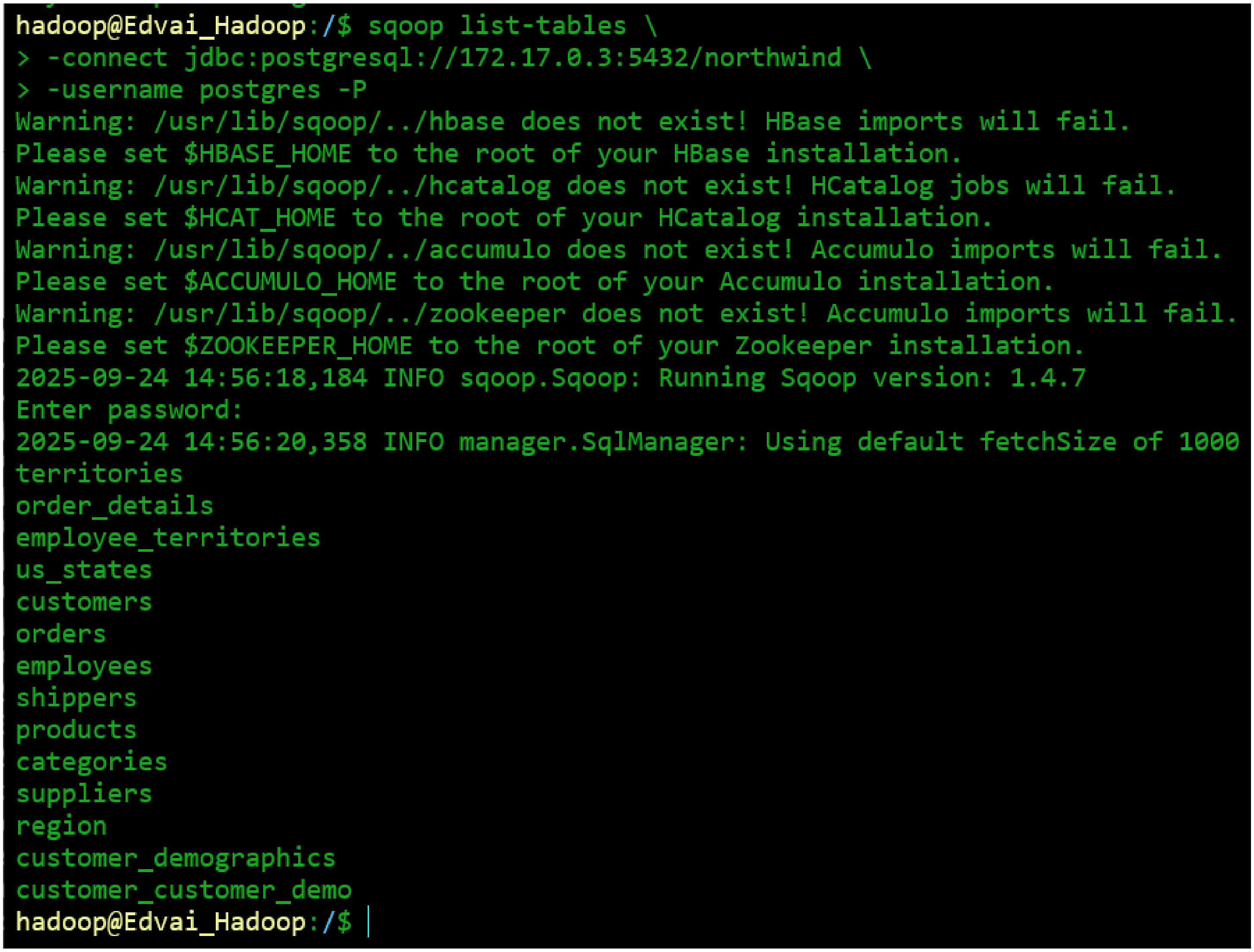
Practica Sqoop

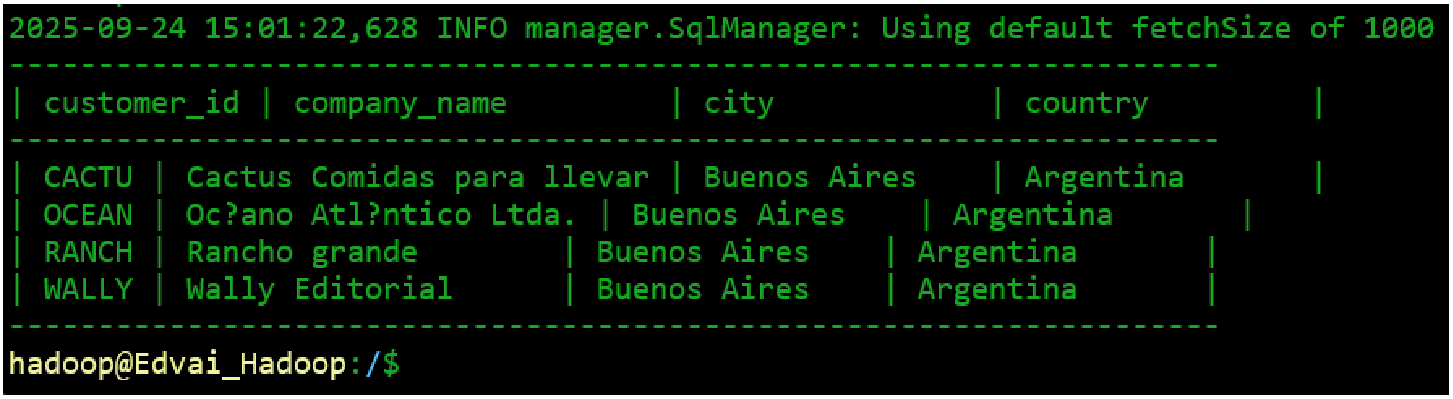
1) Mostrar las tablas de la base de datos northwind



```
hadoop@Edvai_Hadoop:/$ sqoop list-tables \
> -connect jdbc:postgresql://172.17.0.3:5432/northwind \
> -username postgres -P
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2025-09-24 14:56:18,184 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
2025-09-24 14:56:20,358 INFO manager.SqlManager: Using default fetchSize of 1000
territories
order_details
employee_territories
us_states
customers
orders
employees
shippers
products
categories
suppliers
region
customer_demographics
customer_customer_demo
hadoop@Edvai_Hadoop:/$
```

