



Universidad Nacional Autónoma de México

Facultad de Ingeniería

Sistemas Distribuidos

Profesora: Ing. Guadalupe Lizeth Parrales Romay

Propuesta de Proyecto:

Creación de Cluster Kafka en Databricks

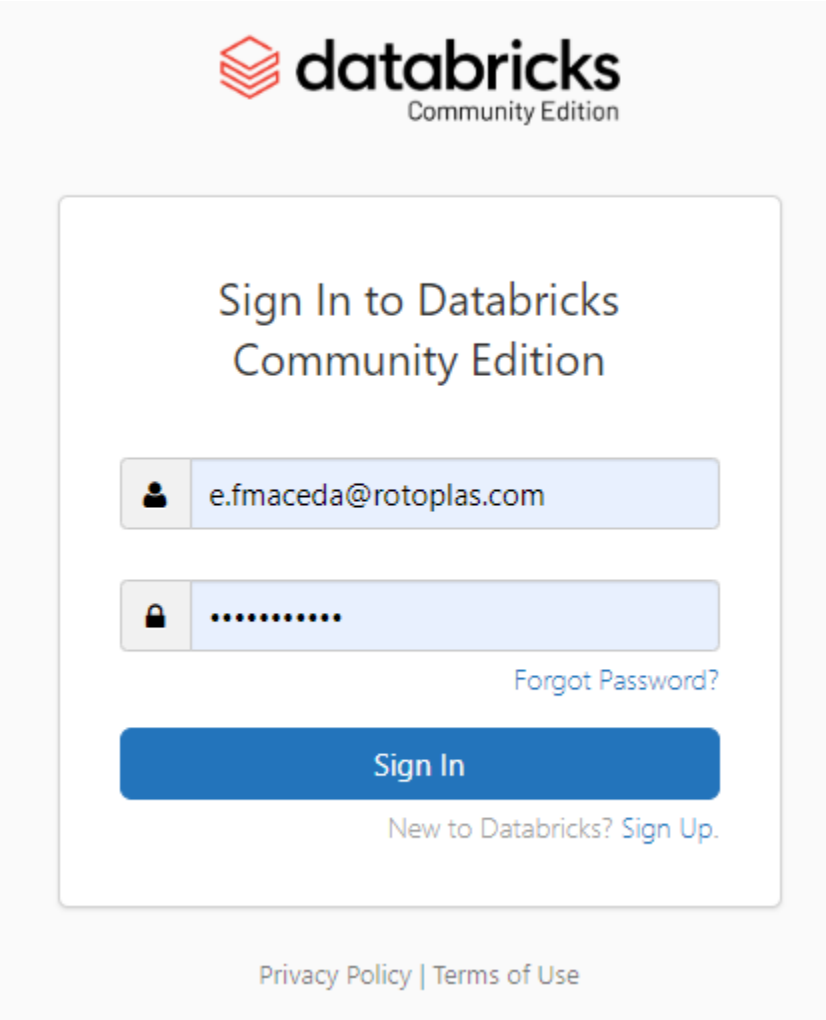
Alumnos: Maceda Patricio Fernando

Calderón Guevara Cesar Yair

Semestre: 2023-1

## 1. Creación de cuenta para Databricks Community

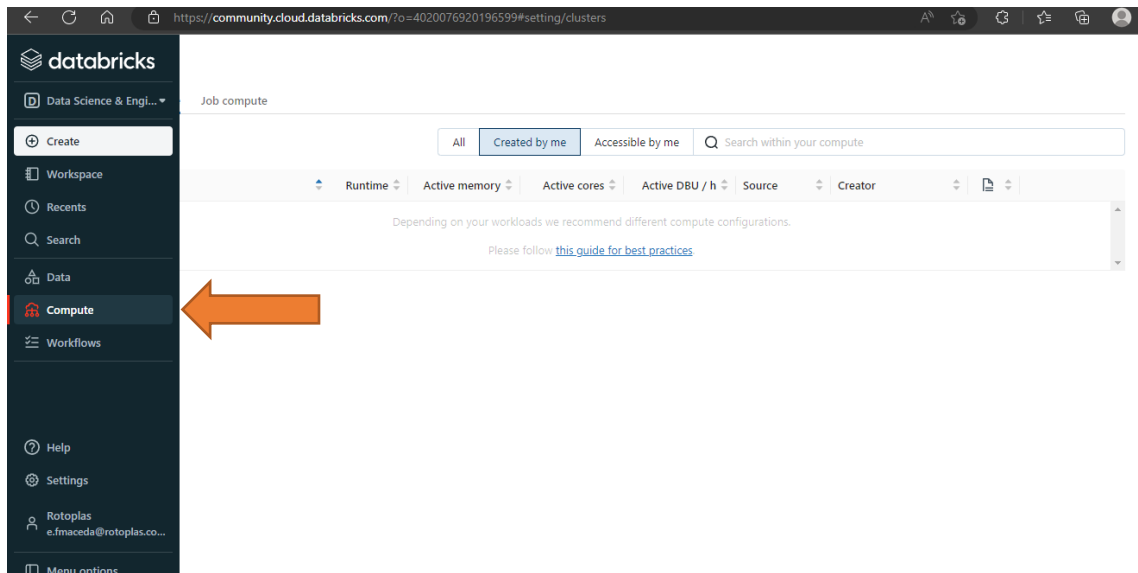
[Login - Databricks Community Edition](#)



