

Untitled2

June 5, 2025

```
[1]: import pyspark
      from pyspark.sql import SparkSession
      from pyspark.sql.types import StructType, StructField, StringType, TimestampType
      from pyspark.sql.functions import col, to_timestamp
```

```
[2]: spark = SparkSession.builder\
      .master("local")\
      .appName("test")\
      .getOrCreate()
```

```
[5]: schema = StructType([
      StructField("customer_nbr", StringType(), True),
      StructField("customer_desc", StringType(), True),
      StructField("start_ts", StringType(), True),
      StructField("end_ts", StringType(), True),
      StructField("create_ts", StringType(), True),
      StructField("last_update_ts", StringType(), True),
      StructField("client_id", StringType(), True)
    ])
```

```
[9]: df_raw = spark.read.csv(
      "gs://datapyspark/customer_data.csv",
      header=True,
      schema=schema
    )

df_raw.show()
```

```
+-----+-----+-----+-----+-----+-----+-----+
|customer_nbr|customer_desc|start_ts|end_ts|create_ts|last_update_ts|client_id|
+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+
only showing top 0 rows
```

```
[11]: df_raw.show()
```

+-----+-----+-----+-----+-----+					
-----+-----+-----+-----+-----+					
customer_nbr customer_desc		start_ts		end_ts	
create_ts last_update_ts		client_id			
+-----+-----+-----+-----+-----+					
	CUST003	Customer C	2022-12-08 22:59:19	2023-11-28 13:23:03	2021-09-22
12:36:06		user3	2022-03-09 00:45:34		
	CUST004	Customer B	2022-09-23 17:31:40	2023-09-25 15:57:53	2021-03-08
11:52:34		user4	2022-08-31 17:57:57		
	CUST002	Customer D	2022-07-08 22:29:18	2023-02-24 03:16:23	2021-07-16
12:57:40		user2	2022-06-24 22:13:39		
	CUST004	Customer D	2022-12-25 20:41:29	2023-05-24 05:42:58	2021-05-30
19:28:21		user2	2022-06-10 18:07:29		
	CUST001	Customer D	2022-12-29 00:06:48	2023-06-03 16:54:36	2021-11-19
20:38:19		user4	2022-04-09 05:20:27		
	CUST004	Customer A	2022-01-20 01:03:28	2023-04-09 04:26:01	2021-12-01
22:47:42		user4	2022-07-02 22:05:32		
	CUST002	Customer C	2022-05-19 15:15:52	2023-09-04 21:57:20	2021-06-14
01:48:49		user4	2022-10-06 04:34:19		
	CUST004	Customer B	2022-09-18 05:45:43	2023-11-11 05:50:45	2021-11-20
11:08:28		user2	2022-12-12 11:56:58		
	CUST001	Customer B	2022-09-22 16:46:56	2023-12-24 09:50:40	2021-12-12
09:21:34		user4	2022-05-25 00:34:22		
	CUST003	Customer D	2022-11-29 23:45:45	2023-03-17 19:42:50	2021-10-20
17:05:13		user4	2022-10-30 07:01:22		
	CUST004	Customer D	2022-04-21 09:46:54	2023-02-15 05:14:24	2021-01-18
12:26:47		user1	2022-04-30 16:57:42		
	CUST004	Customer B	2022-02-26 17:59:11	2023-04-24 23:30:59	2021-02-04
18:24:10		user1	2022-04-01 03:32:01		
	CUST002	Customer C	2022-07-12 23:21:09	2023-12-27 01:53:30	2021-09-20
21:54:35		user3	2022-06-27 06:35:12		
	CUST001	Customer D	2022-10-02 13:36:42	2023-02-19 02:08:02	2021-12-06
00:09:33		user2	2022-04-15 04:45:35		
	CUST003	Customer D	2022-05-05 23:32:11	2023-08-25 11:08:05	2021-07-09
05:51:34		user2	2022-01-06 03:56:05		
	CUST003	Customer C	2022-04-25 09:21:10	2023-11-12 03:24:28	2021-08-01
06:01:45		user2	2022-05-30 17:08:02		
	CUST003	Customer C	2022-09-10 21:59:57	2023-02-07 05:01:21	2021-04-06
15:46:18		user3	2022-12-17 06:45:53		
	CUST001	Customer C	2022-01-05 05:16:43	2023-12-17 04:01:29	2021-12-26
07:47:14		user1	2022-12-06 17:04:53		
	CUST004	Customer C	2022-04-26 10:00:11	2023-05-04 21:14:12	2021-11-19
12:39:04		user1	2022-10-04 02:08:58		
	CUST001	Customer C	2022-04-29 21:19:24	2023-07-09 14:14:15	2021-09-23
17:39:17		user1	2022-07-06 04:23:33		
+-----+-----+-----+-----+-----+					
-----+-----+-----+-----+-----+					

only showing top 20 rows

```
[10]: #-----  
[ ]:  
[ ]:
```