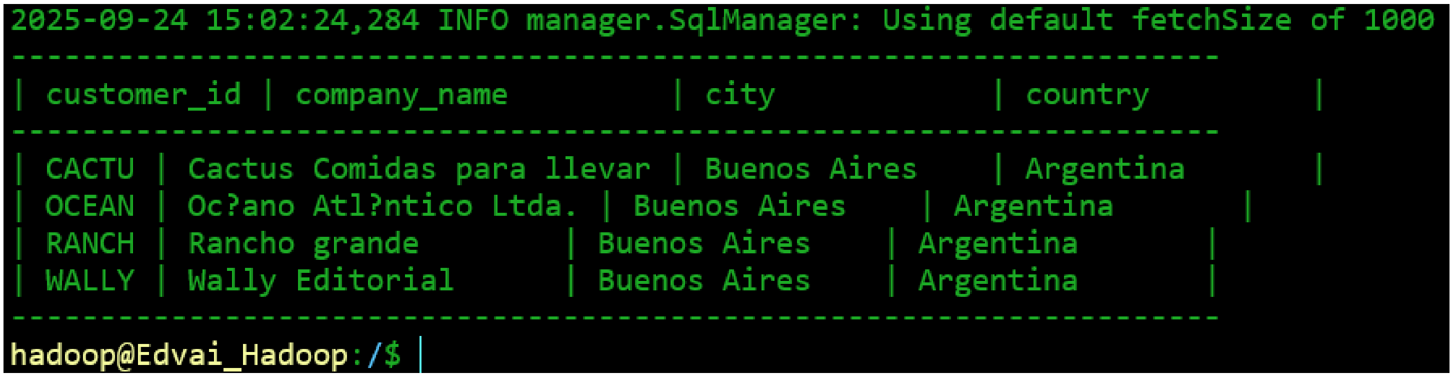
2) Mostrar los clientes de Argentina

1er forma: sqoop eval -connect jdbc:postgresql://172.17.0.3:5432/northwind -username postgres -P -query "select customer_id, company_name, city, country from customers where country = 'Argentina'"



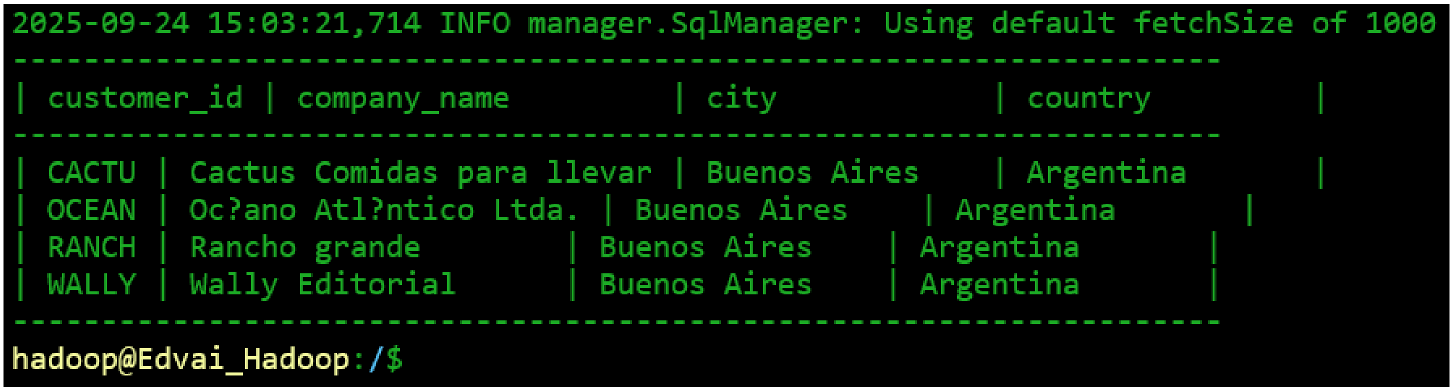
```
2025-09-24 15:01:22,628 INFO manager.SqlManager: Using default fetchSize of 1000
---------------------------------------------------------------------
| customer_id | company_name         | city            | country         |
---------------------------------------------------------------------
| CACTU | Cactus Comidas para llevar | Buenos Aires    | Argentina       |
| OCEAN | Oc?ano Atl?ntico Ltda. | Buenos Aires    | Argentina       |
| RANCH | Rancho grande        | Buenos Aires    | Argentina       |
| WALLY | Wally Editorial      | Buenos Aires    | Argentina       |
---------------------------------------------------------------------
hadoop@Edvai_Hadoop:/$
```