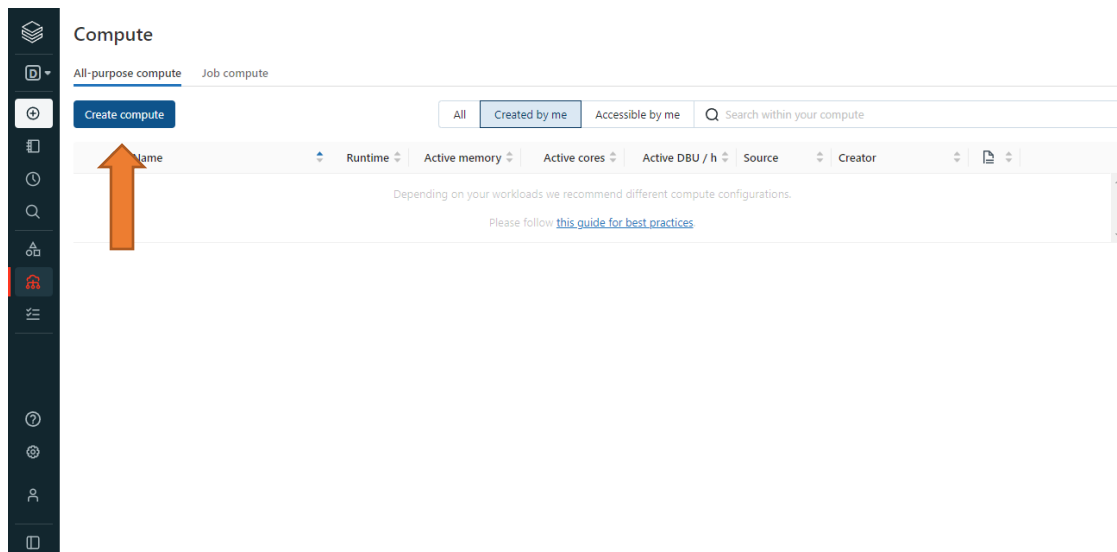
The image shows the login page for Databricks Community Edition. At the top, there is the Databricks logo (a red cube icon) and the text "databricks Community Edition". Below this, the main heading is "Sign In to Databricks Community Edition". There are two input fields: the first is for the email address, containing "e.fmaceda@rotoplas.com", and the second is for the password, represented by a series of dots. To the right of the password field is a link that says "Forgot Password?". Below the input fields is a large blue button labeled "Sign In". Underneath the button is a link that says "New to Databricks? Sign Up.". At the bottom of the page, there are links for "Privacy Policy" and "Terms of Use".

## 2. Creación de Clúster

- a. Dirigirse al apartado de **Compute**



## b. Presionar el botón **Create Clúster**



## c. Dar un nombre al clúster y presionar el botón **Create Clúster**

Clusters / New Compute

New Cluster Cancel Create Cluster 0.0 DBU 1.0 DBU

Cluster name

Databricks runtime version

Instance

Free 15 GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours. For more configuration options, please upgrade your Databricks subscription.

Instances Spark

Availability zone

d. Dirigirse a la pestaña **Settings** y elegir la opción **Admin Console**

databricks

Data Science & Eng...

Job compute

Create

Workspace

Recents

Search

Data

Compute

Workflows

Help

Settings

Rotoplas  
e.fmaceda@rotoplas.co...

