

Nombre: Anahí Andrade

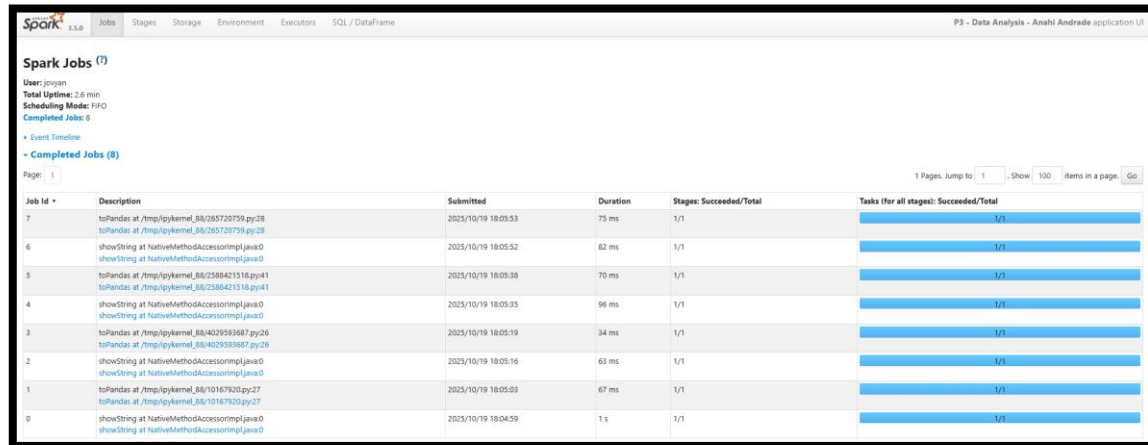
Fecha:19-10-2025

Código: 00323313

## Data Mining – Deber 3

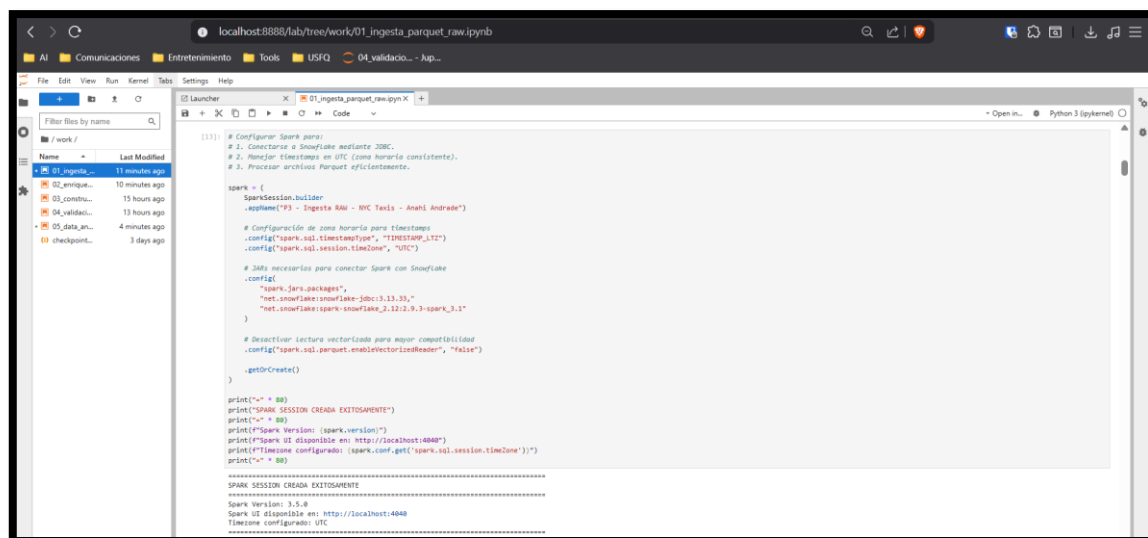
### 1. Evidencia - Capturas de Spark ejecutándose en el puerto designado, al igual que Servidor Jupyter:

**Puerto (<http://localhost:4040>):**



Job ID	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
7	toPandas at /tmp/pykernel_88/265720759.py:28 showString at /tmp/pykernel_88/265720759.py:28	2025/10/19 18:05:53	79 ms	1/1	1/1
6	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2025/10/19 18:05:52	82 ms	1/1	1/1
5	toPandas at /tmp/pykernel_88/2588421518.py:41 toPandas at /tmp/pykernel_88/2588421518.py:41	2025/10/19 18:05:38	70 ms	1/1	1/1
4	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2025/10/19 18:05:35	96 ms	1/1	1/1
3	toPandas at /tmp/pykernel_88/4029591607.py:26 toPandas at /tmp/pykernel_88/4029591607.py:26	2025/10/19 18:05:19	34 ms	1/1	1/1
2	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2025/10/19 18:05:16	63 ms	1/1	1/1
1	toPandas at /tmp/pykernel_88/10167920.py:27 toPandas at /tmp/pykernel_88/10167920.py:27	2025/10/19 18:05:03	67 ms	1/1	1/1
0	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2025/10/19 18:04:59	1 s	1/1	1/1