2da forma: sqoop eval -connect jdbc:postgresql://172.17.0.3:5432/northwind -username postgres -P -query "select customer_id, company_name, city, country from customers where LOWER(country) = 'argentina'"



```
2025-09-24 15:02:24,284 INFO manager.SqlManager: Using default fetchSize of 1000
---------------------------------------------------------------------
| customer_id | company_name         | city            | country         |
---------------------------------------------------------------------
| CACTU | Cactus Comidas para llevar | Buenos Aires    | Argentina       |
| OCEAN | Oc?ano Atl?ntico Ltda. | Buenos Aires    | Argentina       |
| RANCH | Rancho grande        | Buenos Aires    | Argentina       |
| WALLY | Wally Editorial      | Buenos Aires    | Argentina       |
---------------------------------------------------------------------
hadoop@Edvai_Hadoop:/$
```

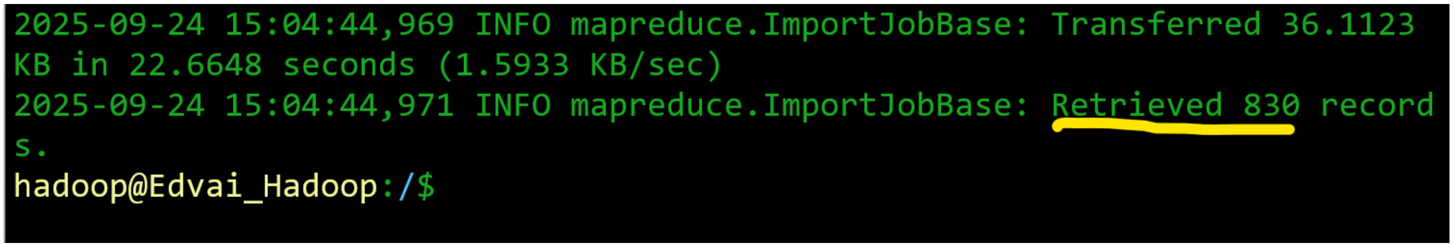
3er forma: sqoop eval -connect jdbc:postgresql://172.17.0.3:5432/northwind -username postgres -P -query "select customer_id, company_name, city, country from customers where country ILIKE 'argentina'"



```
2025-09-24 15:03:21,714 INFO manager.SqlManager: Using default fetchSize of 1000
---------------------------------------------------------------------
| customer_id | company_name         | city            | country         |
---------------------------------------------------------------------
| CACTU | Cactus Comidas para llevar | Buenos Aires    | Argentina       |
| OCEAN | Oc?ano Atl?ntico Ltda. | Buenos Aires    | Argentina       |
| RANCH | Rancho grande        | Buenos Aires    | Argentina       |
| WALLY | Wally Editorial      | Buenos Aires    | Argentina       |
---------------------------------------------------------------------
hadoop@Edvai_Hadoop:/$
```

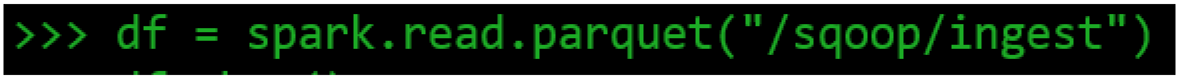
3) Importar un archivo .parquet que contenga toda la tabla orders. Luego ingestar el archivo a HDFS (carpeta /sqoop/ingest)

```
sqoop import -connect jdbc:postgresql://172.17.0.3:5432/northwind -username
    postgres -table orders -m 1 -P -target-dir /sqoop/ingest -as-parquetfile
    -delete-target-dir
```

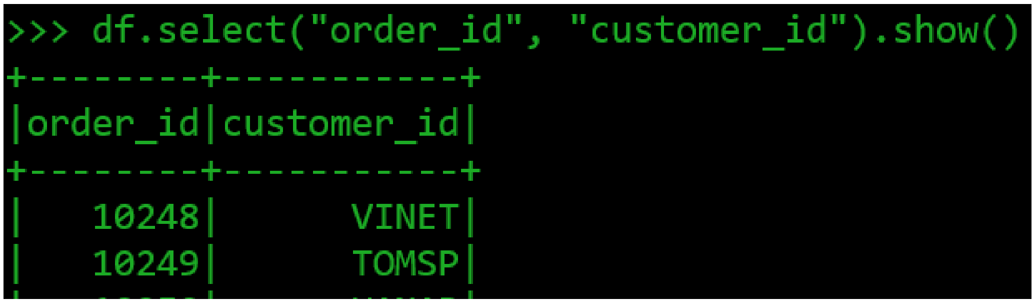


```
2025-09-24 15:04:44,969 INFO mapreduce.ImportJobBase: Transferred 36.1123
KB in 22.6648 seconds (1.5933 KB/sec)
2025-09-24 15:04:44,971 INFO mapreduce.ImportJobBase: Retrieved 830 record
s.
hadoop@Edvai_Hadoop:/$
```

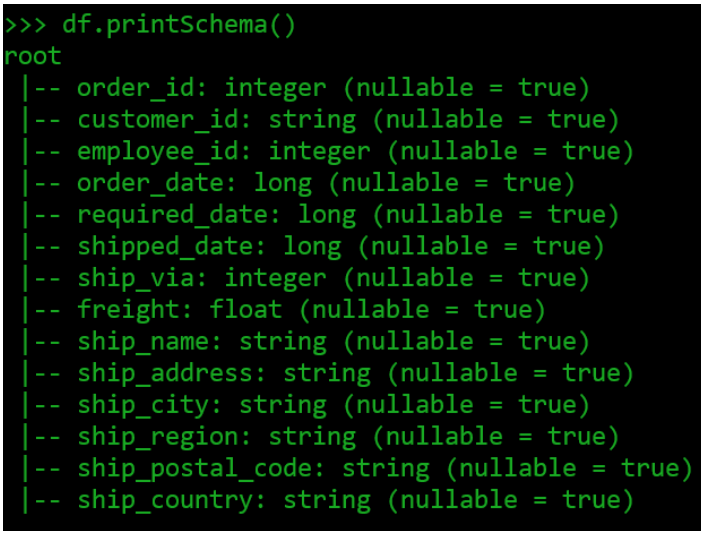
Para confirmar en Spark que el archivo esté ok:



