Caching - DF

February 4, 2025

[9]:
```
from pyspark.sql import SparkSession
```

[8]:
```
spark.stop()
```

[10]:
```
spark = SparkSession.builder \
.appName('DataFrame_caching_demo')\
.enableHiveSupport()\
.getOrCreate()
```

```
25/02/02 05:20:39 INFO SparkEnv: Registering MapOutputTracker
25/02/02 05:20:39 INFO SparkEnv: Registering BlockManagerMaster
25/02/02 05:20:39 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/02/02 05:20:39 INFO SparkEnv: Registering OutputCommitCoordinator
```

[13]:
```
customers_df= spark.read.option('header','true').csv('/tmp/customers_500mb.csv')
```

[14]:
```
customers_df.printSchema()
```

```
root
 |-- customer_id: string (nullable = true)
 |-- name: string (nullable = true)
 |-- city: string (nullable = true)
 |-- state: string (nullable = true)
 |-- country: string (nullable = true)
 |-- registration_date: string (nullable = true)
 |-- is_active: string (nullable = true)
```

[35]:
```
customers_df.count()
```

[35]:
```
8767358
```

[36]:
```
customers_df.cache() # ->customers_df.cache().show()
```

[36]:
```
DataFrame[customer_id: string, name: string, city: string, state: string,
country: string, registration_date: string, is_active: string]
```

[37]:
```
customers_df.count()
```

[37]:
```
8767358
```

I still can't understand why did it only store first partition in customers_df, its possible we could have needed the 2nd partition or the last partition? Or this us the default behavior of saving the 1st partition? then how do we store a certain partition?

[33]:
```
customers_df.show(5)
```

```
+-----------+----------+---------+-----------+-------+-----------------+--------
-+
|customer_id|      name|     city|
state|country|registration_date|is_active|
+-----------+----------+---------+-----------+-------+-----------------+--------
-+
|          0|Customer_0|   Mumbai|  Telangana|  India|       2023-03-21|
True|
|          1|Customer_1|  Chennai|West Bengal|  India|       2023-05-27|
False|
|          2|Customer_2|     Pune|  Karnataka|  India|       2023-10-11|
False|
|          3|Customer_3|Hyderabad|    Gujarat|  India|       2023-11-11|
False|
|          4|Customer_4|   Mumbai|  Karnataka|  India|       2023-05-09|
False|
+-----------+----------+---------+-----------+-------+-----------------+--------
-+
only showing top 5 rows
```

[30]:
```
customers_df.show(5)
```

```
+-----------+----------+---------+-----------+-------+-----------------+--------
-+
|customer_id|      name|     city|
state|country|registration_date|is_active|
+-----------+----------+---------+-----------+-------+-----------------+--------
-+
|          0|Customer_0|   Mumbai|  Telangana|  India|       2023-03-21|
True|
|          1|Customer_1|  Chennai|West Bengal|  India|       2023-05-27|
False|
|          2|Customer_2|     Pune|  Karnataka|  India|       2023-10-11|
False|
|          3|Customer_3|Hyderabad|    Gujarat|  India|       2023-11-11|
False|
```

```
|          4|Customer_4|   Mumbai|  Karnataka|  India|       2023-05-09|
False|
+-----------+----------+---------+-----------+-------+-----------------+--------
-+
only showing top 5 rows
```

[26]:
```
tail_df.unpersist()
```

[26]:
```
DataFrame[customer_id: string, name: string, city: string, state: string,
country: string, registration_date: string, is_active: string]
```

[34]:
```
customers_df.unpersist()
```

[34]:
```
DataFrame[customer_id: string, name: string, city: string, state: string,
country: string, registration_date: string, is_active: string]
```

[21]:
```
tail_df = customers_df.orderBy('customer_id',ascending=False)
```

[22]:
```
tail_df.show(5)
```

```
[Stage 9:================================================>          (4 + 1) / 5]

+-----------+--------------+---------+-----------+-------+-----------------+---
------+
|customer_id|          name|     city|
state|country|registration_date|is_active|
+-----------+--------------+---------+-----------+-------+-----------------+---
------+
|     999999|Customer_999999|Hyderabad|  Karnataka|  India|       2023-05-30|
True|
|     999998|Customer_999998|    Delhi|West Bengal|  India|       2023-07-21|
True|
|     999997|Customer_999997|  Kolkata|  Karnataka|  India|       2023-04-03|
True|
|     999996|Customer_999996|Hyderabad|Maharashtra|  India|       2023-09-03|
True|
|     999995|Customer_999995|     Pune|    Gujarat|  India|       2023-05-03|
False|
+-----------+--------------+---------+-----------+-------+-----------------+---
------+
only showing top 5 rows
```

[23]:
```
tail_df.cache()
```

[23]:
```
DataFrame[customer_id: string, name: string, city: string, state: string,
country: string, registration_date: string, is_active: string]
```

[24]:
```
tail_df.show(5)
```

```
[Stage 11:=============================================>           (4 + 1) / 5]
+-----------+--------------+---------+-----------+-------+-----------------+---
------+
|customer_id|          name|     city|
state|country|registration_date|is_active|
+-----------+--------------+---------+-----------+-------+-----------------+---
------+
|     999999|Customer_999999|Hyderabad|  Karnataka|  India|       2023-05-30|
True|
|     999998|Customer_999998|    Delhi|West Bengal|  India|       2023-07-21|
True|
|     999997|Customer_999997|  Kolkata|  Karnataka|  India|       2023-04-03|
True|
|     999996|Customer_999996|Hyderabad|Maharashtra|  India|       2023-09-03|
True|
|     999995|Customer_999995|     Pune|    Gujarat|  India|       2023-05-03|
False|
+-----------+--------------+---------+-----------+-------+-----------------+---
------+
only showing top 5 rows
```

[25]:
```
tail_df.show(5)
```

```
+-----------+--------------+---------+-----------+-------+-----------------+---
------+
|customer_id|          name|     city|
state|country|registration_date|is_active|
+-----------+--------------+---------+-----------+-------+-----------------+---
------+
|     999999|Customer_999999|Hyderabad|  Karnataka|  India|       2023-05-30|
True|
|     999998|Customer_999998|    Delhi|West Bengal|  India|       2023-07-21|
True|
|     999997|Customer_999997|  Kolkata|  Karnataka|  India|       2023-04-03|
True|
|     999996|Customer_999996|Hyderabad|Maharashtra|  India|       2023-09-03|
True|
|     999995|Customer_999995|     Pune|    Gujarat|  India|       2023-05-03|
False|
+-----------+--------------+---------+-----------+-------+-----------------+---
```

```
------+
only showing top 5 rows
```

[38]:
```
spark.stop()
```