**Servidor Jupyter:**



```
[1]: # Configurar Spark para:
# 1. Conectar a Snowflake mediante JDBC.
# 2. Manejar timestamps en UTC (zona horaria consistente).
# 3. Procesar archivos Parquet eficientemente.

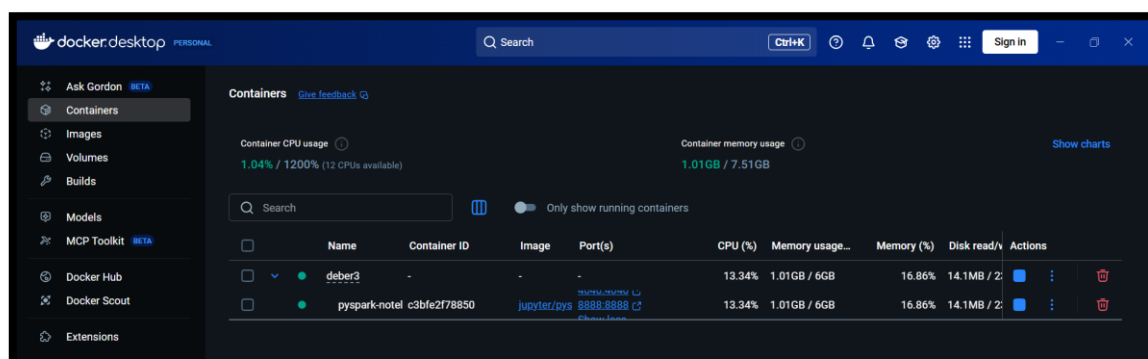
spark = (
    SparkSession.builder
        .appName("P3 - Ingesta RAW - NYC Taxis - Anahí Andrade")
        .config("spark.jars.packages",
            "spark.snowflake:snowflake-jdbc:3.13.33,"
            "net.snowflake:spark-snowflake_2.12:2.9.3-spark_3.3")
        .config("spark.sql.timestampType", "TIMESTAMP_LTZ")
        .config("spark.sql.session.timeZone", "UTC")
        .getOrCreate()

    # Desactivar lectura vectorizada para mayor compatibilidad
    .config("spark.sql.parquet.enableVectorizedReader", "false")
)

print(f"Spark Version: {spark.version}")
print(f"Spark UI disponible en: http://localhost:4040")
print(f"Timestamp configurado: {spark.conf.get('spark.sql.session.timeZone')}")
print(f"Spark UI disponible en: http://localhost:4040")
print(f"Timestamp configurado: UTC")

=====
SPARK SESSION CREADA EXITOSAMENTE
=====
Spark Version: 3.5.0
Spark UI disponible en: http://localhost:4040
Timestamp configurado: UTC
```

### 2. Evidencia – Docker Compose ejecutándose:



Name	Container ID	Image	Port(s)	CPU (%)	Memory usage...	Memory (%)	Disk read/W	Actions
deber3		deber3		13.34%	1.01GB / 6GB	16.86%	14.1MB / 2	
pyspark-notebook	c3bfe2f78850	jupyter/pys	8888:8888	13.34%	1.01GB / 6GB	16.86%	14.1MB / 2	

### 3. Evidencia - Variables de ambiente bien usadas:

```
[11]: # Verificar que todas las variables de ambiente necesarias estén configuradas.

import os
import requests
from io import BytesIO
import json
from datetime import datetime
from pyspark.sql import SparkSession
from pyspark.sql import functions as F
from pyspark.sql import types as T
import snowflake.connector

print("\n" + 88)
print("CONFIGURACIÓN DE AMBIENTE - PROYECTO 3")
print("\n" + 88)

# Variables obligatorias de Snowflake
required_vars = [
    'SNOWFLAKE_ACCOUNT',
    'SNOWFLAKE_DATABASE',
    'SNOWFLAKE_SCHEMA',
    'SNOWFLAKE_WAREHOUSE',
    'SNOWFLAKE_USER',
    'SNOWFLAKE_PASSWORD',
    'SNOWFLAKE_ROLE'
]

# Verificar que todas las variables existan
missing_vars = [var for var in required_vars if not os.getenv(var)]
if missing_vars:
    print("ERROR: Faltan variables de ambiente ('', ' '.join(missing_vars))")
    print("Por favor configura tu archivo .env correctamente.")
else:
    print("Todas las variables de ambiente requeridas están configuradas.")

print("\n" + 88)

=====
CONFIGURACIÓN DE AMBIENTE - PROYECTO 3
=====
Todas las variables de ambiente requeridas están configuradas.
=====
```