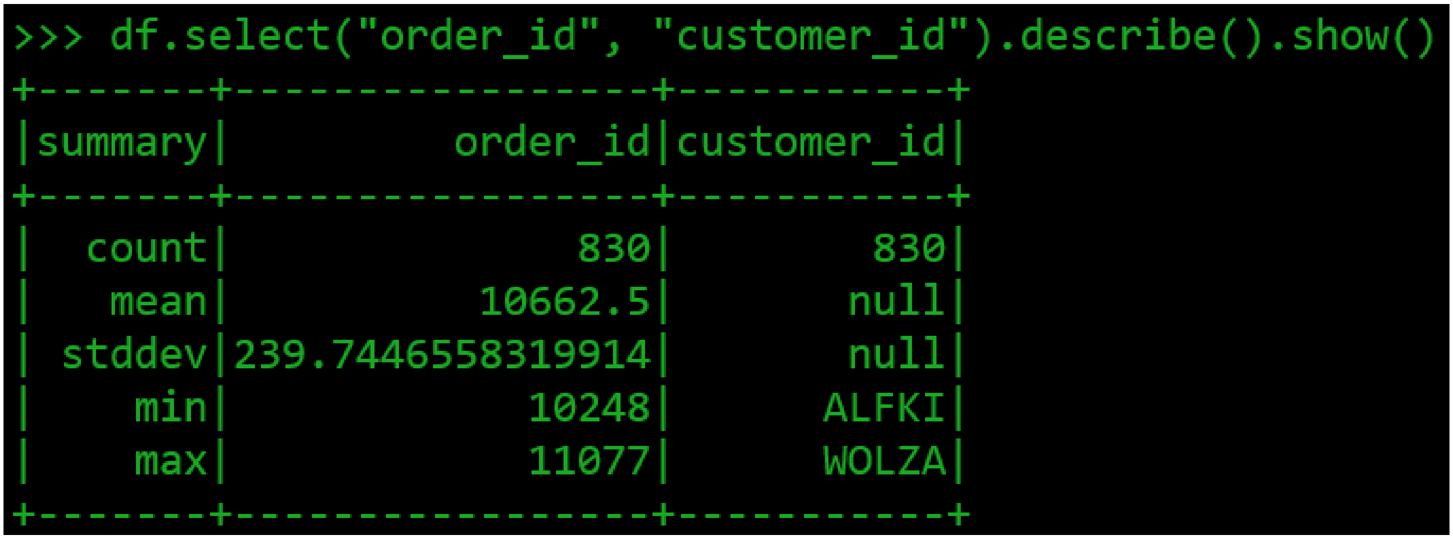
```
>>> df = spark.read.parquet("/sqoop/ingest")
```



```
>>> df.select("order_id", "customer_id").show()
+--------+-----------+
|order_id|customer_id|
+--------+-----------+
|   10248|      VINET|
|   10249|      TOMSP|
```



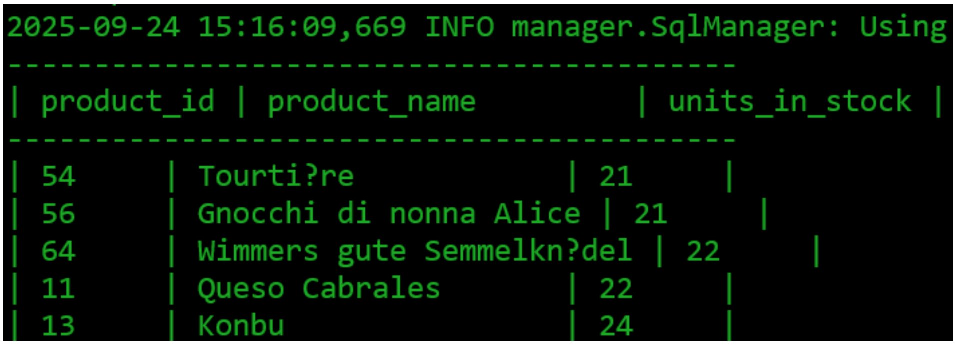
```
>>> df.printSchema()
root
 |-- order_id: integer (nullable = true)
 |-- customer_id: string (nullable = true)
 |-- employee_id: integer (nullable = true)
 |-- order_date: long (nullable = true)
 |-- required_date: long (nullable = true)
 |-- shipped_date: long (nullable = true)
 |-- ship_via: integer (nullable = true)
 |-- freight: float (nullable = true)
 |-- ship_name: string (nullable = true)
 |-- ship_address: string (nullable = true)
 |-- ship_city: string (nullable = true)
 |-- ship_region: string (nullable = true)
 |-- ship_postal_code: string (nullable = true)
 |-- ship_country: string (nullable = true)
```



```
>>> df.select("order_id", "customer_id").describe().show()
+-------+-----------------+-----------+
|summary|         order_id|customer_id|
+-------+-----------------+-----------+
|  count|              830|        830|
|   mean|          10662.5|       null|
| stddev|239.7446558319914|       null|
|    min|            10248|      ALFKI|
|    max|            11077|      WOLZA|
+-------+-----------------+-----------+
```