Menu options

User Settings

Admin Console

Delete Account

	Runtime	Active memory	Active cores	Active DBU / h	Source	Creator	
ejemplo_kafka	10.4	15 GB	2 cores	1	UI	e.fmaceda@rotoplas.com	

e. Ir la opción de **Workspace Settings**



The screenshot shows the 'Admin Console' interface with the 'Advanced' settings tab selected. A list of features is displayed, each with a toggle switch. An orange arrow points to the 'Web Terminal' toggle, which is currently turned on. The sidebar on the left contains various navigation icons.

Feature	Status
> Third-party iFraming prevention:	Enabled
> MIME type sniffing prevention:	Enabled
> XSS attack page rendering prevention:	Enabled
> Download button for notebook results:	Enabled
> Upload data using the UI:	Enabled
> Notebook Exporting:	Enabled
> Notebook Git Versioning:	Enabled
> Notebook Table Clipboard Features:	Enabled
> Web Terminal:	Enabled
> DBFS File Browser:	Enabled
> Databricks Autologging:	Enabled
> MLflow Run Artifact Download:	Enabled
> MLflow Model Registry Email Notifications:	Enabled

3. Dirigirse a la configuración del Clúster e ir a la opción de **Apps**

Clusters / ejemplo\_kafka

**ejemplo\_kafka** ✓

Configuration Notebooks (0) Libraries Event log Spark UI Driver logs Metrics Apps Spark cluster UI - Master ▼

### Web Terminal

Web terminal provides a Bash terminal running in the driver node. See the [documentation](#) for more details.

Launch Web Terminal

### RStudio Server

RStudio is a registered trademark of RStudio PBC.

To use RStudio Server, you need to install the RStudio Server binary package on the Spark driver. See the [documentation](#) for instructions.

Set up RStudio

#### 4. Abrir la Web Terminal

Clusters / ejemplo\_kafka

**ejemplo\_kafka** ✓

Configuration Notebooks (0) Libraries Event log Spark UI Driver logs Metrics Apps Spark cluster UI - Master ▼

### Web Terminal

Web terminal provides a Bash terminal running in the driver node. See the [documentation](#) for more details.

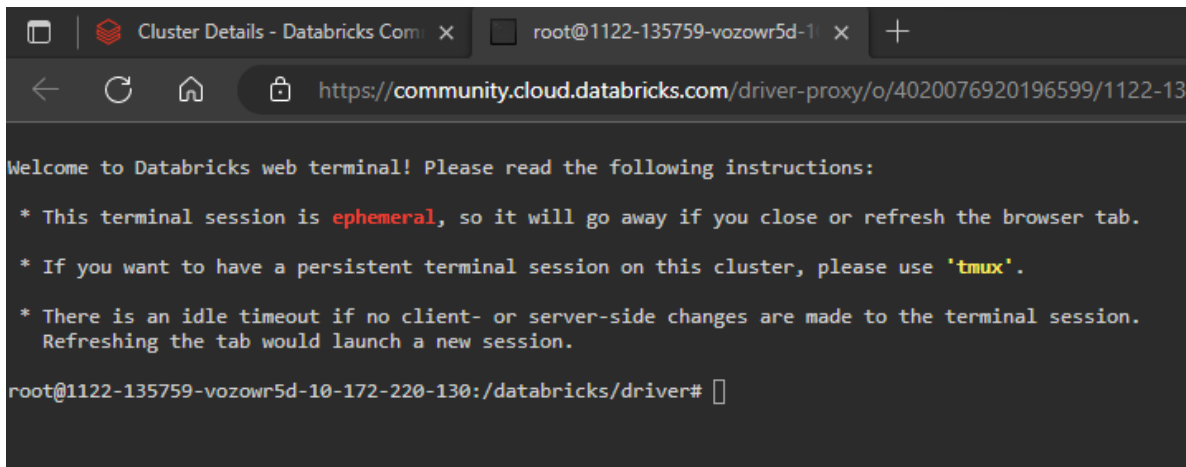
Launch Web Terminal

### RStudio Server

RStudio is a registered trademark of RStudio PBC.

To use RStudio Server, you need to install the RStudio Server binary package on the Spark driver. See the [documentation](#) for instructions.

