assignment_1

June 4, 2025

[2]:
```
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder\
                    . master("local")\
                    .appName("source_step")\
                    .getOrCreate()
path="gs://buck_et01/eq_sample.csv"
df=spark.read.csv(path, header=True)
df.show()
df.printSchema()
```

```
[Stage 1:>                                                          (0 + 1) / 1]

+------------+--------------+-------------------+-------------------+-----------
--------+--------------+-------------------+-------------------+---------+
|customer_nbr| customer_desc|           start_ts|             end_ts|
create_ts|create_user_id|     last_update_ts|last_update_user_id|client_id|
+------------+--------------+-------------------+-------------------+-----------
--------+--------------+-------------------+-------------------+---------+
|     CUST001|     Acme Corp|2024-01-01 08:00:00|2025-01-01 08:00:00|2024-01-01
07:50:00|         admin|2024-06-01 10:20:00|              user1|    CL001|
|     CUST002|   Global Tech|2023-05-15 09:15:00|2025-05-15 09:15:00|2023-05-01
12:00:00|         admin|2024-06-01 11:30:00|              user2|    CL002|
|     CUST003|    Blue Ocean|2024-03-20 10:00:00|2026-03-20 10:00:00|2024-03-18
09:45:00|         user3|2024-06-01 12:15:00|              user3|    CL003|
|     CUST004|   Sunrise Ltd|2024-06-01 00:00:00|2025-06-01 00:00:00|2024-05-28
14:30:00|         admin|2024-06-02 08:30:00|              user4|    CL004|
|     CUST005|   NextGen Inc|2023-11-10 14:00:00|2025-11-10 14:00:00|2023-11-01
10:00:00|         user5|2024-06-01 10:00:00|              user5|    CL005|
|     CUST006|Falcon Systems|2022-01-01 08:00:00|2023-01-01 08:00:00|2021-12-20
16:00:00|         admin|2024-06-01 09:10:00|              user6|    CL006|
|     CUST007|  Green Fields|2024-02-14 06:00:00|2026-02-14 06:00:00|2024-02-01
07:00:00|         user7|2024-06-01 08:00:00|              user7|    CL007|
|     CUST008|  Cloud Matrix|2023-07-01 10:00:00|2024-07-01 10:00:00|2023-06-20
15:30:00|         admin|2024-06-01 12:00:00|              user8|    CL008|
|     CUST009|     Iron Gate|2024-04-01 12:00:00|2025-04-01 12:00:00|2024-03-25
11:00:00|         user9|2024-06-01 14:00:00|              user9|    CL009|
|     CUST010|     Aqua Tech|2024-06-01 09:00:00|2025-06-01 09:00:00|2024-05-30
```

```
08:00:00|        user10|2024-06-01 15:00:00|             user10|    CL010|
+------------+--------------+-------------------+-------------------+----------
--------+--------------+-------------------+-------------------+---------+

root
 |-- customer_nbr: string (nullable = true)
 |-- customer_desc: string (nullable = true)
 |-- start_ts: string (nullable = true)
 |-- end_ts: string (nullable = true)
 |-- create_ts: string (nullable = true)
 |-- create_user_id: string (nullable = true)
 |-- last_update_ts: string (nullable = true)
 |-- last_update_user_id: string (nullable = true)
 |-- client_id: string (nullable = true)
```