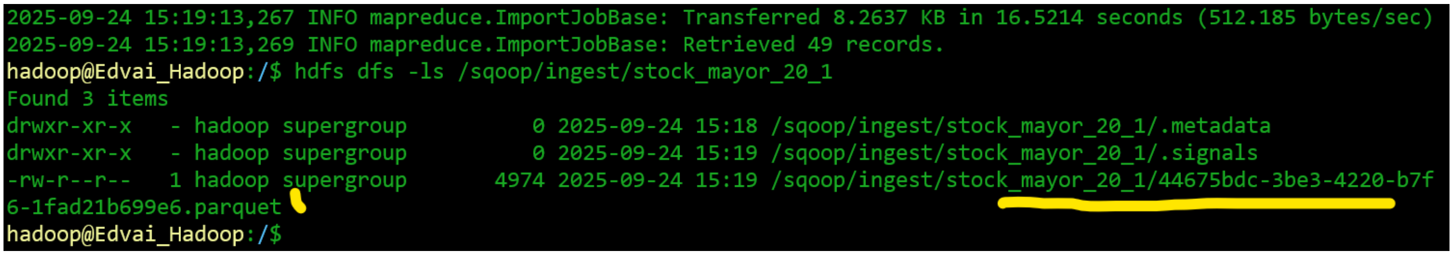
4) Importar un archivo .parquet que contenga solo los productos con mas 20 unidades en stock, de la tabla Products . Luego ingestar el archivo a HDFS (carpeta ingest)

Primero muestro la consulta: sqoop eval -connect jdbc:postgresql://172.17.0.3:5432/northwind -username postgres -P -query "select product_id, product_name, units_in_stock from products where units_in_stock > 20 order by units_in_stock"

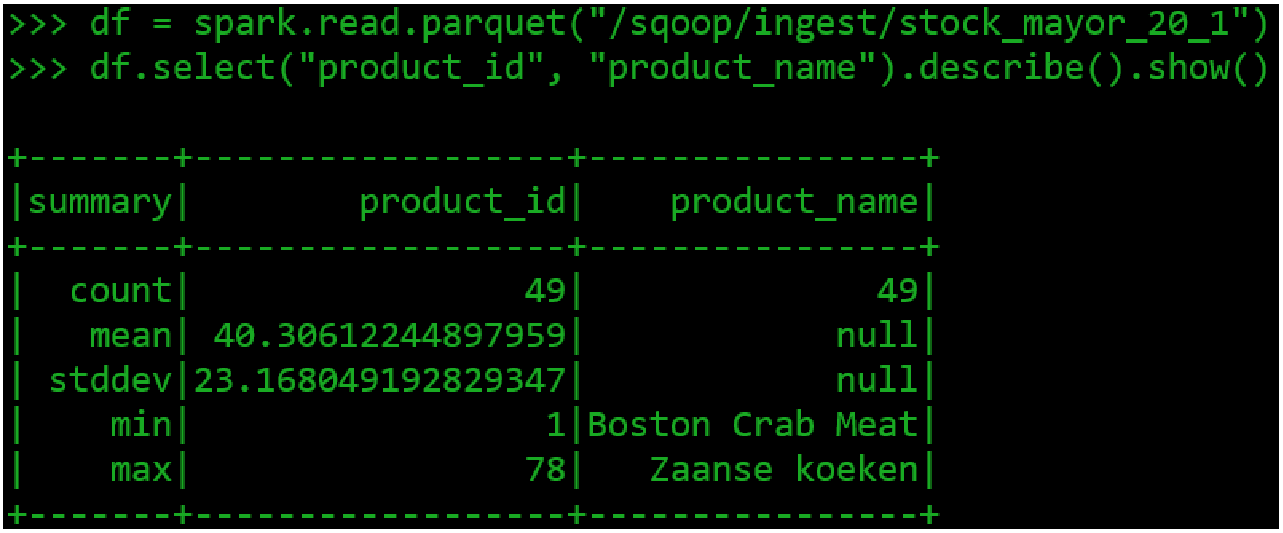


```
2025-09-24 15:16:09,669 INFO manager.SqlManager: Using
-----------------------------------------
| product_id | product_name         | units_in_stock |
-----------------------------------------
| 54         | Tourti?re            | 21         |
| 56         | Gnocchi di nonna Alice | 21         |
| 64         | Wimmers gute Semmelkn?del | 22         |
| 11         | Queso Cabrales       | 22         |
| 13         | Konbu                | 24         |
```

Ingesta 1era forma filtro where: sqoop import -connect jdbc:postgresql://172.17.0.3:5432/northwind -username postgres -table products -m 1 -P -target-dir /sqoop/ingest/stock_mayor_20_1 -as-parquetfile -where "units_in_stock > 20" -delete-target-dir