Set up RStudio



## 5. Ejecutar los siguientes comandos en la terminal

```
wget https://downloads.apache.org/kafka/2.8.2/kafka_2.12-2.8.2.tgz
```

```
tar xzf kafka_2.12-2.8.2.tgz
```

```
cd kafka_2.12-2.8.2
```

```
echo Iniciando Zookeeper ...
```

```
bin/zookeeper-server-start.sh -daemon config/zookeeper.properties > /dev/null 2>&1 & sleep 10
```

```
echo Iniciando Kafka ...
```

```
bin/kafka-server-start.sh -daemon config/server.properties > /dev/null 2>&1 & sleep 10
```

```
bin/kafka-topics.sh --zookeeper localhost:2181 --create --replication-factor 1 --partitions 1 --topic promedios
```

```
bin/kafka-topics.sh --zookeeper localhost:2181 --list
```

```
bin/kafka-console-producer.sh --broker-list localhost:9092 --topic promedios
```



```

root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver# wget https://downloads.apache.org/kafka/2.8.2/kafka_2.12-2.8.2.tgz
--2022-11-22 15:12:12-- https://downloads.apache.org/kafka/2.8.2/kafka_2.12-2.8.2.tgz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 68.99.95.219, 2a01:4f8:10a:201a::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 71748819 (68M) [application/x-gzip]
Saving to: 'kafka_2.12-2.8.2.tgz'

kafka_2.12-2.8.2.tgz          100%[=====] 68.42M  8.89MB/s   in 8.9s

2022-11-22 15:12:21 (7.69 MB/s) - 'kafka_2.12-2.8.2.tgz' saved [71748819/71748819]

root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver# tar xzf kafka_2.12-2.8.2.tgz
root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver# cd kafka_2.12-2.8.2
root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver/kafka_2.12-2.8.2# echo Iniciando Zookeeper ...
Iniciando Zookeeper ...
root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver/kafka_2.12-2.8.2# bin/zookeeper-server-start.sh -daemon config/zookeeper.properties > /dev/null 2>&1 & sleep 10
[1] 3238
[1]+ Done bin/zookeeper-server-start.sh -daemon config/zookeeper.properties > /dev/null 2>&1
root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver/kafka_2.12-2.8.2# echo Iniciando Kafka ...
Iniciando Kafka ...
root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver/kafka_2.12-2.8.2# bin/kafka-server-start.sh -daemon config/server.properties > /dev/null 2>&1 & sleep 10
[1] 3606
[1]+ Done bin/kafka-server-start.sh -daemon config/server.properties > /dev/null 2>&1
root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver/kafka_2.12-2.8.2# bin/kafka-topics.sh --zookeeper localhost:2181 --create --replication-factor 1 --partitions 1 --topic prom
edios
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/databricks/jars/----workspace_spark_3_2--maven-trees--hive-2.3_hadoop-3.2--org.slf4j--slf4j-log4j12--org.slf4j__slf4j-log4j12_1.7.30.jar!/org.slf4j
/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/databricks/driver/kafka_2.12-2.8.2/libs/slf4j-log4j12-1.7.30.jar!/org.slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Created topic promedios.
root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver/kafka_2.12-2.8.2# bin/kafka-topics.sh --zookeeper localhost:2181 --list
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/databricks/jars/----workspace_spark_3_2--maven-trees--hive-2.3_hadoop-3.2--org.slf4j--slf4j-log4j12--org.slf4j__slf4j-log4j12_1.7.30.jar!/org.slf4j
/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/databricks/driver/kafka_2.12-2.8.2/libs/slf4j-log4j12-1.7.30.jar!/org.slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
promedios

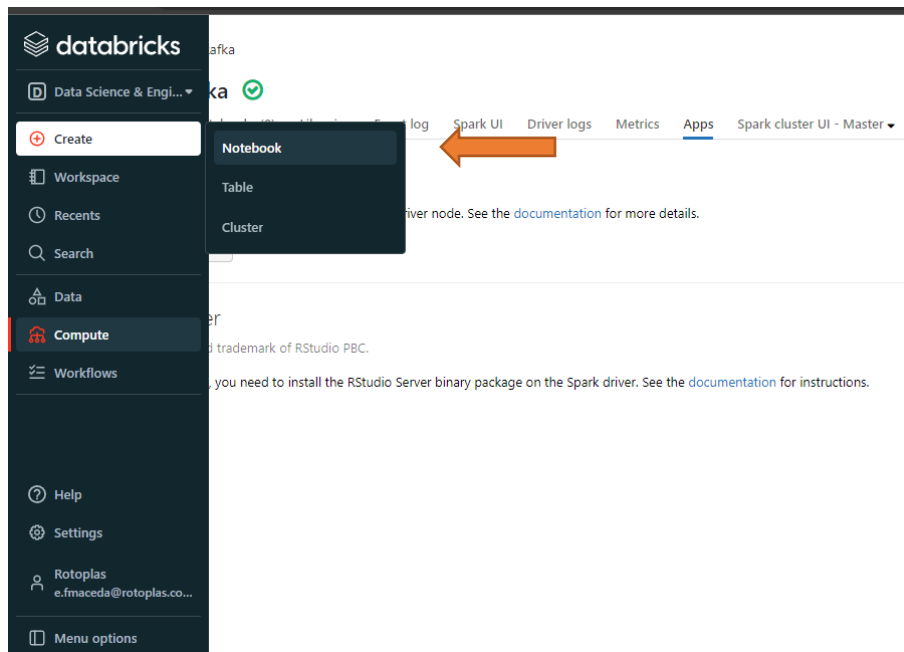
```

```

[2022-11-22 15:14:04,361] WARN An exception was thrown while closing send thread for session 0x100075838830002. (org.apache.zookeeper.ClientCnxn)
EndOfStreamException: Unable to read additional data from server sessionid 0x100075838830002, likely server has closed socket
    at org.apache.zookeeper.ClientCnxnSocketNIO.doIO(ClientCnxnSocketNIO.java:77)
    at org.apache.zookeeper.ClientCnxnSocketNIO.doTransport(ClientCnxnSocketNIO.java:350)
    at org.apache.zookeeper.ClientCnxn$SendThread.run(ClientCnxn.java:1275)
root@1122-135759-vozowr5d-10-172-220-130:/databricks/driver/kafka_2.12-2.8.2# bin/kafka-console-producer.sh --broker-list localhost:9092 --topic promedios
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/databricks/jars/----workspace_spark_3_2--maven-trees--hive-2.3_hadoop-3.2--org.slf4j--slf4j-log4j12--org.slf4j__slf4j-log4j12_1.7.30.jar!/org.slf4j
/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/databricks/driver/kafka_2.12-2.8.2/libs/slf4j-log4j12-1.7.30.jar!/org.slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
>

