Ejercitación Sqoop

1) Mostrar las tablas de la base de datos northwind.

Para mostrar las tablas de la base de datos northwind, se ejecuta lo siguiente en bash:

```
sqoop list-tables \
--connect jdbc:postgresql://172.17.0.3:5432/northwind \
--username postgres -P
```

```
hadoop@401bec58e4c6:/$ 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 $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-04-24 08:11:05,350 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
2024-04-24 08:11:09,291 INFO manager.SqlManager: Using default fetchSize of 1000
territories
order_details
employee_territories
us_states
customers
orders
employees
shippers
products
categories
suppliers
region
customer_demographics
customer_customer_demo
hadoop@401bec58e4c6:/$
```

2) Mostrar los clientes de Argentina.

Para mostrar los clientes de Argentina, se ejecuta lo siguiente en bash:

```
sqoop eval --connect jdbc:postgresql://172.17.0.3:5432/northwind \
--username postgres --P \
--query "select c.company_name , c.customer_id , c.postal_code
from customers c
where c.country like 'Argentina'"
```

```
-----------------------------------------
| company_name         | customer_id | postal_code |
-----------------------------------------
| Cactus Comidas para llevar | CACTU | 1010       |
| Oc?ano Atl?ntico Ltda. | OCEAN | 1010       |
| Rancho grande        | RANCH | 1010       |
-----------------------------------------
hadoop@401bec58e4c6:/$
```

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

Para importar la tabla orders e ingresarla a HDFS (/sqoop/ingest/northwind/orders), se ejecuta lo siguiente en bash:

```
sqoop import \
--connect jdbc:postgresql://172.17.0.3:5432/northwind \
```

```
--username postgres \
--table orders \
--m 1 \
--P \
--target-dir /sqoop/ingest/northwind/orders \
--as-parquetfile \
--delete-target-dir
```

```
hadoop@401bec58e4c6:/$ hdfs dfs -ls /sqoop/ingest
Found 4 items
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 08:45 /sqoop/ingest/.metadata
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 08:45 /sqoop/ingest/.signals
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 09:37 /sqoop/ingest/.temp
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 09:40 /sqoop/ingest/northwind
hadoop@401bec58e4c6:/$ hdfs dfs -ls /sqoop/ingest/northwind
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 09:40 /sqoop/ingest/northwind/orders
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 09:38 /sqoop/ingest/northwind/products_20
hadoop@401bec58e4c6:/$ hdfs dfs -ls /sqoop/ingest/northwind/orders
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 09:40 /sqoop/ingest/northwind/orders/.metadata
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 09:40 /sqoop/ingest/northwind/orders/.signals
-rw-r--r--   1 hadoop supergroup      32390 2024-04-24 09:40 /sqoop/ingest/northwind/orders/48aa5156-b3eb-4ead-9353-5804d5cfea3e.parquet
```

```
df = spark.read.parquet("/sqoop/ingest/northwind/orders/*.parquet")
df.show(5)
```

```
>>> df = spark.read.parquet("/sqoop/ingest/northwind/orders/*.parquet")
>>> df.show(5)
+--------+----------+-----------+-----------+-------------+-------------+--------+-------+--------------------+--------------------+-------------+----------+---------------+-----------+
|order_id|customer_id|employee_id| order_date|required_date|shipped_date|ship_via|freight|           ship_name|        ship_address|    ship_city|ship_region|ship_postal_code|ship_country|
+--------+----------+-----------+-----------+-------------+------------+--------+-------+--------------------+--------------------+-------------+----------+---------------+-----------+
|   10248|     VINET|          5|836449200000|838868400000|837486000000|       3|  32.38|Vins et alcools C...|  59 rue de l'Abbaye|        Reims|      null|          51100|     France|
|   10249|     TOMSP|          6|836535600000|840164400000|836967600000|       1|  11.61|  Toms Spezialitäten|       Luisenstr. 48|      MÜnster|      null|          44087|    Germany|
|   10250|     HANAR|          4|836794800000|839214000000|837140400000|       2|  65.83|       Hanari Carnes|     Rua do Paço, 67|Rio de Janeiro|        RJ|       05454-876|     Brazil|
|   10251|     VICTE|          3|836794800000|839214000000|837399600000|       1|  41.34|Victuailles en stock|  2, rue du Commerce|         Lyon|      null|          69004|     France|
|   10252|     SUPRD|          4|836881200000|839300400000|837054000000|       2|   51.3|    Suprêmes délices|Boulevard Tirou, 255|     Charleroi|      null|           B-6000|    Belgium|
+--------+----------+-----------+-----------+-------------+------------+--------+-------+--------------------+--------------------+-------------+----------+---------------+-----------+
```

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).