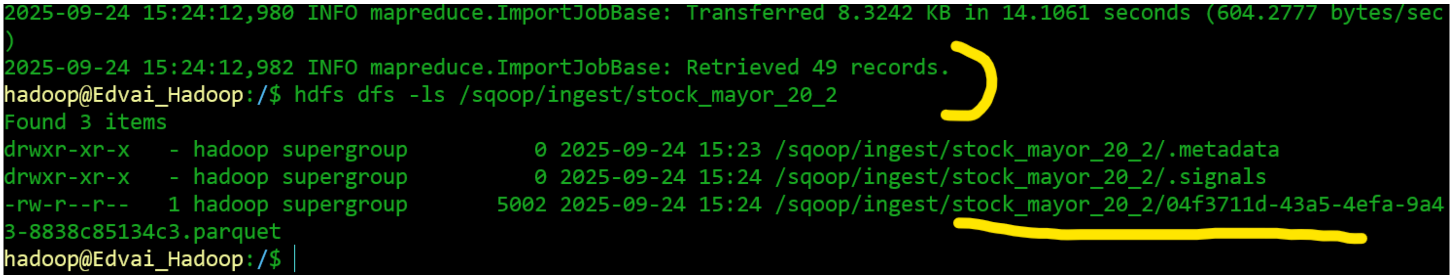


```
2025-09-24 15:19:13,267 INFO mapreduce.ImportJobBase: Transferred 8.2637 KB in 16.5214 seconds (512.185 bytes/sec)
2025-09-24 15:19:13,269 INFO mapreduce.ImportJobBase: Retrieved 49 records.
hadoop@Edvai_Hadoop:/$ hdfs dfs -ls /sqoop/ingest/stock_mayor_20_1
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 15:18 /sqoop/ingest/stock_mayor_20_1/.metadata
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 15:19 /sqoop/ingest/stock_mayor_20_1/.signals
-rw-r--r--   1 hadoop supergroup       4974 2025-09-24 15:19 /sqoop/ingest/stock_mayor_20_1/44675bdc-3be3-4220-b7f
6-1fad21b699e6.parquet
hadoop@Edvai_Hadoop:/$
```

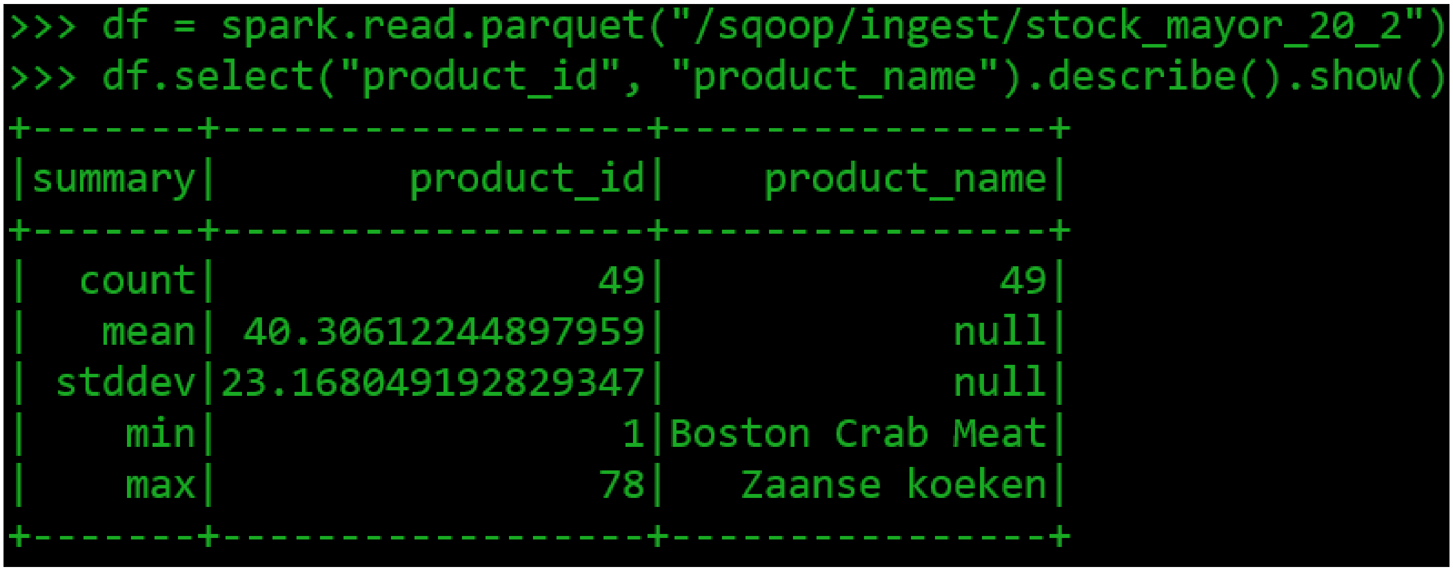


```
>>> df = spark.read.parquet("/sqoop/ingest/stock_mayor_20_1")
>>> df.select("product_id", "product_name").describe().show()

+-------+------------------+----------------+
|summary|        product_id|    product_name|
+-------+------------------+----------------+
|  count|                49|              49|
|   mean| 40.30612244897959|            null|
| stddev|23.168049192829347|            null|
|    min|                 1|Boston Crab Meat|
|    max|                78|   Zaanse koeken|
+-------+------------------+----------------+
```

Ingesta 2da forma con query: sqoop import -connect jdbc:postgresql://172.17.0.3:5432/northwind -username postgres -query "select * from products where units_in_stock > 20 AND \$CONDITIONS" -m 1 -P -target-dir /sqoop/ingest/stock_mayor_20_2 -as-parquetfile -delete-target-dir

