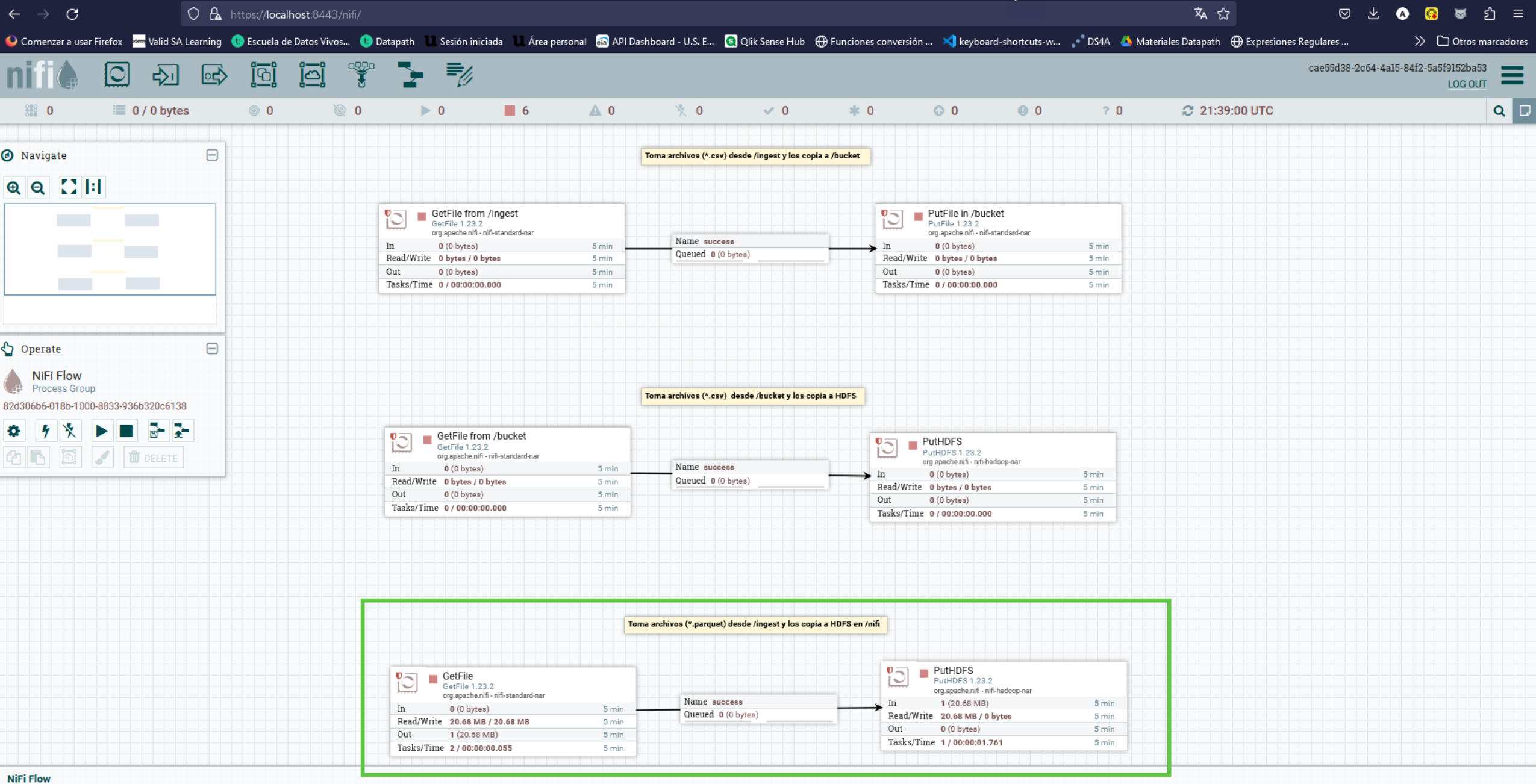
https://localhost:8443/nifi/
Comenzar a usar Firefox
Valid SA Learning
Escuela de Datos Vivos...
Datapath
Sesión iniciada
Área personal
API Dashboard - U.S. E...
Qlik Sense Hub
Funciones conversión ...
keyboard-shortcuts-w...
DS4A
Materiales Datapath
Expresiones Regulares ...
Otros marcadores
nifi
cae55d38-2c64-4a15-84f2-5a5f9152ba53
LOG OUT
0
0 / 0 bytes
0
0
0
6
0
0
0
0
0
0
21:39:00 UTC
Navigate
Operate
NiFi Flow
Process Group
82d306b6-018b-1000-8833-936b320c6138
DELETE
Toma archivos (*.csv) desde /ingest y los copia a /bucket
GetFile from /ingest
GetFile 1.23.2
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
PutFile in /bucket
PutFile 1.23.2
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Toma archivos (*.csv) desde /bucket y los copia a HDFS
GetFile from /bucket
GetFile 1.23.2
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
PutHDFS
PutHDFS 1.23.2
org.apache.nifi - nifi-hadoop-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Toma archivos (*.parquet) desde /ingest y los copia a HDFS en /nifi
GetFile
GetFile 1.23.2
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 20.68 MB / 20.68 MB 5 min
Out 1 (20.68 MB) 5 min
Tasks/Time 2 / 00:00:00.055 5 min
Name success
Queued 0 (0 bytes)
PutHDFS
PutHDFS 1.23.2
org.apache.nifi - nifi-hadoop-nar
In 1 (20.68 MB) 5 min
Read/Write 20.68 MB / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 1 / 00:00:01.761 5 min
NiFi Flow

