Clustering Snowflake

Prerequisite: Create Worksheet with the name clustering

Step 1: Create Database and Schema:

```
CREATE OR REPLACE DATABASE clustering_lab;
USE DATABASE clustering_lab;
CREATE OR REPLACE SCHEMA lab_schema;
USE SCHEMA lab_schema;
```

Step 2: **Create Non Clustered Table**

```
CREATE OR REPLACE TABLE big_sales (
   id INT,
   region STRING,
   sale_date DATE,
   amount NUMBER
);
```

Step 3:

Insert sample data (Below query simulate 1M rows using a loop or Snowflake generator)

```
INSERT INTO big_sales
SELECT
   seq4() AS id,
   CASE MOD(seq4(), 5)
      WHEN 0 THEN 'North'
      WHEN 1 THEN 'South'
      WHEN 2 THEN 'East'
      WHEN 3 THEN 'West'
      ELSE 'Central'
   END AS region,
```

```
    DATEADD(DAY, UNIFORM(0, 365 * 3, RANDOM()), DATE '2022-01-01') AS sale_date,
    UNIFORM(100, 1000, RANDOM()) AS amount
FROM TABLE(GENERATOR(ROWCOUNT => 1000000));
```

Step 4:

Query 1: To retrieve Sales in 'South' region from 2023

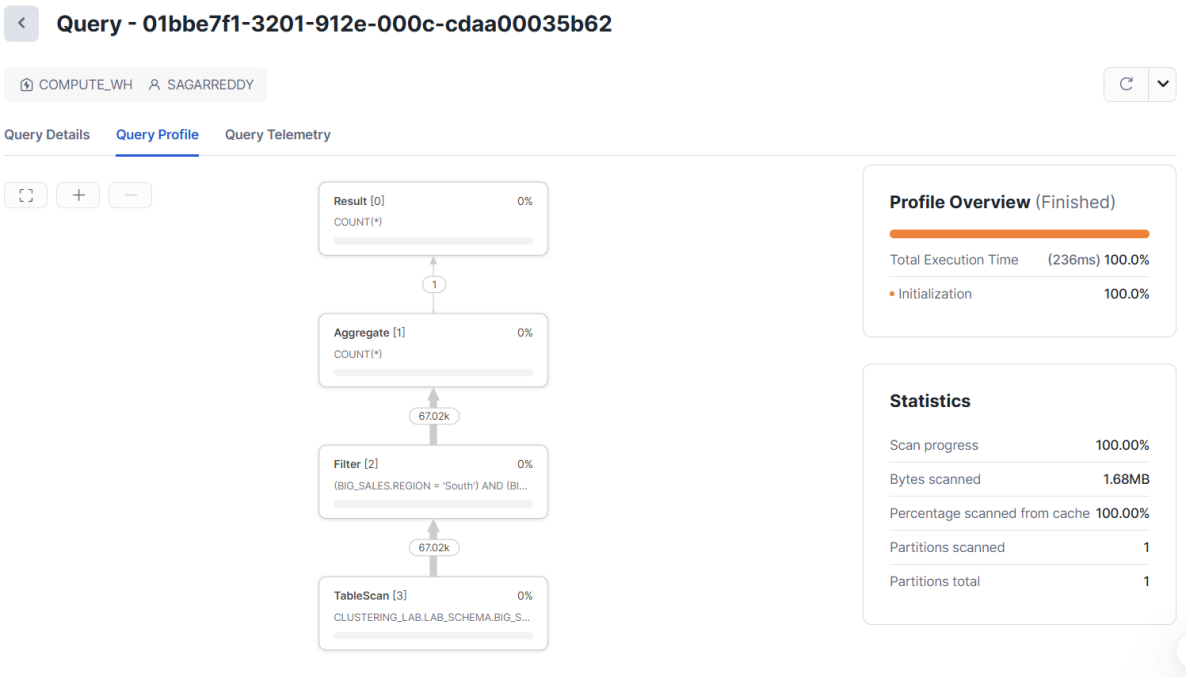
```
SELECT COUNT(*)
FROM big_sales
WHERE region = 'South'
  AND sale_date BETWEEN '2023-01-01' AND '2023-12-31';
```