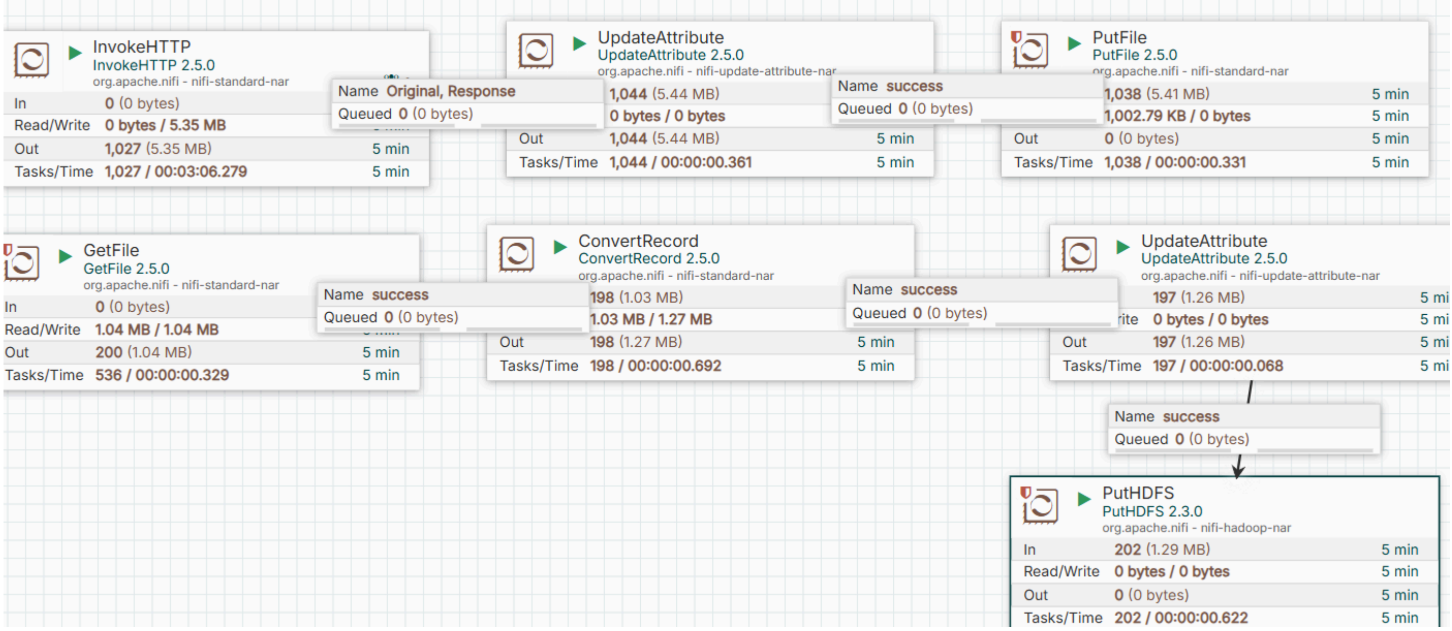


```
2025-09-24 15:24:12,980 INFO mapreduce.ImportJobBase: Transferred 8.3242 KB in 14.1061 seconds (604.2777 bytes/sec
)
2025-09-24 15:24:12,982 INFO mapreduce.ImportJobBase: Retrieved 49 records.
hadoop@Edvai_Hadoop:/$ hdfs dfs -ls /sqoop/ingest/stock_mayor_20_2
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 15:23 /sqoop/ingest/stock_mayor_20_2/.metadata
drwxr-xr-x   - hadoop supergroup          0 2025-09-24 15:24 /sqoop/ingest/stock_mayor_20_2/.signals
-rw-r--r--   1 hadoop supergroup       5002 2025-09-24 15:24 /sqoop/ingest/stock_mayor_20_2/04f3711d-43a5-4efa-9a4
3-8838c85134c3.parquet
hadoop@Edvai_Hadoop:/$
```



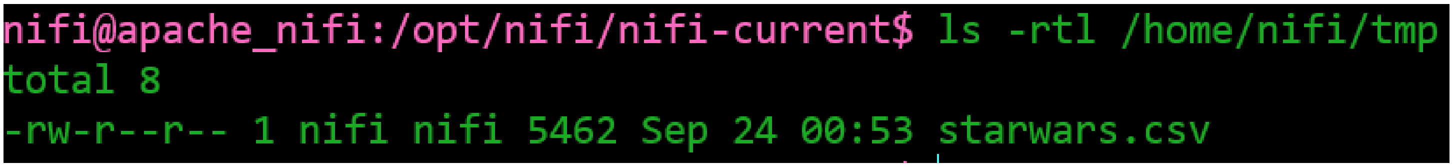
```
>>> df = spark.read.parquet("/sqoop/ingest/stock_mayor_20_2")
>>> df.select("product_id", "product_name").describe().show()
+-------+------------------+----------------+
|summary|        product_id|    product_name|
+-------+------------------+----------------+
|  count|                49|              49|
|   mean| 40.30612244897959|            null|
| stddev|23.168049192829347|            null|
|    min|                 1|Boston Crab Meat|
|    max|                78|   Zaanse koeken|
+-------+------------------+----------------+
```