nifi@d26df9108339: ~/ingest | hadoop@13f54e085e1f: /

```
>>> spark.sql("select VendorID, cast(tpep_pickup_datetime as date), total_amount from yellow_tripdata WHERE total_amount < 10").show()
+--------+--------------------+------------+
|VendorID|tpep_pickup_datetime|total_amount|
+--------+--------------------+------------+
|       1|          2020-12-31|         4.3|
|       2|          2020-12-31|         8.3|
|       2|          2020-12-31|        9.96|
|       2|          2020-12-31|         9.3|
|       2|          2020-12-31|         5.8|
|       1|          2020-12-31|         0.0|
|       1|          2020-12-31|         9.3|
|       2|          2020-12-31|         9.8|
|       2|          2020-12-31|         8.8|
|       2|          2020-12-31|        9.96|
|       2|          2020-12-31|         9.3|
|       2|          2020-12-31|         7.8|
|       1|          2020-12-31|        9.55|
|       2|          2020-12-31|         4.8|
|       1|          2020-12-31|         9.8|
|       2|          2020-12-31|         8.8|
|       1|          2020-12-31|         7.8|
|       2|          2020-12-31|        9.36|
|       2|          2020-12-31|         8.3|
|       1|          2020-12-31|         9.3|
+--------+--------------------+------------+
only showing top 20 rows

>>>
```

nifi@d26df9108339: ~/ingest | hadoop@13f54e085e1f: /

```
>>> spark.sql("select cast(tpep_pickup_datetime as date), SUM(total_amount) AS recaudo from yellow_tripdata GROUP BY cast(tpep_pickup_datetime as date) ORDER BY SUM(total_amount) DESC LIMIT 10").show()
+--------------------+-----------------+
|tpep_pickup_datetime|          recaudo|
+--------------------+-----------------+
|          2021-01-28|961322.5600002451|
|          2021-01-22|942205.9300002148|
|          2021-01-29|937373.5100002222|
|          2021-01-21|932444.4500002082|
|          2021-01-15|931628.1900002063|
|          2021-01-14|926664.0400001821|
|          2021-01-27|  895259.87000017|
|          2021-01-19|890581.4500001629|
|          2021-01-07|887670.1600001527|
|          2021-01-08| 878002.730000146|
+--------------------+-----------------+

>>>
```

