Ejercicios:

1) Mostrar las tablas de la base de datos northwind

```
hadoop@615cf53bef6c:/$ sqoop list-tables \
bc:postgresql://172.17.0.3:5432/> --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 $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2025-09-23 21:32:32,953 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
2025-09-23 21:32:36,761 INFO manager.SqlManager: Using default fetchSize of 1000
territories
order_details
employee_territories
us_states
customers
orders
employees
shippers
products
categories
suppliers
region
customer_demographics
customer_customer_demo
customers3
hadoop@615cf53bef6c:/$
```

2) Mostrar los clientes de Argentina

```
hadoop@615cf53bef6c:/$ sqoop eval --connect jdbc:postgresql://172.17.0.3:5432/northwind --username postgres --P --query "select customer_id, company_name, c
ountry from customers where country = 'Argentina'"
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 $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2025-09-23 21:44:48,375 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
2025-09-23 21:44:51,790 INFO manager.SqlManager: Using default fetchSize of 1000
------------------------------------------------
| customer_id | company_name          | country      |
------------------------------------------------
| CACTU | Cactus Comidas para llevar | Argentina    |
| OCEAN | Oc?ano Atl?ntico Ltda. | Argentina      |
| RANCH | Rancho grande          | Argentina      |
------------------------------------------------
hadoop@615cf53bef6c:/$
```

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

```
hadoop@615cf53bef6c:/$ sqoop import --connect jdbc:postgresql://172.17.0.3:5432/northwind --username postgres --table orders --m 1 --P --target-dir /sqoop/i
ngest --as-parquetfile --delete-target-dir
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.
2025-09-23 22:05:24,238 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
2025-09-23 22:05:27,636 INFO manager.SqlManager: Using default fetchSize of 1000
2025-09-23 22:05:27,637 INFO tool.CodeGenTool: Beginning code generation
2025-09-23 22:05:27,637 INFO tool.CodeGenTool: Will generate java class as codegen_orders
2025-09-23 22:05:28,150 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "orders" AS t LIMIT 1
2025-09-23 22:05:28,240 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-hadoop/compile/9d5e37256e62ed97ae2c76283d9ae577/codegen_orders.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2025-09-23 22:05:36,154 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/9d5e37256e62ed97ae2c76283d9ae577/codegen_orders.jar
2025-09-23 22:05:37,497 INFO tool.ImportTool: Destination directory /sqoop/ingest is not present, hence not deleting.
2025-09-23 22:05:37,497 WARN manager.PostgresqlManager: It looks like you are importing from postgresql.
2025-09-23 22:05:37,497 WARN manager.PostgresqlManager: This transfer can be faster! Use the --direct
2025-09-23 22:05:37,497 WARN manager.PostgresqlManager: option to exercise a postgresql-specific fast path.
2025-09-23 22:05:37,510 INFO mapreduce.ImportJobBase: Beginning import of orders
2025-09-23 22:05:37,512 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-09-23 22:05:37,526 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2025-09-23 22:05:37,591 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "orders" AS t LIMIT 1
2025-09-23 22:05:37,596 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "orders" AS t LIMIT 1
2025-09-23 22:05:38,092 WARN spi.Registration: Not loading URI patterns in org.kitesdk.data.spi.hive.Loader
2025-09-23 22:05:40,098 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2025-09-23 22:05:40,264 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-09-23 22:05:41,301 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1758672703912_00
01
2025-09-23 22:05:44,654 INFO db.DBInputFormat: Using read commited transaction isolation
2025-09-23 22:05:44,741 INFO mapreduce.JobSubmitter: number of splits:1
2025-09-23 22:05:45,076 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1758672703912_0001
2025-09-23 22:05:45,076 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-09-23 22:05:45,538 INFO conf.Configuration: resource-types.xml not found
2025-09-23 22:05:45,538 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-09-23 22:05:46,575 INFO impl.YarnClientImpl: Submitted application application_1758672703912_0001
```

```
2025-09-23 22:05:46,575 INFO impl.YarnClientImpl: Submitted application application_1758672703912_0001
2025-09-23 22:05:46,679 INFO mapreduce.Job: The url to track the job: http://615cf53bef6c:8088/proxy/application_1758672703912_0001/
2025-09-23 22:05:46,681 INFO mapreduce.Job: Running job: job_1758672703912_0001
2025-09-23 22:06:03,069 INFO mapreduce.Job: Job job_1758672703912_0001 running in uber mode : false
2025-09-23 22:06:03,072 INFO mapreduce.Job:   map 0% reduce 0%
2025-09-23 22:06:14,242 INFO mapreduce.Job:   map 100% reduce 0%
2025-09-23 22:06:14,260 INFO mapreduce.Job: Job job_1758672703912_0001 completed successfully
2025-09-23 22:06:14,400 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=276467
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=18876
                HDFS: Number of bytes written=36979
                HDFS: Number of read operations=50
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=10
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=8185
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=8185
                Total vcore-milliseconds taken by all map tasks=8185
                Total megabyte-milliseconds taken by all map tasks=12572160
        Map-Reduce Framework
                Map input records=830
                Map output records=830
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=69
                CPU time spent (ms)=6030
                Physical memory (bytes) snapshot=296792064
                Virtual memory (bytes) snapshot=3007361024
                Total committed heap usage (bytes)=155189248
                Peak Map Physical memory (bytes)=296792064
                Peak Map Virtual memory (bytes)=3007361024
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2025-09-23 22:06:14,413 INFO mapreduce.ImportJobBase: Transferred 36.1123 KB in 34.2966 seconds (1.0529 KB/sec)
2025-09-23 22:06:14,421 INFO mapreduce.ImportJobBase: Retrieved 830 records.
hadoop@615cf53bef6c:/$
```