Práctica Nifi



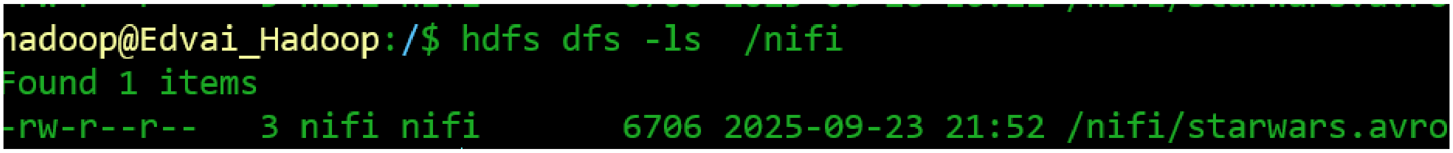
```
InvokeHTTP
InvokeHTTP 2.5.0
org.apache.nifi - nifi-standard-nar
In          0 (0 bytes)
Read/Write  0 bytes / 5.35 MB
Out         1,027 (5.35 MB)              5 min
Tasks/Time  1,027 / 00:03:06.279         5 min

Name  Original, Response
Queued  0 (0 bytes)

UpdateAttribute
UpdateAttribute 2.5.0
org.apache.nifi - nifi-update-attribute-nar
            1,044 (5.44 MB)
            0 bytes / 0 bytes
Out         1,044 (5.44 MB)              5 min
Tasks/Time  1,044 / 00:00:00.361         5 min

Name  success
Queued  0 (0 bytes)

PutFile
PutFile 2.5.0
org.apache.nifi - nifi-standard-nar
            1,038 (5.41 MB)              5 min
            1,002.79 KB / 0 bytes        5 min
Out         0 (0 bytes)                  5 min
Tasks/Time  1,038 / 00:00:00.331         5 min

GetFile
GetFile 2.5.0
org.apache.nifi - nifi-standard-nar
In          0 (0 bytes)
Read/Write  1.04 MB / 1.04 MB
Out         200 (1.04 MB)                5 min
Tasks/Time  536 / 00:00:00.329           5 min

Name  success
Queued  0 (0 bytes)

ConvertRecord
ConvertRecord 2.5.0
org.apache.nifi - nifi-standard-nar
            198 (1.03 MB)
            1.03 MB / 1.27 MB
Out         198 (1.27 MB)                5 min
Tasks/Time  198 / 00:00:00.692           5 min

Name  success
Queued  0 (0 bytes)

UpdateAttribute
UpdateAttribute 2.5.0
org.apache.nifi - nifi-update-attribute-nar
            197 (1.26 MB)                5 mi
       rite 0 bytes / 0 bytes            5 mi
Out         197 (1.26 MB)                5 mi
Tasks/Time  197 / 00:00:00.068           5 mi

Name  success
Queued  0 (0 bytes)

PutHDFS
PutHDFS 2.3.0
org.apache.nifi - nifi-hadoop-nar
In          202 (1.29 MB)                5 min
Read/Write  0 bytes / 0 bytes            5 min
Out         0 (0 bytes)                  5 min
Tasks/Time  202 / 00:00:00.622           5 min
```

starwars.csv :



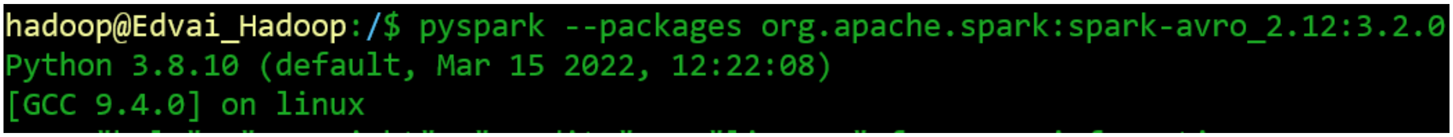
```
nifi@apache_nifi:/opt/nifi/nifi-current$ ls -rtl /home/nifi/tmp
total 8
-rw-r--r-- 1 nifi nifi 5462 Sep 24 00:53 starwars.csv
```

starwars.avro:

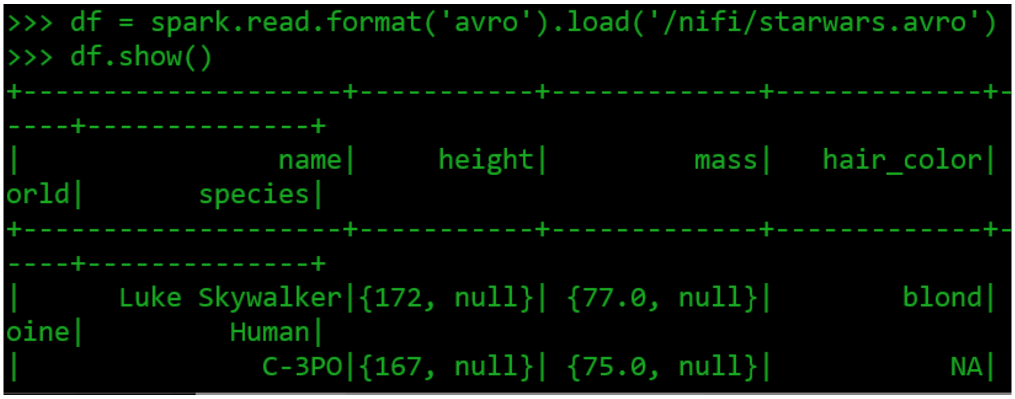


```
hadoop@Edvai_Hadoop:/$ hdfs dfs -ls  /nifi
Found 1 items
-rw-r--r--   3 nifi nifi       6706 2025-09-23 21:52 /nifi/starwars.avro
```

Lectura con pyspark:



```
hadoop@Edvai_Hadoop:/$ pyspark --packages org.apache.spark:spark-avro_2.12:3.2.0
Python 3.8.10 (default, Mar 15 2022, 12:22:08)
[GCC 9.4.0] on linux
```



```
>>> df = spark.read.format('avro').load('/nifi/starwars.avro')
>>> df.show()
+--------------------+-----------+-------------+-------------+-
----+--------------+
|                name|     height|         mass|   hair_color|
orld|       species|
+--------------------+-----------+-------------+-------------+-
----+--------------+
|      Luke Skywalker|{172, null}| {77.0, null}|        blond|
oine|         Human|
|               C-3PO|{167, null}| {75.0, null}|           NA|
```