### 4. Evidencia - Cobertura real 2015–2025 (Yellow/Green):

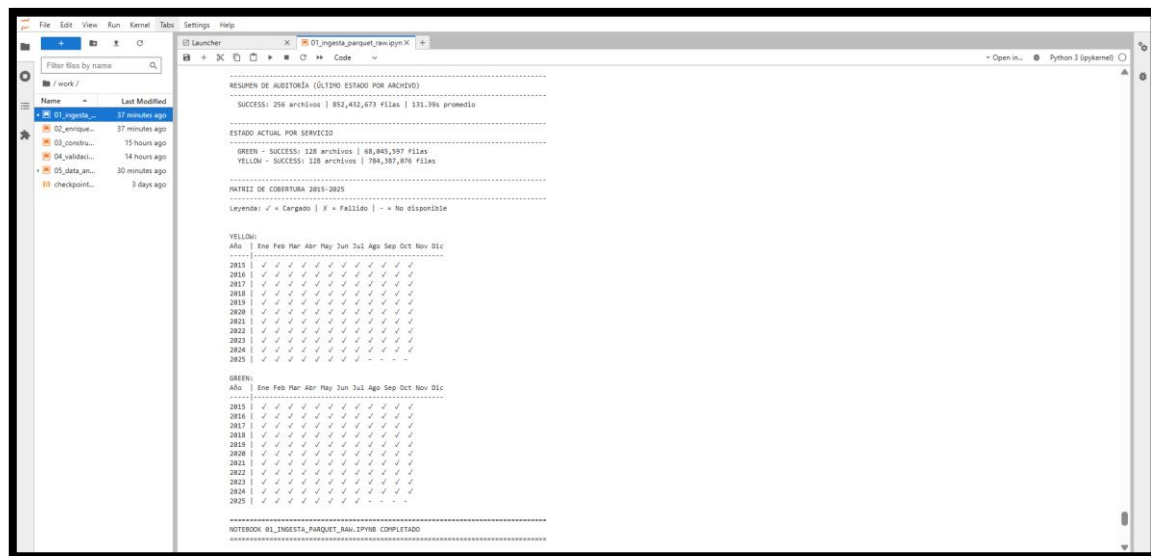
#### Tabla INGESTA\_AUDIT:

Se creó en Snowflake una tabla de auditoría llamada **INGESTA\_AUDIT** dentro del esquema **RAW**. Esta tabla registra información como el tipo de servicio del archivo Parquet (YELLOW/GREEN), el año y mes correspondientes, la URL del archivo, el número de lote, la cantidad de filas contenidas, el tiempo de ejecución, el estado del proceso (exitoso o fallido), entre otros datos. Su propósito fue facilitar la identificación de los años y meses que se procesaron correctamente.

AUDIT_ID	RUN_ID	SERVICE	TYPE	SOURCE_YEAR	SOURCE_MONTH	SOURCE_PATH	BATCH_NUMBER	ROWS
1	6	1	yellow	2015	1	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
2	704	1	yellow	2015	2	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
3	603	1	yellow	2015	3	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
4	7	1	yellow	2015	4	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
5	406	1	yellow	2015	5	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
6	105	1	yellow	2015	6	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
7	407	1	yellow	2015	7	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
8	705	1	yellow	2015	8	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
9	604	1	yellow	2015	9	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
10	8	1	yellow	2015	10	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1
11	9	1	yellow	2015	11	https://d37c6vzurychv.cloudfront.net/trip-data-yellow-trips	1	1

#### Jupyter:

De forma similar, en el notebook **01\_ingesta\_parquet\_raw.ipynb** se incluyeron impresiones de auditoría.