[6]:
```
from pyspark.sql.functions import lit
from pyspark.sql.types import StructType, StructField, StringType,␣
 ↪TimestampType, DateType
from pyspark.sql.functions import col, to_timestamp, to_date

df1 = df.withColumn("start_timestamp", to_timestamp(col("start_ts"),␣
 ↪"yyyy-MM-dd HH:mm:ss")) \
        .withColumn("end_timestamp", to_timestamp(col("end_ts"), "yyyy-MM-dd HH:
 ↪mm:ss")) \
        .withColumn("create_timestamp", to_timestamp(col("create_ts"),␣
 ↪"yyyy-MM-dd HH:mm:ss")) \
        .withColumn("last_update_timestamp",␣
 ↪to_timestamp(col("last_update_ts"),"yyyy-MM-dd HH:mm:ss"))\
        .withColumn("start_ts", to_date(df["start_ts"], "yyyy-MM-dd HH:mm:ss").
 ↪cast(DateType()))\
        .withColumn("end_ts", to_date(df["end_ts"], "yyyy-MM-dd HH:mm:ss").
 ↪cast(DateType()))\
        .withColumn("create_ts", to_date(df["create_ts"], "yyyy-MM-dd HH:mm:
 ↪ss").cast(DateType()))\
        .withColumn("last_update_ts", to_date(df["last_update_ts"], "yyyy-MM-dd␣
 ↪HH:mm:ss").cast(DateType()))
df1.show()
df1.printSchema()
```

```
+------------+--------------+---------+---------+---------+--------------+---
-----------+------------------+---------+------------------+-----------------
-+------------------+--------------------+
|customer_nbr| customer_desc|  start_ts|    end_ts|
create_ts|create_user_id|last_update_ts|last_update_user_id|client_id|
start_timestamp|      end_timestamp|   create_timestamp|last_update_timestamp|
```

```
+------------+--------------+----------+----------+----------+--------------+---
-----------+------------------+---------+------------------+-----------------
-+------------------+--------------------+
|     CUST001|     Acme Corp|2024-01-01|2025-01-01|2024-01-01|         admin|
2024-06-01|              user1|    CL001|2024-01-01 08:00:00|2025-01-01
08:00:00|2024-01-01 07:50:00|  2024-06-01 10:20:00|
|     CUST002|   Global Tech|2023-05-15|2025-05-15|2023-05-01|         admin|
2024-06-01|              user2|    CL002|2023-05-15 09:15:00|2025-05-15
09:15:00|2023-05-01 12:00:00|  2024-06-01 11:30:00|
|     CUST003|    Blue Ocean|2024-03-20|2026-03-20|2024-03-18|         user3|
2024-06-01|              user3|    CL003|2024-03-20 10:00:00|2026-03-20
10:00:00|2024-03-18 09:45:00|  2024-06-01 12:15:00|
|     CUST004|   Sunrise Ltd|2024-06-01|2025-06-01|2024-05-28|         admin|
2024-06-02|              user4|    CL004|2024-06-01 00:00:00|2025-06-01
00:00:00|2024-05-28 14:30:00|  2024-06-02 08:30:00|
|     CUST005|   NextGen Inc|2023-11-10|2025-11-10|2023-11-01|         user5|
2024-06-01|              user5|    CL005|2023-11-10 14:00:00|2025-11-10
14:00:00|2023-11-01 10:00:00|  2024-06-01 10:00:00|
|     CUST006|Falcon Systems|2022-01-01|2023-01-01|2021-12-20|         admin|
2024-06-01|              user6|    CL006|2022-01-01 08:00:00|2023-01-01
08:00:00|2021-12-20 16:00:00|  2024-06-01 09:10:00|
|     CUST007|  Green Fields|2024-02-14|2026-02-14|2024-02-01|         user7|
2024-06-01|              user7|    CL007|2024-02-14 06:00:00|2026-02-14
06:00:00|2024-02-01 07:00:00|  2024-06-01 08:00:00|
|     CUST008|  Cloud Matrix|2023-07-01|2024-07-01|2023-06-20|         admin|
2024-06-01|              user8|    CL008|2023-07-01 10:00:00|2024-07-01
10:00:00|2023-06-20 15:30:00|  2024-06-01 12:00:00|
|     CUST009|     Iron Gate|2024-04-01|2025-04-01|2024-03-25|         user9|
2024-06-01|              user9|    CL009|2024-04-01 12:00:00|2025-04-01
12:00:00|2024-03-25 11:00:00|  2024-06-01 14:00:00|
|     CUST010|     Aqua Tech|2024-06-01|2025-06-01|2024-05-30|        user10|
2024-06-01|             user10|    CL010|2024-06-01 09:00:00|2025-06-01
09:00:00|2024-05-30 08:00:00|  2024-06-01 15:00:00|
+------------+--------------+----------+----------+----------+--------------+---
-----------+------------------+---------+------------------+-----------------
-+------------------+--------------------+

root
 |-- customer_nbr: string (nullable = true)
 |-- customer_desc: string (nullable = true)
 |-- start_ts: date (nullable = true)
 |-- end_ts: date (nullable = true)
 |-- create_ts: date (nullable = true)
 |-- create_user_id: string (nullable = true)
 |-- last_update_ts: date (nullable = true)
 |-- last_update_user_id: string (nullable = true)
 |-- client_id: string (nullable = true)
 |-- start_timestamp: timestamp (nullable = true)
```

```
 |-- end_timestamp: timestamp (nullable = true)
 |-- create_timestamp: timestamp (nullable = true)
 |-- last_update_timestamp: timestamp (nullable = true)
```

[ ]: