random subset of rows from a table. We can specify either a percentage of rows or a fixed number of rows. For example, SAMPLE (10) returns 10% of rows, while SAMPLE ROW (1000) returns 1000 rows. We can also use a SEED value to make the sampling deterministic. I’ve used this in my previous projects to quickly validate data quality and test transformations without scanning the entire table.

There are 2 types of Sample methods, ROW OR Bernoulli method and System or Block method

**1. ROW / BERNOULLI Method**

* **Row-level sampling**.
* Each row is **independently included or excluded** based on probability.
* Can return **different number of rows each time**.

**2. SYSTEM / BLOCK Method**

* **Block-level sampling** (based on micro-partitions).
* Snowflake decides which blocks of data to read, so results are **faster** but may not be evenly random.
* Better for **large datasets**

Useful for:

* **Testing queries** on large tables
* **Quick data exploration**
* **Improving performance**

**Seed or Repeatable**(0 to 2147483647) used to produce the same data set if run again

while debugging

-- 50,000 -- 5000

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCDS\_SF100TCL.CATALOG\_PAGE sample row(10); -- Percentage --4,987

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCDS\_SF100TCL.CATALOG\_PAGE tablesample block(10); -- 5,082

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCDS\_SF100TCL.CATALOG\_PAGE sample (1000000 rows);

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCDS\_SF100TCL.CATALOG\_PAGE limit 10;

// Creating tables with sample data

// Bernoulli or Row

create table data\_sampling\_1

as

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCH\_SF10.CUSTOMER sample bernoulli(10); -- 1,500,000

create table data\_sampling\_2

as

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCH\_SF10.CUSTOMER tablesample row(10) seed(452);

select count(\*) from data\_sampling\_1; -- 150398

select count(\*) from data\_sampling\_2; -- 150084

--drop table data\_sampling\_2;

// System or block

create or replace table data\_sampling\_3

as

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCH\_SF10.CUSTOMER tablesample system(20);

create or replace table data\_sampling\_4

as

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCH\_SF10.CUSTOMER sample block(20) seed(11);

select count(\*) from data\_sampling\_3; -- 277500

select count(\*) from data\_sampling\_4; -- 309000

//testing sampling data

select \* from data\_sampling\_1 -- 150514

minus

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCH\_SF10.CUSTOMER sample bernoulli(10); -- 135,274

select \* from data\_sampling\_2

minus

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCH\_SF10.CUSTOMER sample row(10) seed(452);--same

select \* from data\_sampling\_3

minus

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCH\_SF10.CUSTOMER tablesample system(20); -- 225,000

select \* from data\_sampling\_4

minus

select \* from SNOWFLAKE\_SAMPLE\_DATA.TPCH\_SF10.CUSTOMER sample block(20 rows) repeatable(11); -- same