## 5. Evidencia – Datos y Metadatos de la cobertura real 2015–2025 (Yellow/Green):

### Tabla RAW YELLOW TRIPS:

Object Details for: NY\_TAXI.RAW.RAW\_YELLOW...

	name	type	kind
	AIRPORT_FEE	3.8% FLOAT	46.2%
	BATCH_NUMBER	3.8% NUMBER(38,0)	30.8%
	+24 more	+4 more	CO... 100.0%
1	VENDORID	NUMBER(38,0)	COLUMN
2	TPEP_PICKUP_DATETIME	TIMESTAMP_NTZ(9)	COLUMN
3	TPEP_DROPOFF_DATETIME	TIMESTAMP_NTZ(9)	COLUMN
4	PASSENGER_COUNT	FLOAT	COLUMN
5	TRIP_DISTANCE	FLOAT	COLUMN
6	RATECODEID	FLOAT	COLUMN
7	STORE_AND_FWD_FLAG	VARCHAR(1)	COLUMN
8	PULocationID	NUMBER(38,0)	COLUMN
9	DOLocationID	NUMBER(38,0)	COLUMN
10	PAYMENT_TYPE	NUMBER(38,0)	COLUMN
11	FARE_AMOUNT	FLOAT	COLUMN
12	EXTRA	FLOAT	COLUMN
13	MTA_TAX	FLOAT	COLUMN
14	TIP_AMOUNT	FLOAT	COLUMN
15	TOLLS_AMOUNT	FLOAT	COLUMN
16	IMPROVEMENT_SURCHARGE	FLOAT	COLUMN
17	TOTAL_AMOUNT	FLOAT	COLUMN
18	CONGESTION_SURCHARGE	FLOAT	COLUMN
19	AIRPORT_FEE	FLOAT	COLUMN
20	RUN_ID	NUMBER(38,0)	COLUMN
21	SERVICE_TYPE	VARCHAR(10)	COLUMN
22	SOURCE_YEAR	NUMBER(38,0)	COLUMN
23	SOURCE_MONTH	NUMBER(38,0)	COLUMN
24	INGESTED_AT_UTC	TIMESTAMP_NTZ(9)	COLUMN
25	SOURCE_PATH	VARCHAR(500)	COLUMN
26	BATCH_NUMBER	NUMBER(38,0)	COLUMN

### Tabla RAW GREEN TRIPS:

Object Details for: NY\_TAXI.RAW.RAW\_GREEN\_T...

Table Chart

	name	type	kind
	BATCH_NUMBER	3.7% FLOAT	44.4%
	CONGESTION_SURC...	3.7% NUMBER(38,0)	33.3%
	+25 more	+4 more	CO... 100.0%
1	VENDORID	NUMBER(38,0)	COLUMN
2	LPEP_PICKUP_DATETIME	TIMESTAMP_NTZ(9)	COLUMN
3	LPEP_DROPOFF_DATETIME	TIMESTAMP_NTZ(9)	COLUMN
4	PASSENGER_COUNT	FLOAT	COLUMN
5	TRIP_DISTANCE	FLOAT	COLUMN
6	RATECODEID	FLOAT	COLUMN
7	STORE_AND_FWD_FLAG	VARCHAR(1)	COLUMN
8	PULocationID	NUMBER(38,0)	COLUMN
9	DOLocationID	NUMBER(38,0)	COLUMN
10	PAYMENT_TYPE	NUMBER(38,0)	COLUMN
11	FARE_AMOUNT	FLOAT	COLUMN
12	EXTRA	FLOAT	COLUMN
13	MTA_TAX	FLOAT	COLUMN
14	TIP_AMOUNT	FLOAT	COLUMN
15	TOLLS_AMOUNT	FLOAT	COLUMN
16	IMPROVEMENT_SURCHARGE	FLOAT	COLUMN
17	TOTAL_AMOUNT	FLOAT	COLUMN
18	TRIP_TYPE	NUMBER(38,0)	COLUMN
19	CONGESTION_SURCHARGE	FLOAT	COLUMN
20	EHAIL_FEE	FLOAT	COLUMN
21	RUN_ID	NUMBER(38,0)	COLUMN
22	SERVICE_TYPE	VARCHAR(10)	COLUMN
23	SOURCE_YEAR	NUMBER(38,0)	COLUMN
24	SOURCE_MONTH	NUMBER(38,0)	COLUMN
25	INGESTED_AT_UTC	TIMESTAMP_NTZ(9)	COLUMN
26	SOURCE_PATH	VARCHAR(500)	COLUMN
27	BATCH_NUMBER	NUMBER(38,0)	COLUMN

## 6. Evidencia – Snapshot de OBT:

### Tabla OBT\_TRIPS:

Object Details for: NY\_TAXI.ANALYTICS.OBT\_T...