nifi@d26df9108339: ~/ingest | hadoop@13f54e085e1f: /

```
>>> spark.sql("select trip_distance, total_amount AS recaudo from yellow_tripdata WHERE trip_distance > 10 ORDER BY total_amount ASC LIMIT 10").show()
+-------------+-------+
|trip_distance|recaudo|
+-------------+-------+
|        12.68| -252.3|
|        34.35|-176.42|
|        14.75| -152.8|
|        33.96|-127.92|
|         29.1| -119.3|
|        26.94| -111.3|
|        20.08| -107.8|
|        19.55| -102.8|
|        19.16| -90.55|
|        25.83| -88.54|
+-------------+-------+

>>>
```

nifi@d26df9108339: ~/ingest | hadoop@13f54e085e1f: /

```
>>> spark.sql("SELECT trip_distance, cast(tpep_pickup_datetime as date) FROM yellow_tripdata WHERE passenger_count >= 2 AND payment_type = 1").show()
+-------------+--------------------+
|trip_distance|tpep_pickup_datetime|
+-------------+--------------------+
|          2.7|          2020-12-31|
|         6.11|          2020-12-31|
|         1.21|          2020-12-31|
|          1.7|          2020-12-31|
|         1.16|          2020-12-31|
|         3.15|          2020-12-31|
|         0.64|          2020-12-31|
|        10.74|          2020-12-31|
|         2.01|          2020-12-31|
|         3.45|          2020-12-31|
|         2.85|          2020-12-31|
|         1.68|          2020-12-31|
|         0.77|          2020-12-31|
|         0.52|          2020-12-31|
|          0.4|          2020-12-31|
|         1.05|          2020-12-31|
|         5.85|          2020-12-31|
|          3.7|          2020-12-31|
|        16.54|          2020-12-31|
|          4.0|          2020-12-31|
+-------------+--------------------+
only showing top 20 rows

>>>
```

nifi@d26df9108339: ~/ingest | hadoop@13f54e085e1f: /

```
>>> spark.sql("SELECT cast(tpep_pickup_datetime as date), trip_distance, passenger_count, tip_amount FROM yellow_tripdata WHERE trip_distance > 10 ORDER BY tip_amount DESC LIMIT 7").show()
+--------------------+-------------+---------------+----------+
|tpep_pickup_datetime|trip_distance|passenger_count|tip_amount|
+--------------------+-------------+---------------+----------+
|          2021-01-20|        427.7|            1.0|   1140.44|
|          2021-01-03|        267.7|            1.0|     369.4|
|          2021-01-12|        326.1|            0.0|    192.61|
|          2021-01-19|        260.5|            1.0|    149.03|
|          2021-01-31|         11.1|            0.0|     100.0|
|          2021-01-01|        14.86|            2.0|      99.0|
|          2021-01-18|         13.0|            0.0|      90.0|
+--------------------+-------------+---------------+----------+

>>>
```

nifi@d26df9108339: ~/ingest | hadoop@13f54e085e1f: /

```
>>> spark.sql("SELECT RatecodeID, SUM(total_amount), AVG(total_amount) FROM yellow_tripdata WHERE RateCodeID <> '6' GROUP BY RatecodeID ").show()
+----------+--------------------+------------------+
|RatecodeID|   sum(total_amount)| avg(total_amount)|
+----------+--------------------+------------------+
|       1.0|1.9496468430212937E7|15.606626116946773|
|       4.0|   90039.93000000082| 74.90842762063296|
|       3.0|   67363.26000000043| 78.69539719626219|
|       2.0|   973635.4700000732| 65.52937609369182|
|      99.0|  1748.0699999999997| 48.55749999999999|
|       5.0|  255075.08999999086|48.939963545662096|
+----------+--------------------+------------------+

>>>
```