# 04-building-a-data-lake

## ETL with Spark (Local)

```
from pyspark.sql import SparkSession
#from pyspark.sql.types import StructType, StructField, DoubleType, StringType, IntegerType, DateType, TimestampType

#import pyspark.sql.functions as F
```

```
data = "github_events_01.json"
```

```
spark = SparkSession.builder \
    .appName("ETL") \
    .getOrCreate()
```

```
data = spark.read.option("multiline", "true").json(data)
```

```
data.printSchema()
```

```
root
 |-- actor: struct (nullable = true)
 |    |-- avatar_url: string (nullable = true)
 |    |-- display_login: string (nullable = true)
```

---

```
data.createOrReplaceTempView("staging_events")
```

```
table = spark.sql("""
    select
        *
    from
        staging_events
""").show()
```

```
+--------------------+--------------------+----------+--------------------+--------------------+------+----------
---------+----------------+
|               actor|          created_at|        id|                 org|             payload|public|
repo|            type|
+--------------------+--------------------+----------+--------------------+--------------------+------+----------
---------+----------------+
|{https://avatars....|2022-08-17T15:51:05Z|23487929637|{https://avatars....|{created, {COLLAB...|  true|{75340147,
350org...|IssueCommentEvent|
+--------------------+--------------------+----------+--------------------+--------------------+------+----------
---------+----------------+
```

```
table = spark.sql("""
    select
        id
        , type
        , created_at
        , to_date(created_at) as date
        , day(created_at) as day
        , month(created_at) as month
        , year(created_at) as year
        , actor.id as login_id
        , actor.login as login_name
        , repo.name as repo_name
        , payload.action as action
        , payload.issue.user.login as username
        , org.login as org_name

    from
        staging_events
""")
```

---

```
table.show()
```

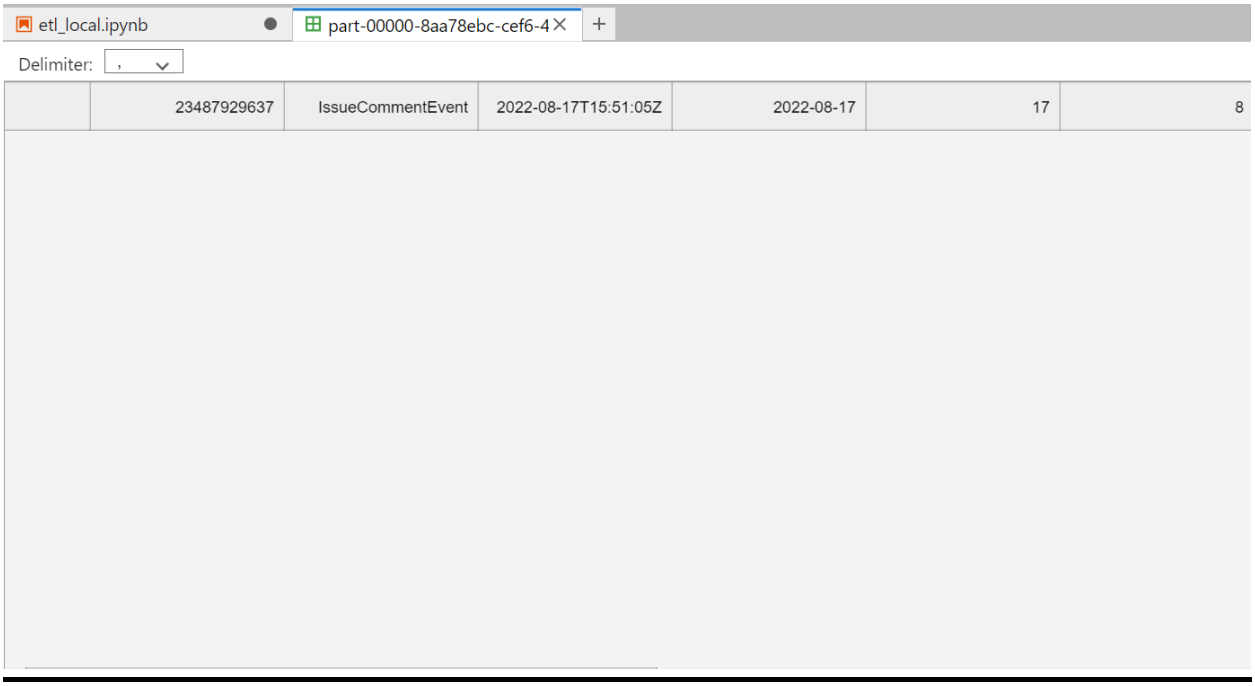
```
+-----------+-----------------+--------------------+----------+---+-----+----+--------+----------+-----------------
+-------+---------+--------+
|         id|             type|          created_at|      date|day|month|year|login_id|login_name|        repo_name
| action| username|org_name|
+-----------+-----------------+--------------------+----------+---+-----+----+--------+----------+-----------------
+-------+---------+--------+
|23487929637|IssueCommentEvent|2022-08-17T15:51:05Z|2022-08-17| 17|    8|2022| 1696078|   sukhada|350org/ak_intl_v3
|created|rachelhbd|  350org|
+-----------+-----------------+--------------------+----------+---+-----+----+--------+----------+-----------------
+-------+---------+--------+
```

```
output_csv = "../output_csv"
output_parquet = "../output_parquet"
```

```
table.write.partitionBy("year").mode("overwrite").csv(output_csv)
```

```
table.write.partitionBy("year").mode("overwrite").parquet(output_parquet)
```

---


etl_local.ipynb
part-00000-8aa78ebc-cef6-4
Delimiter: ,
23487929637
IssueCommentEvent
2022-08-17T15:51:05Z
2022-08-17
17
8