A continuación mostramos el contenido de /sqoop/ingest y luego renombramos el archivo como **orders.parquet**

```
hadoop@615cf53bef6c:~$ hdfs dfs -ls /sqoop/ingest
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2025-09-23 22:05 /sqoop/ingest/.metadata
drwxr-xr-x   - hadoop supergroup          0 2025-09-23 22:06 /sqoop/ingest/.signals
-rw-r--r--   1 hadoop supergroup      32390 2025-09-23 22:06 /sqoop/ingest/5b41c8b3-c874-40a1-a769-ddb88c338636.parquet
hadoop@615cf53bef6c:~$ hdfs dfs -mv /sqoop/ingest/5b41c8b3-c874-40a1-a769-ddb88c338636.parquet /sqoop/ingest/orders.parquet
hadoop@615cf53bef6c:~$ hdfs dfs -ls /sqoop/ingest
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2025-09-23 22:05 /sqoop/ingest/.metadata
drwxr-xr-x   - hadoop supergroup          0 2025-09-23 22:06 /sqoop/ingest/.signals
-rw-r--r--   1 hadoop supergroup      32390 2025-09-23 22:06 /sqoop/ingest/orders.parquet
hadoop@615cf53bef6c:~$
```

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)

```
hadoop@615cf53bef6c:~$ sqoop import --connect jdbc:postgresql://172.17.0.3:5432/northwind --username postgres --table products -m 1 -P --target-dir /sqoop/ingest --as-parquetfile --where
"
units_in_stock > 20" --delete-target-dir
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.
2025-09-23 23:13:31,631 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
2025-09-23 23:13:34,947 INFO manager.SqlManager: Using default fetchSize of 1000
2025-09-23 23:13:34,948 INFO tool.CodeGenTool: Beginning code generation
2025-09-23 23:13:34,948 INFO tool.CodeGenTool: Will generate java class as codegen_products
2025-09-23 23:13:35,463 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "products" AS t LIMIT 1
2025-09-23 23:13:35,546 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-hadoop/compile/c08538e762b209460334f92dc229326f/codegen_products.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2025-09-23 23:13:39,718 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/c08538e762b209460334f92dc229326f/codegen_products.jar
2025-09-23 23:13:41,341 INFO tool.ImportTool: Destination directory /sqoop/ingest deleted.
2025-09-23 23:13:41,342 WARN manager.PostgresqlManager: It looks like you are importing from postgresql.
2025-09-23 23:13:41,342 WARN manager.PostgresqlManager: This transfer can be faster! Use the --direct
2025-09-23 23:13:41,342 WARN manager.PostgresqlManager: option to exercise a postgresql-specific fast path.
2025-09-23 23:13:41,357 INFO mapreduce.ImportJobBase: Beginning import of products
2025-09-23 23:13:41,358 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-09-23 23:13:41,370 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2025-09-23 23:13:41,461 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "products" AS t LIMIT 1
2025-09-23 23:13:41,466 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "products" AS t LIMIT 1
2025-09-23 23:13:41,803 WARN spi.Registration: Not loading URI patterns in org.kitesdk.data.spi.hive.Loader
2025-09-23 23:13:43,088 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2025-09-23 23:13:43,253 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2025-09-23 23:13:43,942 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1758672703912_0002
2025-09-23 23:13:45,786 INFO db.DBInputFormat: Using read committed transaction isolation
2025-09-23 23:13:45,875 INFO mapreduce.JobSubmitter: number of splits:1
2025-09-23 23:13:46,180 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1758672703912_0002
2025-09-23 23:13:46,180 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-09-23 23:13:46,612 INFO conf.Configuration: resource-types.xml not found
2025-09-23 23:13:46,612 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-09-23 23:13:46,784 INFO impl.YarnClientImpl: Submitted application application_1758672703912_0002
2025-09-23 23:13:46,846 INFO mapreduce.Job: The url to track the job: http://615cf53bef6c:8088/proxy/application_1758672703912_0002/
2025-09-23 23:13:46,848 INFO mapreduce.Job: Running job: job_1758672703912_0002
2025-09-23 23:13:59,222 INFO mapreduce.Job: Job job_1758672703912_0002 running in uber mode : false
2025-09-23 23:13:59,225 INFO mapreduce.Job:  map 0% reduce 0%
2025-09-23 23:14:09,596 INFO mapreduce.Job:  map 100% reduce 0%
2025-09-23 23:14:10,622 INFO mapreduce.Job: Job job_1758672703912_0002 completed successfully
2025-09-23 23:14:10,795 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
```

Mostramos el contenido de /sqoop/ingest y renombramos el archivo parquet a **productos_mas_de_20.parquet**

```
hadoop@615cf53bef6c:~$ hdfs dfs -ls /sqoop/ingest
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2025-09-23 23:13 /sqoop/ingest/.metadata
drwxr-xr-x   - hadoop supergroup          0 2025-09-23 23:14 /sqoop/ingest/.signals
-rw-r--r--   1 hadoop supergroup       4974 2025-09-23 23:14 /sqoop/ingest/62dfddfd-43a0-4395-ae1c-28bd72da56c1.parquet
hadoop@615cf53bef6c:~$ hdfs dfs -mv /sqoop/ingest/62dfddfd-43a0-4395-ae1c-28bd72da56c1.parquet /sqoop/ingest/productos_mas_de_20.parquet
hadoop@615cf53bef6c:~$ hdfs dfs -ls /sqoop/ingest
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2025-09-23 23:13 /sqoop/ingest/.metadata
drwxr-xr-x   - hadoop supergroup          0 2025-09-23 23:14 /sqoop/ingest/.signals
-rw-r--r--   1 hadoop supergroup       4974 2025-09-23 23:14 /sqoop/ingest/productos_mas_de_20.parquet
hadoop@615cf53bef6c:~$
```

```
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=276233
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=14444
                HDFS: Number of bytes written=8455
                HDFS: Number of read operations=50
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=10
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7856
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=7856
                Total vcore-milliseconds taken by all map tasks=7856
                Total megabyte-milliseconds taken by all map tasks=12066816
        Map-Reduce Framework
                Map input records=48
                Map output records=48
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=70
                CPU time spent (ms)=4010
                Physical memory (bytes) snapshot=273584128
                Virtual memory (bytes) snapshot=3001712640
                Total committed heap usage (bytes)=154140672
                Peak Map Physical memory (bytes)=273584128
                Peak Map Virtual memory (bytes)=3001712640
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2025-09-23 23:14:10,809 INFO mapreduce.ImportJobBase: Transferred 8.2568 KB in 27.6976 seconds (305.2608 bytes/sec)
2025-09-23 23:14:10,817 INFO mapreduce.ImportJobBase: Retrieved 48 records.
hadoop@615cf53bef6c:~$
```

Practica Nifi

Crear el ambiente de nifi e ingestar el archivo starwars.csv, luego convertirlo a starwars.avro y guardarlo en la carpeta de HDFS: /nifi
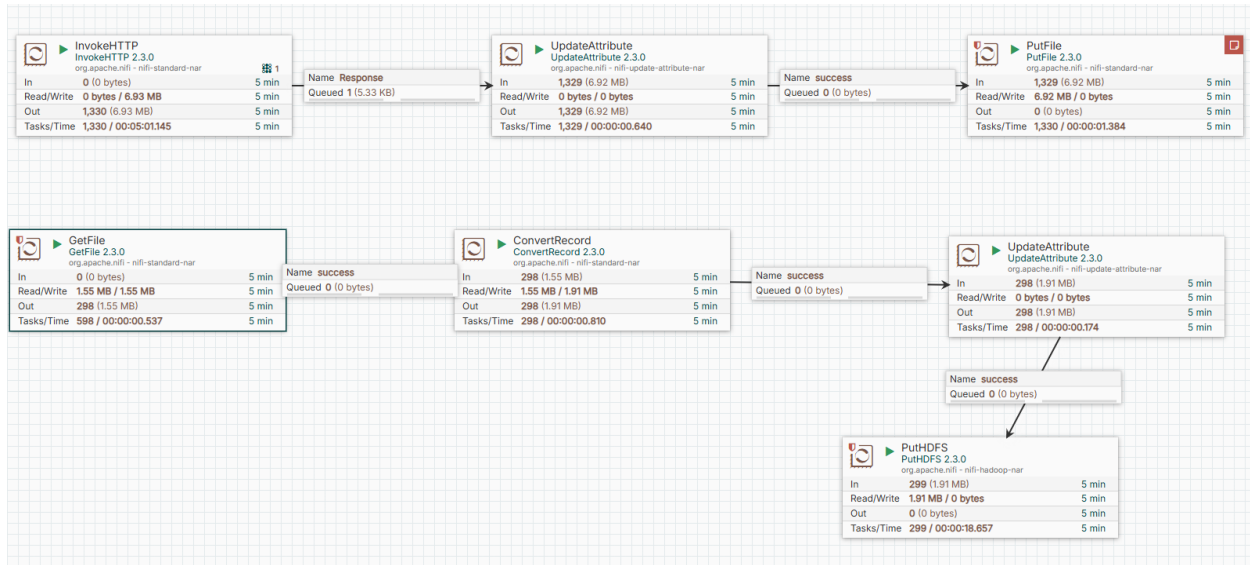
El siguiente screenshot confirma que funciona la primera rama el piplane:


```
nifi@1647add278ce:~/hdfs$ ls
core-site.xml   hdfs-site.xml
nifi@1647add278ce:~/hdfs$ cd ..
nifi@1647add278ce:~$ ls
hdfs   tmp
nifi@1647add278ce:~$ cd tmp
nifi@1647add278ce:~/tmp$ ls
starwars.csv
```

El ultimo screenshot confirma que funciona la segunda rama del piplane:


```
hadoop@615cf53bef6c:~$ hdfs dfs -ls /nifi
Found 1 items
-rw-r--r--   1 nifi supergroup       6706 2025-09-24 20:57 /nifi/starwars.avro
hadoop@615cf53bef6c:~$
```