Note: Observe Query Profile as we are running this query before clustering , and We will run this query after creating clustering table.

check the Query Profile in the Snowflake UI:

- Bytes scanned: likely high
- Partitions scanned: most or all
- Execution time: longer

Before Clustering



Step 5:

Creating Clustering Table on columns region and sale_date

```
CREATE OR REPLACE TABLE big_sales_clustered
CLUSTER BY (region, sale_date)
AS
SELECT * FROM big_sales;
```

Step 6:

```
SELECT COUNT(*)
FROM big_sales_clustered
WHERE region = 'South'
```

AND sale_date BETWEEN '2023-01-01' AND '2023-12-31';

Note: Observe Query Profile Now , To understand how clustering improved performance

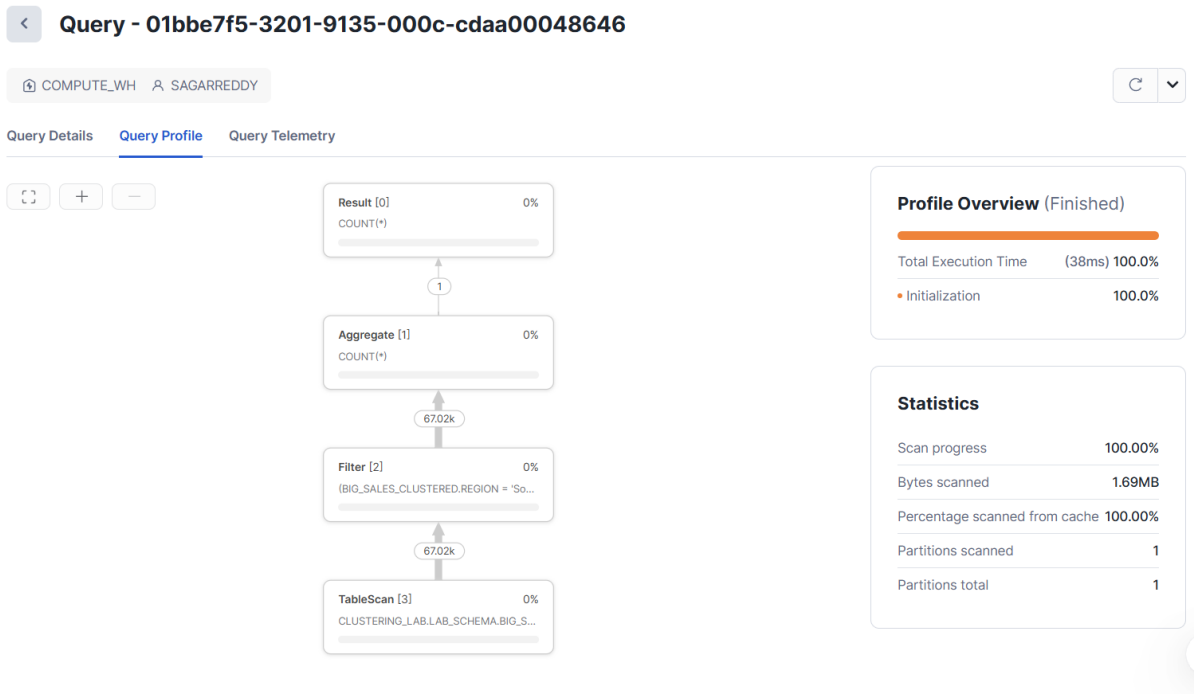
compare the Query Profile again:

- Bytes scanned: significantly lower

- Partitions scanned: fewer

- Execution time: improved

After Clustering



To View Clustering Depth:

SELECT SYSTEM$CLUSTERING_INFORMATION('big_sales_clustered');

Question:

Write an SQL query to retrieve the id, region, and amount of the first 5 sales records in the "big_sales_clustered" table from the "South" region that occurred in the year 2023. Ensure the results are ordered by sale_date in ascending order.