Para importar la tabla orders e ingresarla a HDFS (/sqoop/ingest/northwind/products_20), se ejecuta lo siguiente en bash:

```
sqoop import \
--connect jdbc:postgresql://172.17.0.3:5432/northwind \
-username postgres \
--table products \
--m 1 \
--P \
--target-dir /sqoop/ingest/northwind/products_20 \
--as-parquetfile \
--where "units_in_stock > 20" \
--delete-target-dir
```

```
hadoop@401bec58e4c6:/$ hdfs dfs -ls /sqoop/ingest/northwind/products_20
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 09:38 /sqoop/ingest/northwind/products_20/.metadata
drwxr-xr-x   - hadoop supergroup          0 2024-04-24 09:38 /sqoop/ingest/northwind/products_20/.signals
-rw-r--r--   1 hadoop supergroup       4974 2024-04-24 09:38 /sqoop/ingest/northwind/products_20/e7da9e85-7911-4899-8455-5a4a77e56cf6.parquet
```

```
>>> df = spark.read.parquet("/sqoop/ingest/northwind/products_20/*.parquet")
>>> df.show(5)
+----------+--------------------+-----------+-----------+--------------------+----------+--------------+-------------+-------------+------------+
|product_id|        product_name|supplier_id|category_id| quantity_per_unit|unit_price|units_in_stock|units_on_order|reorder_level|discontinued|
+----------+--------------------+-----------+-----------+--------------------+----------+--------------+-------------+-------------+------------+
|         1|                Chai|          8|          1|10 boxes x 30 bags|      18.0|            39|            0|           10|           1|
|         4|Chef Anton's Caju...|          2|          2|    48 - 6 oz jars|      22.0|            53|            0|            0|           0|
|         6|Grandma's Boysenb...|          3|          2|    12 - 8 oz jars|      25.0|           120|            0|           25|           0|
|         9|     Mishi Kobe Niku|          4|          6|  18 - 500 g pkgs.|      97.0|            29|            0|            0|           1|
|        10|               Ikura|          4|          8|  12 - 200 ml jars|      31.0|            31|            0|            0|           0|
+----------+--------------------+-----------+-----------+--------------------+----------+--------------+-------------+-------------+------------+
```

Ejercitación Nifi

Se crea el script ingest.sh en /home/nifi/bucket, este se encarga de descargar el archivo starwars.csv:

```
wget https://raw.githubusercontent.com/fpineyro/homework-0/master/starwars.csv
```

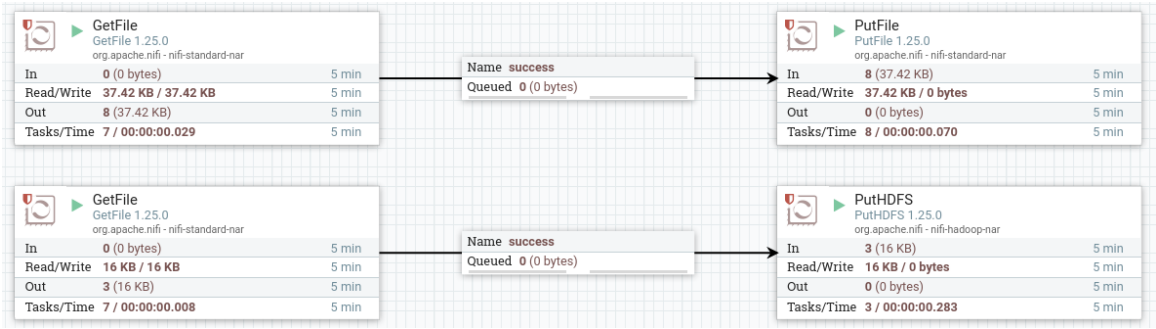
En /home/nifi se crean los directorios bucket, ingest y hadoop:

```
mkdir hadoop
mkdir ingest
mkdir bucket
```

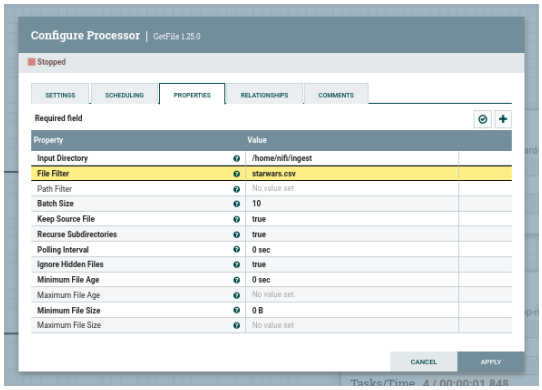
Se crearon los archivos core-site.xml y hdfs-site.xml en /home/nifi/hadoop y se empenzó a trabajar en la plataforma de nifi accediendo a:

https://localhost:8443/nifi

Donde se crearon los procesos de Nifi solicitados:



Tener en cuenta que al haber creado el script ingest.sh donde también está el archivo starwars.csv, en el filtro del GetFile hay que poner que solamente acepte archivos .csv, o solamente el archivo starwars.csv.



Se puede acceder a los archivos en hdfs contenidos en /nifi con el comando

```
hdfs dfs -ls /nifi
```

Cuya salida es:

```
hadoop@401bec58e4c6:~$ hdfs dfs -ls /nifi
Found 1 items
-rw-r--r--   1 nifi supergroup       5462 2024-04-24 10:50 /nifi/starwars.csv
```