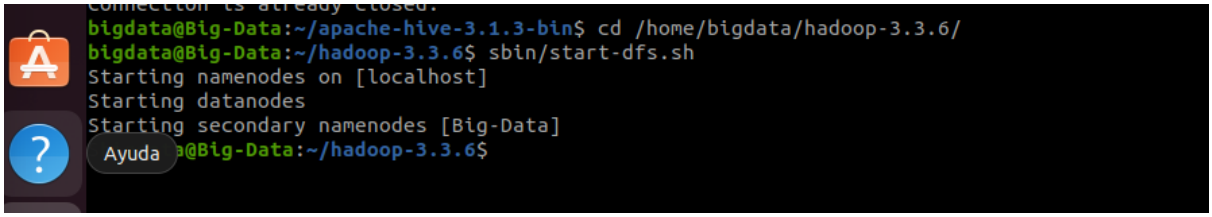
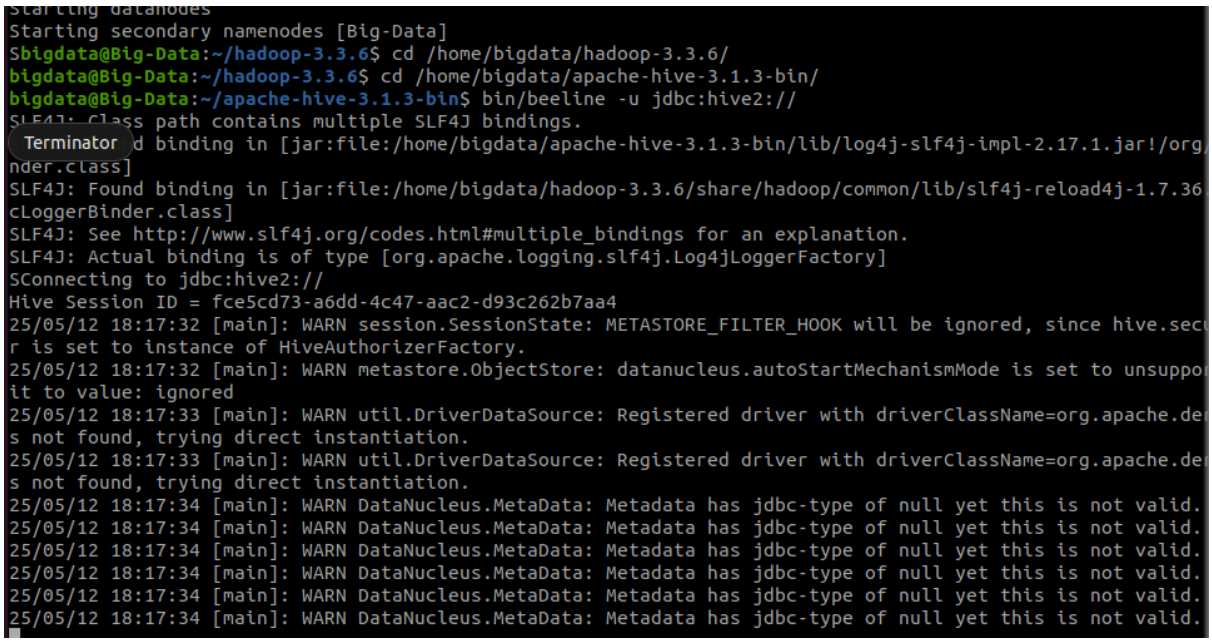
1. Sprint 1 - Pascal Seihon

El primer paso es **Arrancar HDFS**

Parea ello nos ubicamos en el directorio de hadoop y arrancamos el servicio dfs

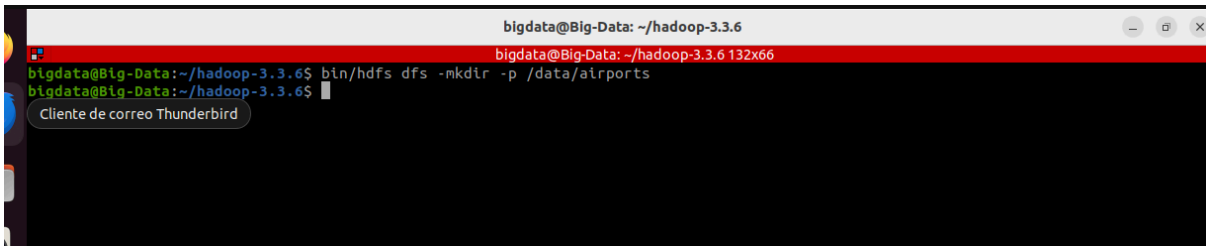


```
Connection is already closed.
bigdata@Big-Data:~/apache-hive-3.1.3-bin$ cd /home/bigdata/hadoop-3.3.6/
bigdata@Big-Data:~/hadoop-3.3.6$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [Big-Data]
a@Big-Data:~/hadoop-3.3.6$
```



```
Starting datanodes
Starting secondary namenodes [Big-Data]
Sbigdata@Big-Data:~/hadoop-3.3.6$ cd /home/bigdata/hadoop-3.3.6/
bigdata@Big-Data:~/hadoop-3.3.6$ cd /home/bigdata/apache-hive-3.1.3-bin/
bigdata@Big-Data:~/apache-hive-3.1.3-bin$ bin/beeline -u jdbc:hive2://
SLF4J: Class path contains multiple SLF4J bindings.
         d binding in [jar:file:/home/bigdata/apache-hive-3.1.3-bin/lib/log4j-slf4j-impl-2.17.1.jar!/org
nder.class]
SLF4J: Found binding in [jar:file:/home/bigdata/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36
cLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
SConnecting to jdbc:hive2://
Hive Session ID = fce5cd73-a6dd-4c47-aac2-d93c262b7aa4
25/05/12 18:17:32 [main]: WARN session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.sec
r is set to instance of HiveAuthorizerFactory.
25/05/12 18:17:32 [main]: WARN metastore.ObjectStore: datanucleus.autoStartMechanismMode is set to unsuppo
it to value: ignored
25/05/12 18:17:33 [main]: WARN util.DriverDataSource: Registered driver with driverClassName=org.apache.de
s not found, trying direct instantiation.
25/05/12 18:17:33 [main]: WARN util.DriverDataSource: Registered driver with driverClassName=org.apache.de
s not found, trying direct instantiation.
25/05/12 18:17:34 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid.
25/05/12 18:17:34 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid.
25/05/12 18:17:34 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid.
25/05/12 18:17:34 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid.
25/05/12 18:17:34 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid.
25/05/12 18:17:34 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid.
```

- ○ Crear la carpeta `/data/airports/` en HDFS ejecutando el comando correspondiente.



```
bigdata@Big-Data:~/hadoop-3.3.6$ bin/hdfs dfs -mkdir -p /data/airports
bigdata@Big-Data:~/hadoop-3.3.6$
```

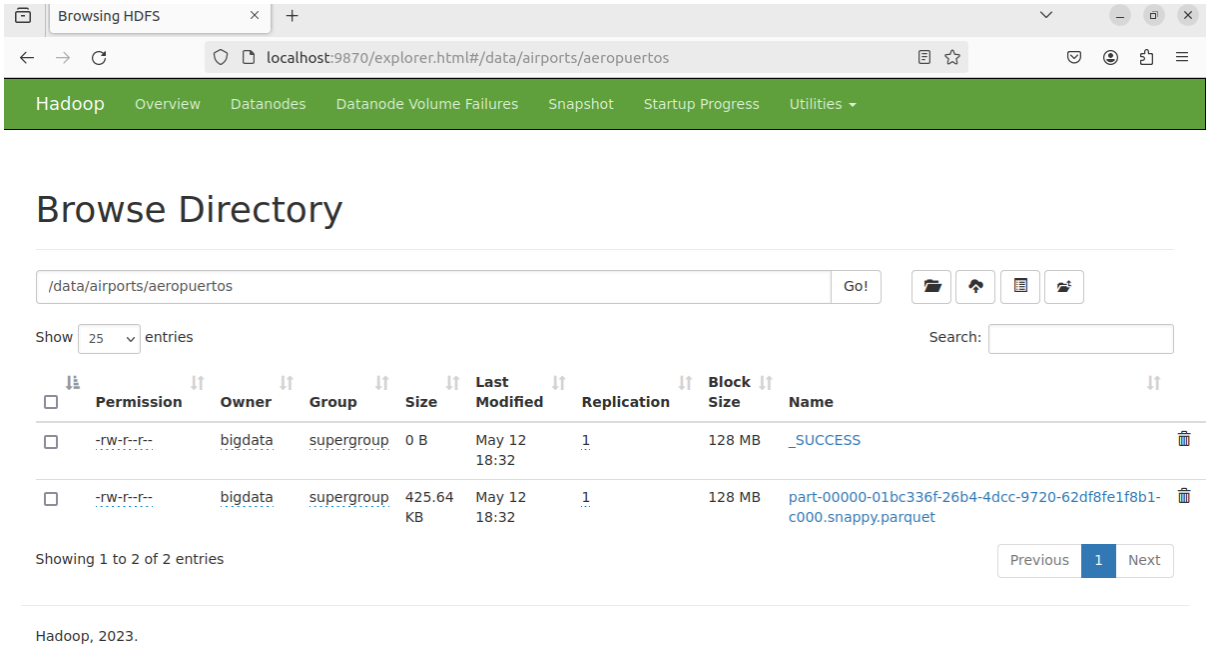
- ○ Mover el fichero parquet contenido en la carpeta `/home/bigdata/Descargas/aeropuertos/` del sistema de ficheros en local de Ubuntu a la carpeta creada anteriormente en el sistema de ficheros distribuído de HDFS.

```
bigdata@Big-Data: ~/Descargas/aeropuertos
bigdata@Big-Data: ~/Descargas/aeropuertos 132x66
bigdata@Big-Data:~/hadoop-3.3.6$ bin/hdfs dfs -mkdir -p /data/airports
bigdata@Big-Data:~/hadoop-3.3.6$ ls /home/bigdata/D
Descargas/  Documentos/
bigdata@Big-Data:~/hadoop-3.3.6$ ls /home/bigdata/Descargas/aeropuertos
part-00000-01bc336f-26b4-4dcc-9720-62df8fe1f8b1-c000.snappy.parquet  _SUCCESS
```

```
bigdata@Big-Data: ~/hadoop-3.3.6
bigdata@Big-Data: ~/hadoop-3.3.6 132x66
bigdata@Big-Data:~/hadoop-3.3.6$ bin/hdfs dfs -copyFromLocal /home/bigdata/Descargas/aeropuertos/
part-00000-01bc336f-26b4-4dcc-9720-62df8fe1f8b1-c000.snappy.parquet
.part-00000-01bc336f-26b4-4dcc-9720-62df8fe1f8b1-c000.snappy.parquet.crc
_SUCCESS
._SUCCESS.crc
bigdata@Big-Data:~/hadoop-3.3.6$ bin/hdfs dfs -copyFromLocal /home/bigdata/Descargas/aeropuertos /data/airports
bigdata@Big-Data:~/hadoop-3.3.6$
```

- Listar el contenido de la carpeta `/data/airports/` en HDFS ejecutando el comando correspondiente.

```
bigdata@Big-Data: ~/hadoop-3.3.6
bigdata@Big-Data: ~/hadoop-3.3.6 132x66
bigdata@Big-Data:~/hadoop-3.3.6$ bin/hdfs dfs -copyFromLocal /home/bigdata/Descargas/aeropuertos/
part-00000-01bc336f-26b4-4dcc-9720-62df8fe1f8b1-c000.snappy.parquet
.part-00000-01bc336f-26b4-4dcc-9720-62df8fe1f8b1-c000.snappy.parquet.crc
_SUCCESS
._SUCCESS.crc
bigdata@Big-Data:~/hadoop-3.3.6$ bin/hdfs dfs -copyFromLocal /home/bigdata/Descargas/aeropuertos /data/airports
bigdata@Big-Data:~/hadoop-3.3.6$ bin/hdfs dfs -ls /data/airports
Found 1 items
drwxr-xr-x   - bigdata supergroup          0 2025-05-12 18:32 /data/airports/aeropuertos
bigdata@Big-Data:~/hadoop-3.3.6$
```



Browse Directory

/data/airports/aeropuertos Go!

Show 25 entries Search:

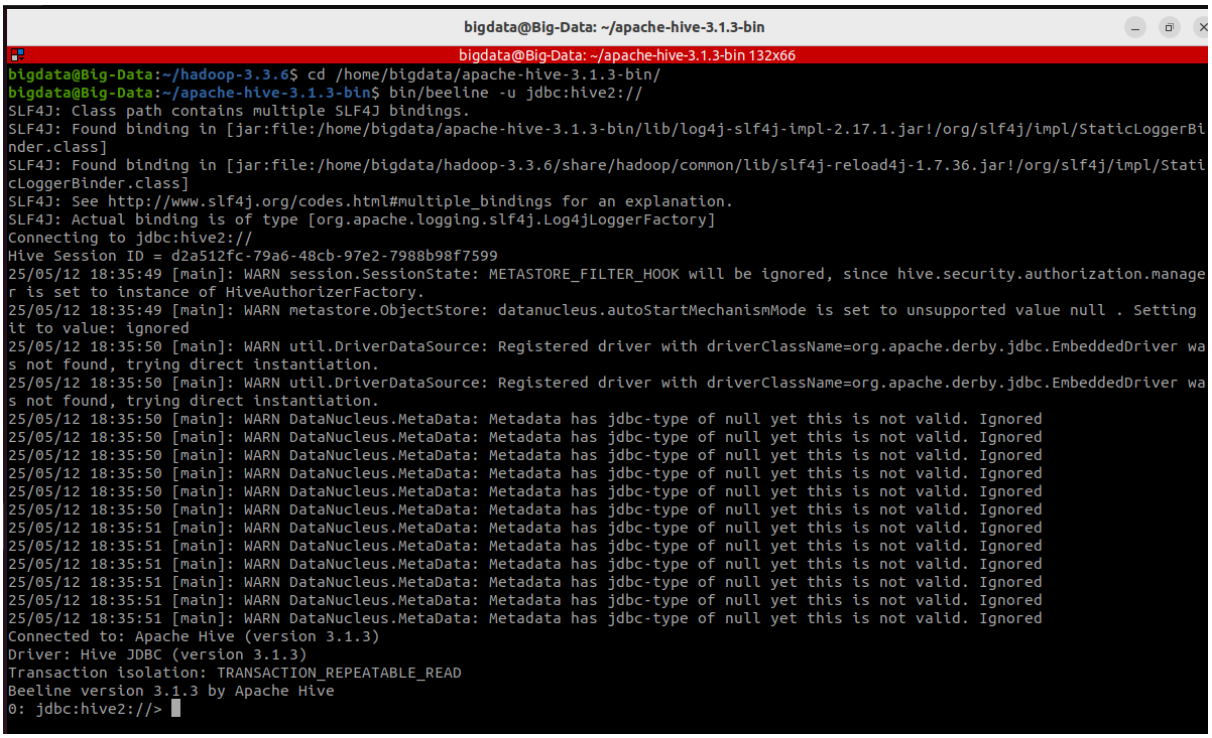
☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	-rw-r--r--	bigdata	supergroup	0 B	May 12 18:32	1	128 MB	_SUCCESS
☐	-rw-r--r--	bigdata	supergroup	425.64 KB	May 12 18:32	1	128 MB	part-00000-01bc336f-26b4-4dcc-9720-62df8fe1f8b1-c000.snappy.parquet

Showing 1 to 2 of 2 entries

Previous 1 Next

Hadoop, 2023.

- Crear una tabla en Hive llamada que apunte a la carpeta creada anteriormente en HDFS y que contiene el fichero parquet. NOTA: La información sobre la estrucutura de la tabla se puede encontrar en: https://web.archive.org/web/20230930101821/https://openflights.org/data.html



```
bigdata@Big-Data:~/hadoop-3.3.6$ cd /home/bigdata/apache-hive-3.1.3-bin/
bigdata@Big-Data:~/apache-hive-3.1.3-bin$ bin/beeline -u jdbc:hive2://
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/bigdata/apache-hive-3.1.3-bin/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBi
nder.class]
SLF4J: Found binding in [jar:file:/home/bigdata/hadoop-3.3.6/share/hadoop/common/lib/slf4j-reload4j-1.7.36.jar!/org/slf4j/impl/Stati
cLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://
Hive Session ID = d2a512fc-79a6-48cb-97e2-7988b98f7599
25/05/12 18:35:49 [main]: WARN session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manage
r is set to instance of HiveAuthorizerFactory.
25/05/12 18:35:49 [main]: WARN metastore.ObjectStore: datanucleus.autoStartMechanismMode is set to unsupported value null . Setting
it to value: ignored
25/05/12 18:35:50 [main]: WARN util.DriverDataSource: Registered driver with driverClassName=org.apache.derby.jdbc.EmbeddedDriver wa
s not found, trying direct instantiation.
25/05/12 18:35:50 [main]: WARN util.DriverDataSource: Registered driver with driverClassName=org.apache.derby.jdbc.EmbeddedDriver wa
s not found, trying direct instantiation.
25/05/12 18:35:50 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:50 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:50 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:50 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:50 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:50 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:51 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:51 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:51 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:51 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:51 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
25/05/12 18:35:51 [main]: WARN DataNucleus.MetaData: Metadata has jdbc-type of null yet this is not valid. Ignored
Connected to: Apache Hive (version 3.1.3)
Driver: Hive JDBC (version 3.1.3)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3 by Apache Hive
0: jdbc:hive2://>
```

DROP TABLE aeropuertos;

```
CREATE EXTERNAL TABLE aeropuertos (
  airport_id INT,
  name STRING,
  city STRING,
  country STRING,
  iata STRING,
  icao STRING,
  latitude DOUBLE,
  longitude DOUBLE,
  altitude INT,
  timezone DOUBLE,
  dst STRING,                    ls
  tz_database_time_zone STRING,
  type STRING,
  source STRING
) STORED AS PARQUET LOCATION '/data/airports/aeropuertos/';
SELECT * FROM aeropuertos LIMIT 10;
```

```
Driver: Hive JDBC (version 3.1.3)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.3 by Apache Hive
0: jdbc:hive2://> CREATE EXTERNAL TABLE aeropuertos (
. . . . . .>   airport_id INT,
. . . . . .>   name STRING,
. . . . . .>   city STRING,
. . . . . .>   country STRING,
. . . . . .>   iata STRING,
. . . . . .>   icao STRING,
. . . . . .>   latitude DOUBLE,
. . . . . .>   longitude DOUBLE,
. . . . . .>   altitude INT,
. . . . . .>   timezone DOUBLE,
. . . . . .>   dst STRING,
. . . . . .>   tz_database_time_zone STRING,
. . . . . .>   type STRING,
. . . . . .>   source STRING
. . . . . .> ) STORED AS PARQUET LOCATION '/data/airports/aeropuertos/';
```

```
0: jdbc:hive2://> SELECT * FROM aeropuertos LIMIT 10;
OK
+-------------------------+-------------------+-------------------+---------------------+-------------------+-----------
--------+----------------------+----------------------+----------------------+----------------------+---------------
---+-----------------------------------+-------------------+---------------------+
| aeropuertos.airport_id  | aeropuertos.name  | aeropuertos.city  | aeropuertos.country  | aeropuertos.iata  | aeropuerto
s.icao  | aeropuertos.latitude  | aeropuertos.longitude  | aeropuertos.altitude  | aeropuertos.timezone  | aeropuertos.ds
t  | aeropuertos.tz_database_time_zone  | aeropuertos.type  | aeropuertos.source  |
+-------------------------+-------------------+-------------------+---------------------+-------------------+-----------
--------+----------------------+----------------------+----------------------+----------------------+---------------
---+-----------------------------------+-------------------+---------------------+
| NULL                    | 24                | JFK               | 3797                | LAX               | 3484
        | NULL                  | 0.0                   | 738                  | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
| NULL                    | 197               | ATL               | 3682                | ORD               | 3830
        | NULL                  | 0.0                   | NULL                 | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
| NULL                    | 135               | LHR               | 507                 | JFK               | 3797
        | NULL                  | 0.0                   | 777                  | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
| NULL                    | 131               | CDG               | 1382                | NRT               | 2359
        | NULL                  | 0.0                   | 772                  | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
| NULL                    | 193               | MAD               | 1229                | BCN               | 1218
        | NULL                  | 0.0                   | 320                  | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
| NULL                    | 124               | FRA               | 340                 | MUC               | 346
        | NULL                  | 0.0                   | 321                  | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
| NULL                    | 203               | SFO               | 3469                | ORD               | 3830
        | NULL                  | 0.0                   | 738                  | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
| NULL                    | 156               | NRT               | 2359                | HND               | 2356
        | NULL                  | 0.0                   | 767                  | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
| NULL                    | 70                | SVO               | 2985                | LED               | 2948
        | NULL                  | 0.0                   | 320                  | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
| NULL                    | 81                | SYD               | 3361                | MEL               | 3339
        | NULL                  | 0.0                   | 738                  | NULL                 | NULL
   | NULL                              | NULL              | NULL                |
+-------------------------+-------------------+-------------------+---------------------+-------------------+-----------
--------+----------------------+----------------------+----------------------+----------------------+---------------
---+-----------------------------------+-------------------+---------------------+
10 rows selected (1,199 seconds)
0: jdbc:hive2://>
```

- Responder a las siguientes preguntas a través de consultas SQL sobre la tabla aeropuertos de Hive:
  - ¿Qué aeropuerto está a mayor altitud (columna altitude)?

    SELECT * FROM aeropuertos ORDER BY altitude DESC LIMIT 1;

```
Ended Job = job_local1144649934_0001
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 7290 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
+-------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+---------
-------------+-----------------+-----------------+-----------------+-----------------+-----------------+
| aeropuertos.airport_id  | aeropuertos.name  | aeropuertos.city  | aeropuertos.country  | aeropuertos.iata  | aeropuertos.icao  | aeropuertos.latitude  | aeropuertos.longitude  | aeropuertos.altitude  | aeropuer
tos.timezone  | aeropuertos.dst  | aeropuertos.tz_database_time_zone  | aeropuertos.type  | aeropuertos.source  |
+-------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+---------
-------------+-----------------+-----------------+-----------------+-----------------+-----------------+
| 3127            | Pokhara Airport   | Pokhara           | Nepal             | PKR               | VNPK              | 28.200899             | 83.982101             | 2712                  | 5.75
          | N               | Asia/Katmandu                     | airport           | OurAirports         |
+-------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+---------
-------------+-----------------+-----------------+-----------------+-----------------+-----------------+
1 row selected (2,038 seconds)
0: jdbc:hive2://> S
```

- ¿Cuántos aeropuertos hay en España (Spain)?

  SELECT COUNT(*) AS total_aeropuertos FROM aeropuertos WHERE country = 'Spain';

```
Job running in-process (local Hadoop)
25/05/12 21:03:43 [pool-21-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
2025-05-12 21:03:43,969 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local776025961_0002
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 13318 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
+--------------------+
| total_aeropuertos  |
+--------------------+
| 2                  |
+--------------------+
1 row selected (1,667 seconds)
0: jdbc:hive2://>
```

- ¿Qué países tienen aeropuertos cuyo horario de verano (columna dst) sea el de Europa (E)?

  SELECT DISTINCT COUNTRY FROM aeropuertos WHERE DST= “ E”;

```
25/05/12 21:04:30 [pool-27-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X rel
2025-05-12 21:04:31,424 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local547141957_0003
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 19346 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
+-----------------+
|     country     |
+-----------------+
| France          |
| Germany         |
| Netherlands     |
| Spain           |
| United Kingdom  |
+-----------------+
5 rows selected (1,46 seconds)
0: jdbc:hive2://>
```

1. Sprint 2 Pascal Seihon

La entrega de este sprint consiste en un informe técnico en formato PDF en el que se aporten capturas de pantalla de los pasos seguidos además de una explicación sobre qué se está haciendo y el por qué de cada paso.
El escenario de este sprint conlleva la realización de las siguientes tareas:

```
bigdata@Big-Data: ~/confluent-7.5.1
bigdata@Big-Data: ~/confluent-7.5.1 132x27
bigdata@Big-Data:~/confluent-7.5.1$ confluent local services connect start
The local commands are intended for a single-node development environment only, NOT for production usage. See more: https://docs.con
fluent.io/current/cli/index.html
As of Confluent Platform 8.0, Java 8 will no longer be supported.

Using CONFLUENT_CURRENT: /tmp/confluent.096938
Starting ZooKeeper
ZooKeeper is [UP]
Starting Kafka
Kafka is [UP]
Starting Schema Registry
Schema Registry is [UP]
Starting Connect
Connect is [UP]
bigdata@Big-Data:~/confluent-7.5.1$
```

- Cargar el conector Hdfs3Sink en Kafka Connect con el fichero /home/bigdata/Descargas/hdfs3-parquet.json

```
bigdata@Big-Data: ~/confluent-7.5.1
bigdata@Big-Data: ~/confluent-7.5.1 132x27
bigdata@Big-Data:~/confluent-7.5.1$ bin/confluent local services connect connector load Hdfs3Sink --config /home/bigdata/Descargas/h
dfs3-parquet.json
The local commands are intended for a single-node development environment only, NOT for production usage. See more: https://docs.con
fluent.io/current/cli/index.html
As of Confluent Platform 8.0, Java 8 will no longer be supported.

{
  "name": "hdfs3-parquet",
  "config": {
    "connector.class": "io.confluent.connect.hdfs3.Hdfs3SinkConnector",
    "tasks.max": "1",
    "topics": "airlines_topic",
    "hdfs.url": "hdfs://localhost:9000",
    "flush.size": "10",
    "key.converter": "org.apache.kafka.connect.storage.StringConverter",
    "value.converter": "io.confluent.connect.avro.AvroConverter",
    "value.converter.schema.registry.url": "http://localhost:8081",
    "confluent.topic.bootstrap.servers": "localhost:9092",
    "confluent.topic.replication.factor": "1",
    "format.class": "io.confluent.connect.hdfs3.parquet.ParquetFormat",
    "name": "hdfs3-parquet"
  },
  "tasks": [],
  "type": "sink"
}
bigdata@Big-Data:~/confluent-7.5.1$
```

- Comprobar el estado de dicho conector con el comando correspondiente desde la shell.

```
bigdata@Big-Data: ~/confluent-7.5.1
bigdata@Big-Data: ~/confluent-7.5.1 132x27
  "config": {
    "connector.class": "io.confluent.connect.hdfs3.Hdfs3SinkConnector",
    "tasks.max": "1",
    "topics": "airlines_topic",
    "hdfs.url": "hdfs://localhost:9000",
    "flush.size": "10",
    "key.converter": "org.apache.kafka.connect.storage.StringConverter",
    "value.converter": "io.confluent.connect.avro.AvroConverter",
    "value.converter.schema.registry.url": "http://localhost:8081",
    "confluent.topic.bootstrap.servers": "localhost:9092",
    "confluent.topic.replication.factor": "1",
    "format.class": "io.confluent.connect.hdfs3.parquet.ParquetFormat",
    "name": "hdfs3-parquet"
  },
  "tasks": [],
  "type": "sink"
}
bigdata@Big-Data:~/confluent-7.5.1$ bin/confluent local services connect connector status Hdfs3Sink
The local commands are intended for a single-node development environment only, NOT for production usage. See more: https://docs.con
fluent.io/current/cli/index.html
As of Confluent Platform 8.0, Java 8 will no longer be supported.

{
  "error_code": 404,
  "message": "No status found for connector Hdfs3Sink"
}
bigdata@Big-Data:~/confluent-7.5.1$ S
```

- Ejecutar un consumidor de Avro del topic `airlines_topic` de Kafka con el comando correspondiente desde la shell.

```
bigdata@Big-Data:~/confluent-7.5.1$ bin/kafka-avro-console-consumer --bootstrap-server localhost:9092 --topic airlines_topic
[2025-05-21 00:41:54,926] INFO KafkaAvroDeserializerConfig values:
        auto.register.schemas = true
        avro.reflection.allow.null = false
        avro.use.logical.type.converters = false
        basic.auth.credentials.source = URL
        basic.auth.user.info = [hidden]
        bearer.auth.cache.expiry.buffer.seconds = 300
        bearer.auth.client.id = null
        bearer.auth.client.secret = null
        bearer.auth.credentials.source = STATIC_TOKEN
        bearer.auth.custom.provider.class = null
        bearer.auth.identity.pool.id = null
        bearer.auth.issuer.endpoint.url = null
        bearer.auth.logical.cluster = null
        bearer.auth.scope = null
        bearer.auth.scope.claim.name = scope
        bearer.auth.sub.claim.name = sub
        bearer.auth.token = [hidden]
        context.name.strategy = class io.confluent.kafka.serializers.context.NullContextNameStrategy
        http.connect.timeout.ms = 60000
        http.read.timeout.ms = 60000
        id.compatibility.strict = true
        key.subject.name.strategy = class io.confluent.kafka.serializers.subject.TopicNameStrategy
        latest.cache.size = 1000
        latest.cache.ttl.sec = -1
```

- Ejecutar un productor de Avro sobre el topic `airlines_topic` de Kafka aportando el esquema de los datos (NOTA: La información sobre la estrucutura de la tabla se puede encontrar en: https://web.archive.org/web/20230930101821/https://openflights.org/data.html) y volcando los datos contenidos en la carpeta `/home/bigdata/Descargas/aerolineas.json` en el sistema de ficheros en local a través del comando correspondiente desde la shell.

cat /home/bigdata/Descargas/aerolineas.json |
bin/kafka-avro-console-producer --broker-list

localhost:9092 --topic airlines_topic --property

value.schema='{

```
  "type": "record",
  "name": "Airline",
  "fields": [
    { "name": "airlineID", "type": "int" },
    { "name": "name", "type": "string" },
    { "name": "alias", "type": "string" },
    { "name": "iata", "type": "string" },
    { "name": "icao", "type": "string" },
    { "name": "callsign", "type": "string" },
    { "name": "country", "type": "string" },
    { "name": "active", "type": "boolean" }
  ]
}'
```

```
cat /home/bigdata/Descargas/aerolineas.json | \
bin/kafka-avro-console-producer \
  --broker-list localhost:9092 \
  --topic airlines_topic \
  --property schema.registry.url=http://localhost:8081 \
  --property
value.schema="{\"type\":\"record\",\"name\":\"Airline\",\"fields\":[{\"name\":\"airlineID\",\"type\":\"int\"},{\"name\":\"name\",\"type\":\"string\"},{\"name\":\"alias\",\"type\":\"string\"},{\"name\":\"iata\",\"type\":\"string\"},{\"name\":\"icao\",\"type\":\"string\"},{\"name\":\"callsign\",\"type\":\"string\"},{\"name\":\"country\",\"type\":\"string\"},{\"name\":\"active\",\"type\":\"boolean\"}]}"
```

```
bigdata@Big-Data: ~/confluent-7.5.1
bigdata@Big-Data: ~/confluent-7.5.1 132x28
        schema.reflection = false
        schema.registry.basic.auth.user.info = [hidden]
        schema.registry.ssl.cipher.suites = null
        schema.registry.ssl.enabled.protocols = [TLSv1.2]
        schema.registry.ssl.endpoint.identification.algorithm = https
        schema.registry.ssl.engine.factory.class = null
        schema.registry.ssl.key.password = null
        schema.registry.ssl.keymanager.algorithm = SunX509
        schema.registry.ssl.keystore.certificate.chain = null
        schema.registry.ssl.keystore.key = null
        schema.registry.ssl.keystore.location = null
        schema.registry.ssl.keystore.password = null
        schema.registry.ssl.keystore.type = JKS
        schema.registry.ssl.protocol = TLSv1.2
        schema.registry.ssl.provider = null
        schema.registry.ssl.secure.random.implementation = null
        schema.registry.ssl.trustmanager.algorithm = PKIX
        schema.registry.ssl.truststore.certificates = null
        schema.registry.ssl.truststore.location = null
        schema.registry.ssl.truststore.password = null
        schema.registry.ssl.truststore.type = JKS
        schema.registry.url = [http://localhost:8081]
        use.latest.version = false
        use.latest.with.metadata = null
        use.schema.id = -1
        value.subject.name.strategy = class io.confluent.kafka.serializers.subject.TopicNameStrategy
 (io.confluent.kafka.serializers.KafkaAvroSerializerConfig:370)
bigdata@Big-Data:~/confluent-7.5.1$ S
```

- Listar el contenido de la carpeta `/topics/airlines_topic/partition=0` en HDFS ejecutando el comando correspondiente.

```
bigdata@Big-Data: ~/hadoop-3.3.6
bigdata@Big-Data: ~/hadoop-3.3.6 80x24
bigdata@Big-Data:~/hadoop-3.3.6$ hdfs dfs -ls /topics/airlines_topic/partition=0
ls: `/topics/airlines_topic/partition=0': No such file or directory
bigdata@Big-Data:~/hadoop-3.3.6$
```

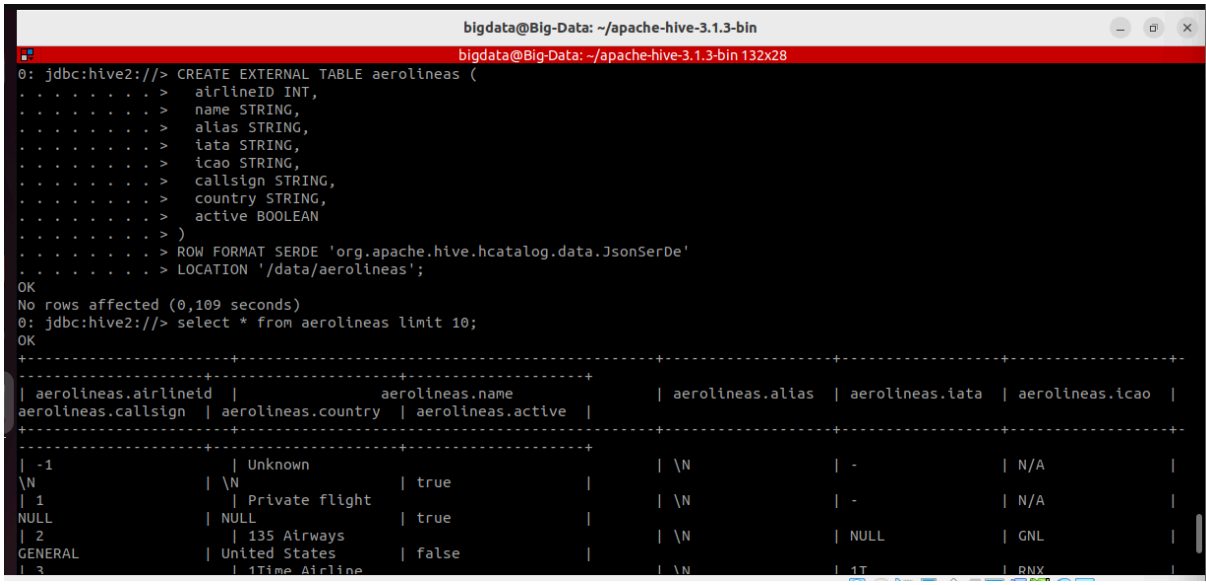
- Crear una tabla en Hive llamada aerolineas que apunte a la carpeta mencionada anteriormente en HDFS y que contiene los ficheros parquet.

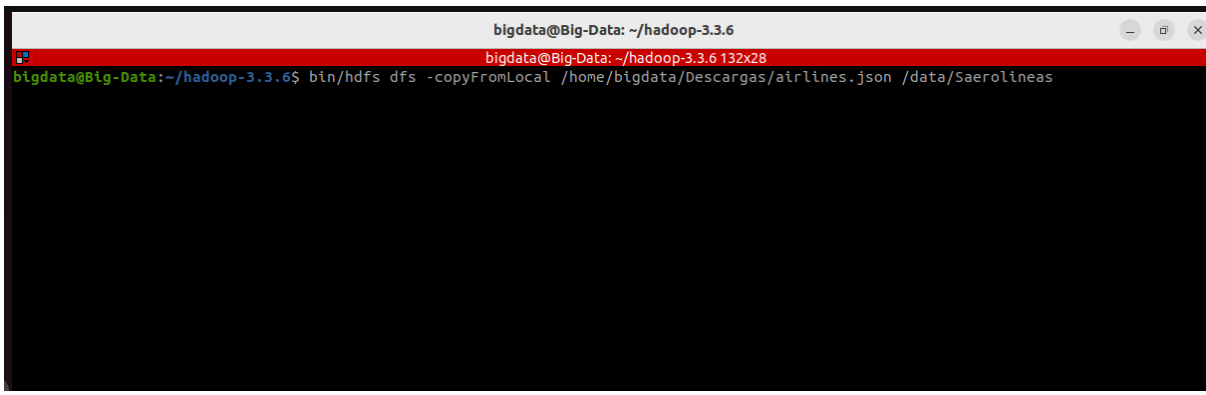
```
DROP TABLE IF EXISTS aerolineas;


CREATE EXTERNAL TABLE aerolineas (

  airlineID INT,
```

```
  name STRING,
  alias STRING,
  iata STRING,
  icao STRING,
  callsign STRING,
  country STRING,
  active BOOLEAN
)
ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe'
LOCATION '/data/aerolineas';
select * from aerolineas limit 10;
```





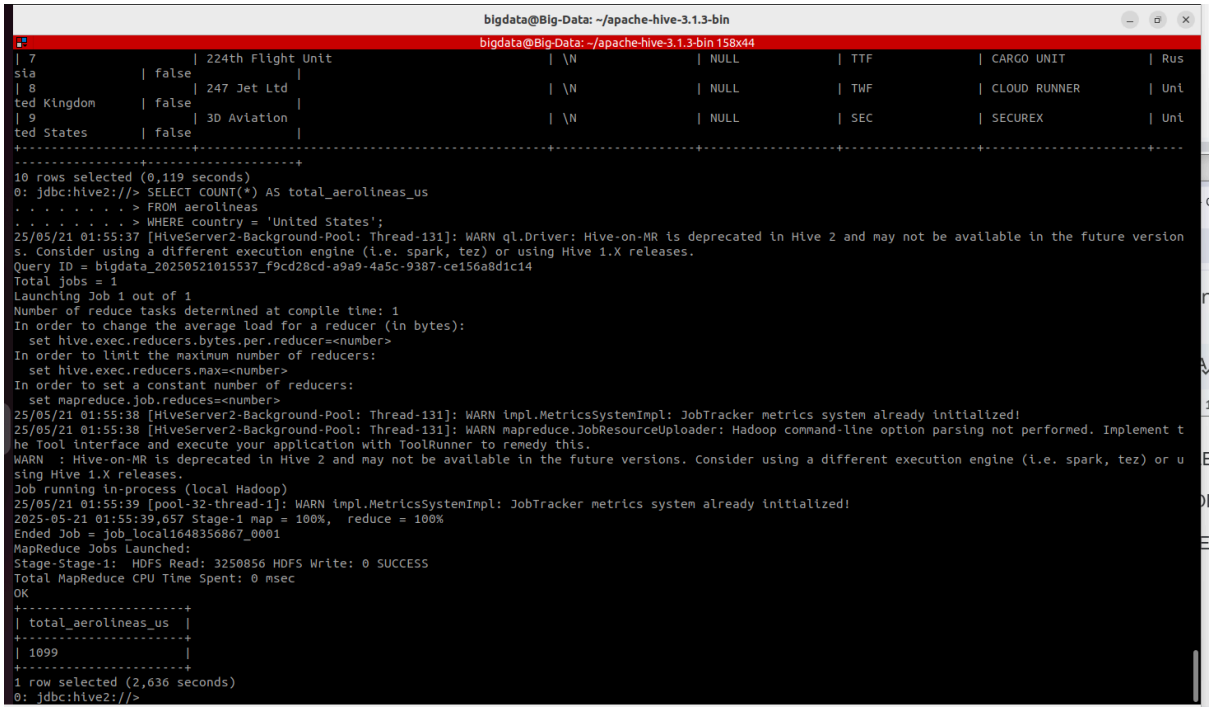
- Responder a las siguientes preguntas a través de consultas SQL sobre la tabla aeropuertos de Hive:

  - ¿Cuántas aerolíneas tiene en total EEUU (United States)?

SELECT COUNT(*) AS total_aerolineas_us

FROM aerolineas

WHERE country = 'United States';



- ¿Cuales son los 10 países con más aerolíneas inactivas (active=false)?

SELECT country, COUNT(*) AS aerolineas_inactivas

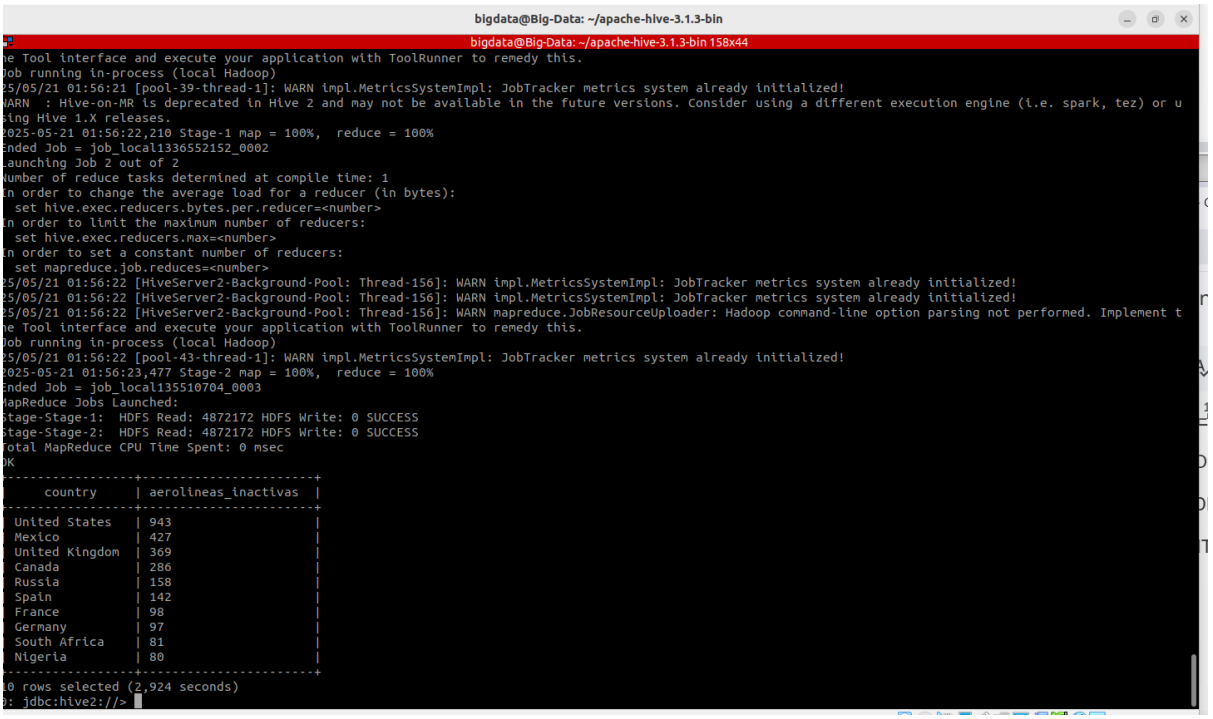
FROM aerolineas

WHERE active = false

GROUP BY country

ORDER BY aerolineas_inactivas DESC

LIMIT 10;



```
bigdata@Big-Data: ~/apache-hive-3.1.3-bin
bigdata@Big-Data: ~/apache-hive-3.1.3-bin 158x44
he Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/05/21 01:56:21 [pool-39-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or u
sing Hive 1.X releases.
2025-05-21 01:56:22,210 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1336552152_0002
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
25/05/21 01:56:22 [HiveServer2-Background-Pool: Thread-156]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/05/21 01:56:22 [HiveServer2-Background-Pool: Thread-156]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
25/05/21 01:56:22 [HiveServer2-Background-Pool: Thread-156]: WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement t
he Tool interface and execute your application with ToolRunner to remedy this.
Job running in-process (local Hadoop)
25/05/21 01:56:22 [pool-43-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-21 01:56:23,477 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local135510704_0003
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 4872172 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 4872172 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
+-----------------+-----------------------+
|     country     | aerolineas_inactivas  |
+-----------------+-----------------------+
| United States   | 943                   |
| Mexico          | 427                   |
| United Kingdom  | 369                   |
| Canada          | 286                   |
| Russia          | 158                   |
| Spain           | 142                   |
| France          | 98                    |
| Germany         | 97                    |
| South Africa    | 81                    |
| Nigeria         | 80                    |
+-----------------+-----------------------+
10 rows selected (2,924 seconds)
0: jdbc:hive2://>
```

- ¿Qué países tienen aerolíneas en activo (active=true) y aeropuertos con una latitud (latitude) mayor a 80?

SELECT DISTINCT a.country

FROM aerolineas a

INNER JOIN aeropuertos a3p ON a.country = aep.country

WHERE a.active = true AND aep.latitude > 80;

```
bigdata@Big-Data: ~/apache-hive-3.1.3-bin
bigdata@Big-Data: ~/apache-hive-3.1.3-bin 105x20
cs system already initialized!
25/05/21 23:41:50 [HiveServer2-Background-Pool: Thread-52]: WARN mapreduce.JobResourceUploader: Hadoop co
mmand-line option parsing not performed. Implement the Tool interface and execute your application with T
oolRunner to remedy this.
Job running in-process (local Hadoop)
25/05/21 23:41:51 [pool-12-thread-1]: WARN impl.MetricsSystemImpl: JobTracker metrics system already init
ialized!
2025-05-21 23:41:51,724 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1111760078_0001
MapReduce Jobs Launched:
Stage-Stage-2:  HDFS Read: 819200 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
+------------+
| a.country  |
+------------+
+------------+
No rows selected (9,453 seconds)
0: jdbc:hive2://>
0: jdbc:hive2://> S
```

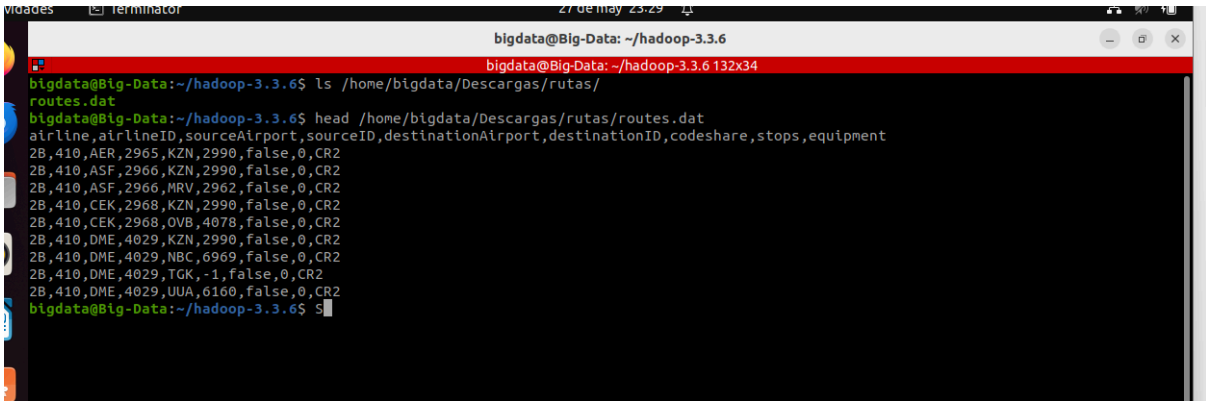
3. Sprint 3 Pascal

- Realizar en NiFi un Process Group para cargar el fichero csv routes.dat (NOTA: La información sobre la estrucutura de la tabla se puede encontrar en:

  https://web.archive.org/web/20230930101821/https://openflights.org/data.html) almacenado en carpeta `/home/bigdata/Descargas/rutas/` (existe un fichero de backup /home/bigdata/Descargas/routes.dat) del sistema de ficheros en local, transformarlo a formato parquet y cargarlo en la ruta `/data/routes/` de HDFS.

Inicialización de HDFS

```
bigdata@Big-Data: ~/hadoop-3.3.6
bigdata@Big-Data: ~/hadoop-3.3.6 132x47
bigdata@Big-Data:~$ cd /home/bigdata/hadoop-3.3.6/
bigdata@Big-Data:~/hadoop-3.3.6$ sbin/start-dfs.sh
Starting namenodes on [localhost]
SStarting datanodes
```



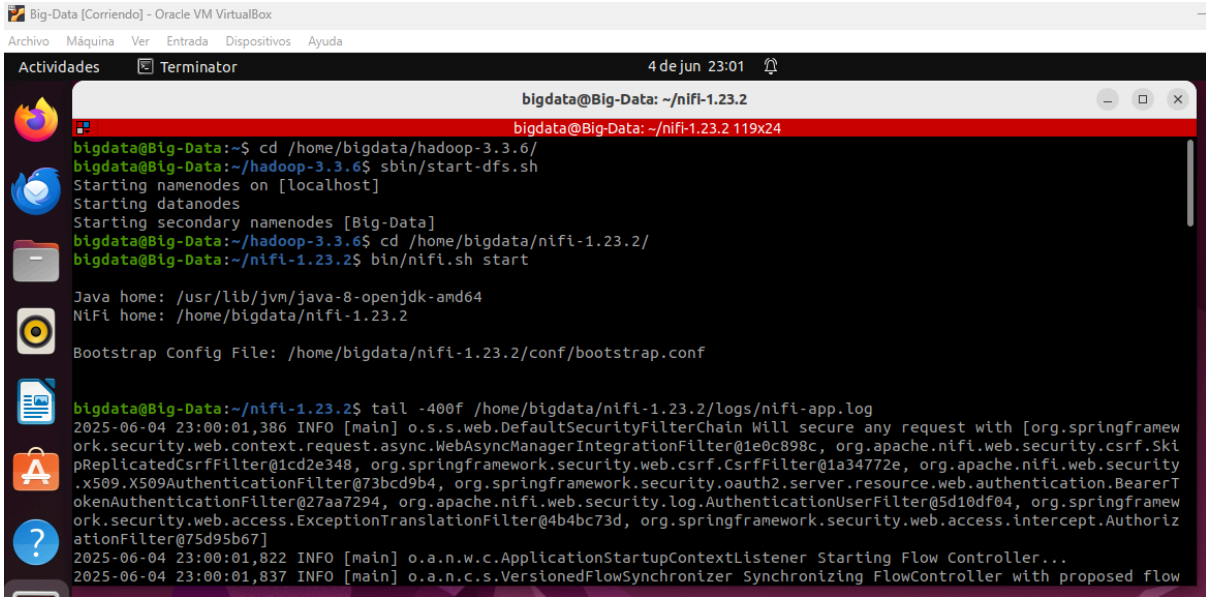
Converti CSV a Parquet

Vamos a arrancar el servicio de NiFi

cd /home/bigdata/nifi-1.23.2/

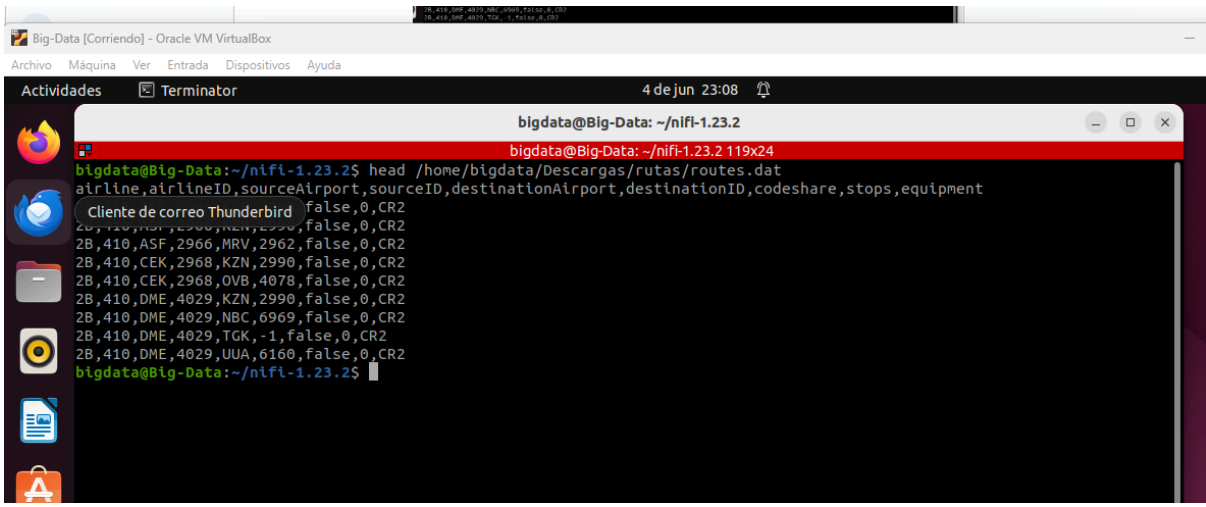
bin/nifi.sh start

para observar el log: tail -400f /home/bigdata/nifi-1.23.2/logs/nifi-app.log

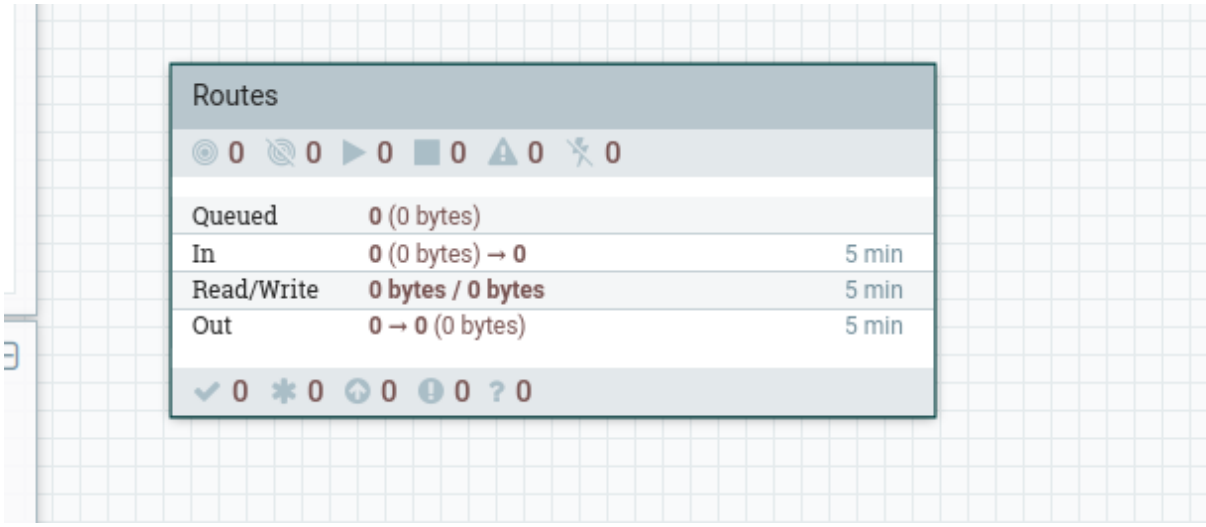


Vamos a observar las primeras lineas del fichero .dat(csv) que vamos a convertir a Parquet

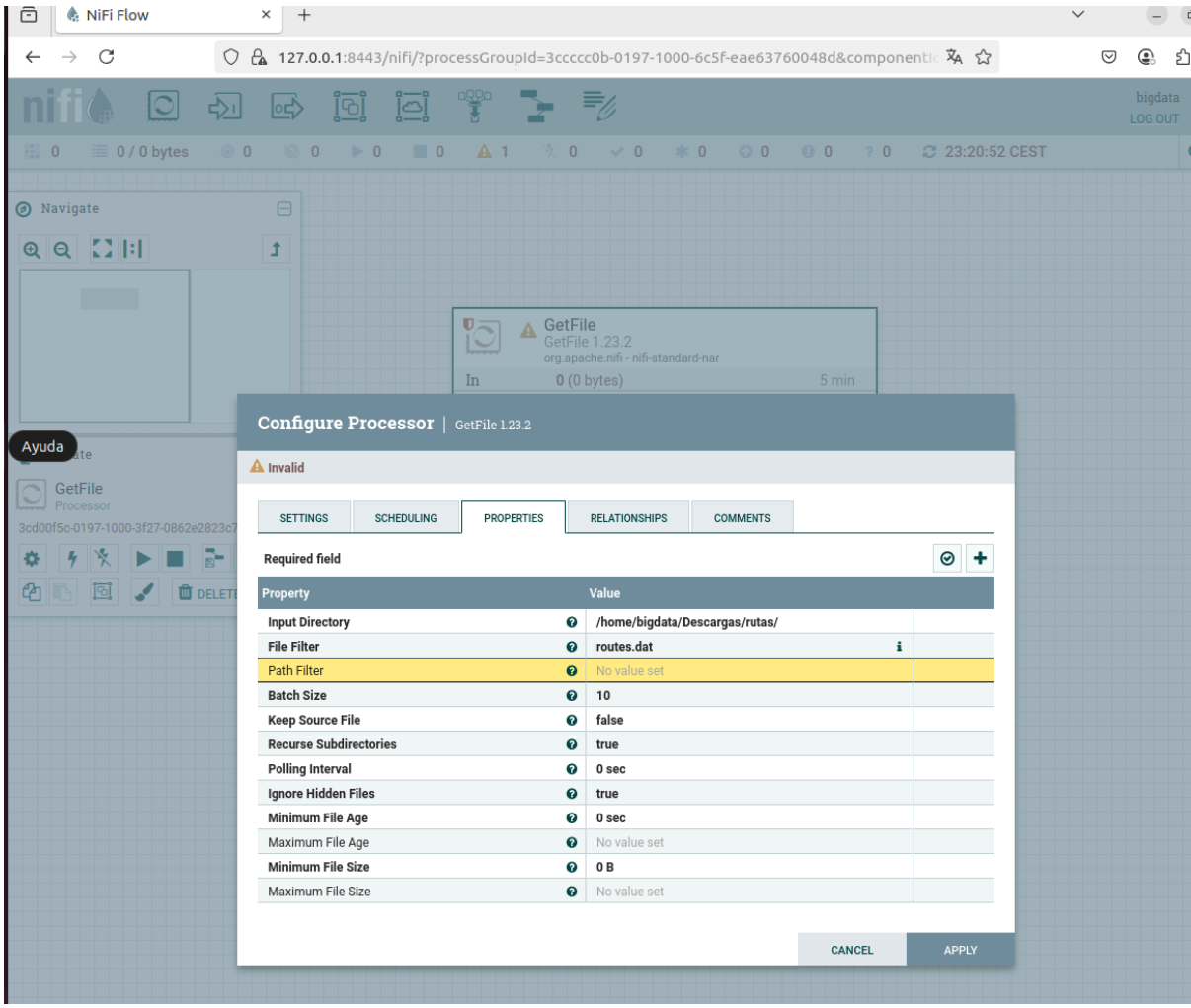
head /home/Descargas/turas/routes.dat



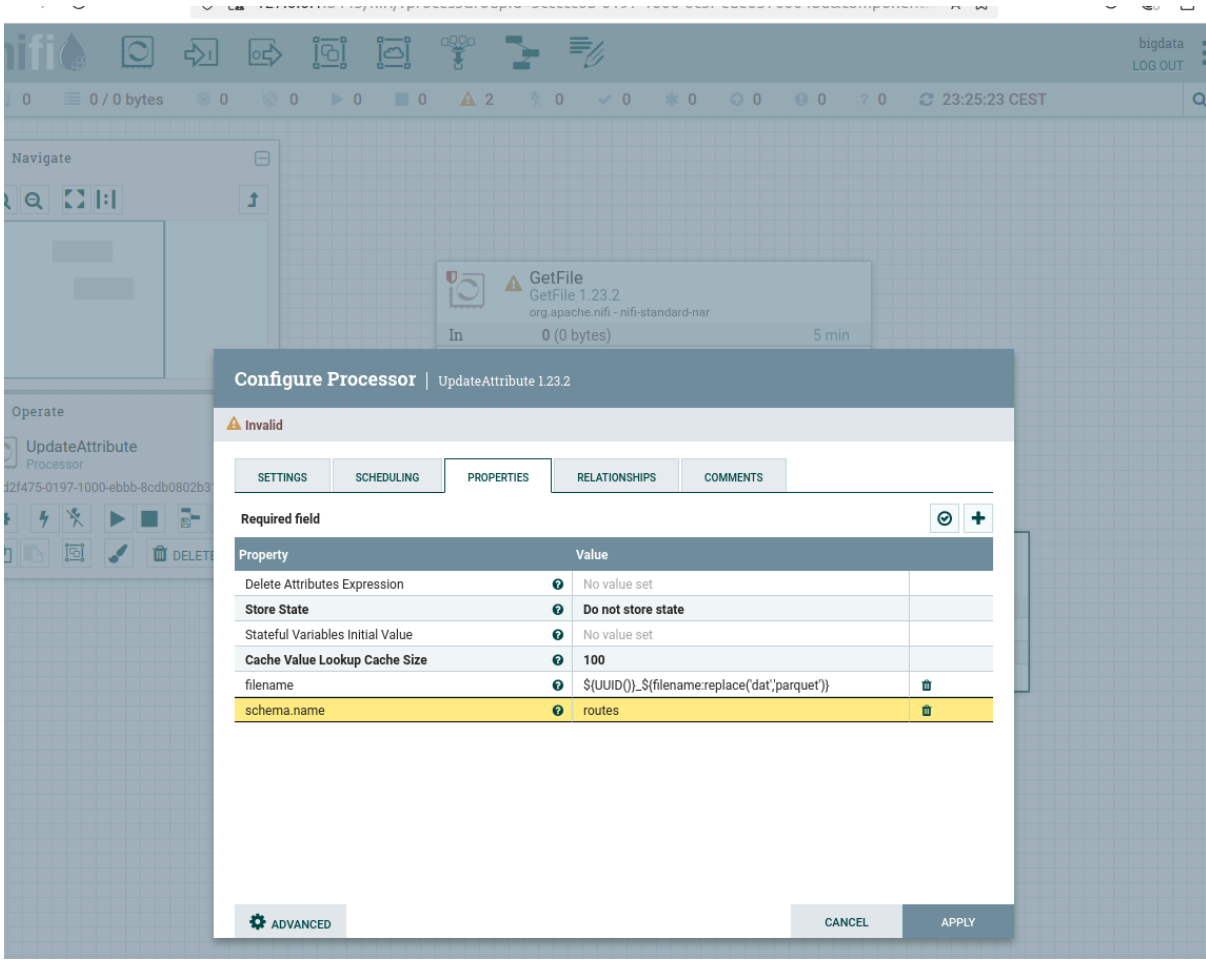
. Creación del Processor Group en NiFi

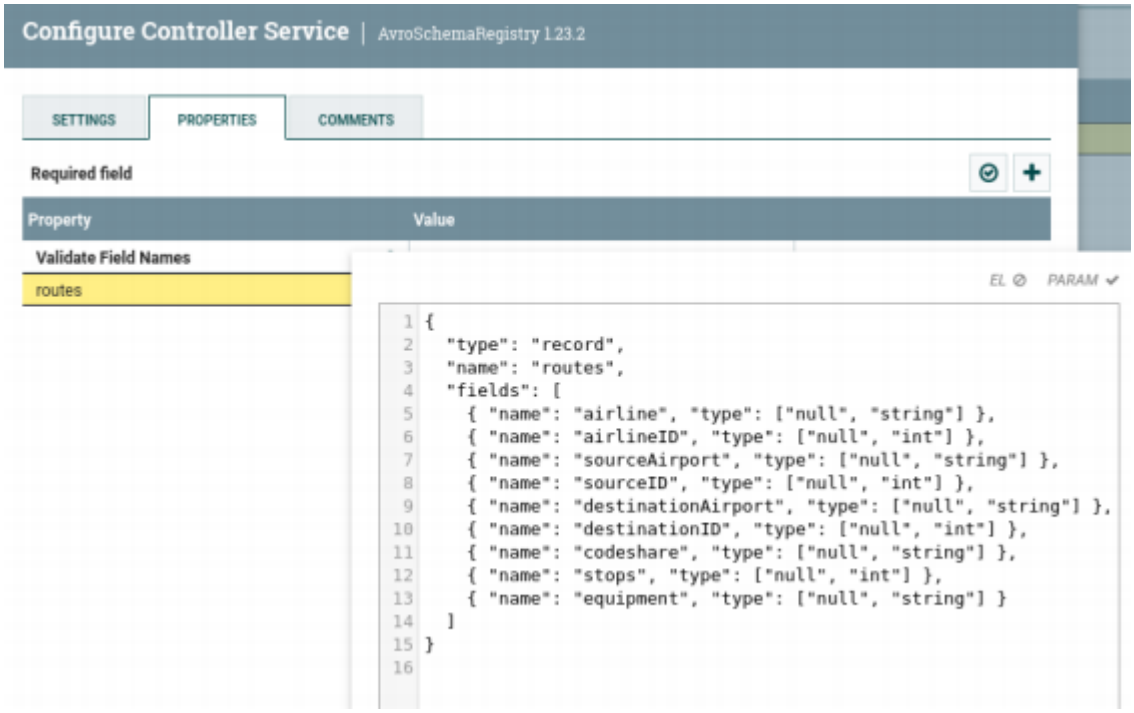


Después configyramos nuestro GetFile

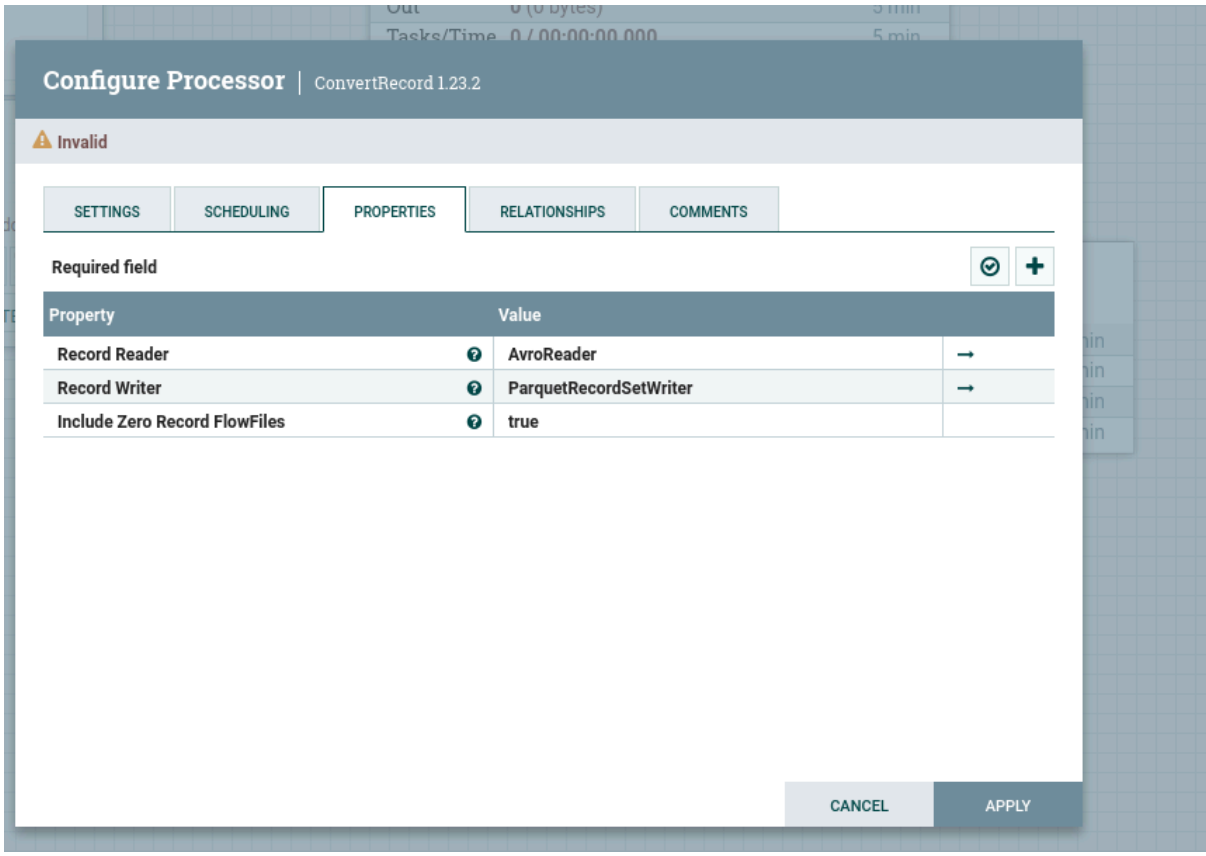


Configuracion **Update Atributos**





Creando el Proceso Convert Record



- Crear una tabla en Hive llamada rutas que apunte a la carpeta mencionada anteriormente en HDFS y que contiene los ficheros parquet.

Objetivo de la actividad: Introducirse en el manejo de la interfaz de NiFi y en el aprendizaje del diseño de flujos de datos tanto Batch como Streaming.