Table Chart

	name	
	AIRPORT_FEE	2.2%
	AVG_SPEED_MPH	2.2%
	+43 more	
1	TRIP_ID	
2	PICKUP_DATETIME	
3	DROPOFF_DATETIME	
4	PICKUP_DATE	
5	PICKUP_HOUR	
6	DROPOFF_DATE	
7	DROPOFF_HOUR	
8	DAY_OF_WEEK	
9	MONTH	
10	YEAR	
11	PU_LOCATION_ID	

My Workspace

Search for files

+ Add new

Database Explorer

Objects Data Products

Search

Filter

NY\_TAXI

ANALYTICS

Tables 1

OBT\_TRIPS

CLEAN

INFORMATION\_SCHEMA

RAW

SNOWFLAKE

SNOWFLAKE\_LEARNING\_DB

SNOWFLAKE\_SAMPLE\_DATA

	name		type	kind
	AIRPORT_FEE	2.2%	FLOAT 33.3%	
	AVG_SPEED_MPH	2.2%	NUMBE... 28.9%	
	+43 more			CO... 100.0%
1	TRIP_ID		VARCHAR(64)	COLUMN
2	PICKUP_DATETIME		TIMESTAMP_NTZ(9)	COLUMN
3	DROPOFF_DATETIME		TIMESTAMP_NTZ(9)	COLUMN
4	PICKUP_DATE		DATE	COLUMN
5	PICKUP_HOUR		NUMBER(38,0)	COLUMN
6	DROPOFF_DATE		DATE	COLUMN
7	DROPOFF_HOUR		NUMBER(38,0)	COLUMN
8	DAY_OF_WEEK		NUMBER(38,0)	COLUMN
9	MONTH		NUMBER(38,0)	COLUMN
10	YEAR		NUMBER(38,0)	COLUMN
11	PU_LOCATION_ID		NUMBER(38,0)	COLUMN
12	PU_ZONE		VARCHAR(100)	COLUMN
13	PU_BOROUGH		VARCHAR(50)	COLUMN
14	DO_LOCATION_ID		NUMBER(38,0)	COLUMN
15	DO_ZONE		VARCHAR(100)	COLUMN
16	DO_BOROUGH		VARCHAR(50)	COLUMN
17	SERVICE_TYPE		VARCHAR(10)	COLUMN
18	VENDOR_ID		NUMBER(38,0)	COLUMN
19	VENDOR_NAME		VARCHAR(100)	COLUMN
20	RATE_CODE_ID		FLOAT	COLUMN
21	RATE_CODE_DESC		VARCHAR(50)	COLUMN
22	PAYMENT_TYPE		NUMBER(38,0)	COLUMN
23	PAYMENT_TYPE_DESC		VARCHAR(50)	COLUMN
24	TRIP_TYPE		NUMBER(38,0)	COLUMN

	name		type	kind
	AIRPORT_FEE	2.2%	FLOAT 33.3%	
	AVG_SPEED_MPH	2.2%	NUMBE... 28.9%	
	+43 more			CO... 100.0%
21	RATE_CODE_DESC		VARCHAR(50)	COLUMN
22	PAYMENT_TYPE		NUMBER(38,0)	COLUMN
23	PAYMENT_TYPE_DESC		VARCHAR(50)	COLUMN
24	TRIP_TYPE		NUMBER(38,0)	COLUMN
25	TRIP_TYPE_DESC		VARCHAR(50)	COLUMN
26	PASSENGER_COUNT		FLOAT	COLUMN
27	TRIP_DISTANCE		FLOAT	COLUMN
28	STORE_AND_FWD_FLAG		VARCHAR(1)	COLUMN
29	FARE_AMOUNT		FLOAT	COLUMN
30	EXTRA		FLOAT	COLUMN
31	MTA_TAX		FLOAT	COLUMN
32	TIP_AMOUNT		FLOAT	COLUMN
33	TOLLS_AMOUNT		FLOAT	COLUMN
34	IMPROVEMENT_SURCHARGE		FLOAT	COLUMN
35	CONGESTION_SURCHARGE		FLOAT	COLUMN
36	AIRPORT_FEE		FLOAT	COLUMN
37	TOTAL_AMOUNT		FLOAT	COLUMN
38	TRIP_DURATION_MIN		FLOAT	COLUMN
39	AVG_SPEED_MPH		FLOAT	COLUMN
40	TIP_PCT		FLOAT	COLUMN
41	RUN_ID		NUMBER(38,0)	COLUMN
42	INGESTED_AT_UTC		TIMESTAMP_NTZ(9)	COLUMN
43	SOURCE_SERVICE		VARCHAR(10)	COLUMN
44	SOURCE_YEAR		NUMBER(38,0)	COLUMN
45	SOURCE_MONTH		NUMBER(38,0)	COLUMN