```

## 6. Crear un nuevo notebook en el apartado **Create**



## 7. Asignarle el Clúster creado al notebook

test01 Python


File Edit View Run Help Last edit was 5 days ago Give feedback

Interrupt ejemplo\_kafka Publish

Cmd 1

### 1. Definición de función

Cmd 2



## 8. Agregar el siguiente código al notebook

test01 Python

File Edit View Run Help Last edit was 5 days ago Give feedback

Interrupt ejemplo\_kafka Publish

Cmd 1

### 1. Definición de función

Cmd 2

```
1 def promediarValores(df):
2     df.createOrReplaceTempView("resultadoMedio")
3     promedios = spark.sql("""SELECT tipo, AVG(total) AS promedio FROM resultadoMedio GROUP BY tipo ORDER BY promedio DESC""")
4     return promedios
```

Command took 0.53 seconds -- by e.fmaceda@rotoplas.com at 22/11/2022, 09:18:49 on ejemplo\_kafka

Cmd 3

### 2. Subscripción al Topic

Cmd 4

```
1 tiposStreamingDF = (spark
2     .readStream
3     .format("kafka")
4     .option("kafka.bootstrap.servers", "127.0.0.1:9092")
5     .option("subscribe", "promedios")
6     .load())
```

Command took 3.59 seconds -- by e.fmaceda@rotoplas.com at 22/11/2022, 09:18:49 on ejemplo\_kafka

Cmd 5

### 3. Definición del esquema de los datos a recibir

Cmd 6

```
1 from pyspark.sql.types import StructType, StructField, StringType, DoubleType
2 import pyspark.sql.functions as F
3
4 esquema = StructType([\
5     StructField("tipo", StringType()),\
6     StructField("total", DoubleType())\
7 ])
8
9 parsedDF = tiposStreamingDF.select("value").withColumn("value", F.col("value").cast(StringType())).withColumn("parejas",
10 F.from_json(F.col("value"), esquema)).withColumn("tipo", F.col("parejas.tipo")).withColumn("total", F.col("parejas.total"))
```

Command took 1.13 seconds -- by e.fmaceda@rotoplas.com at 22/11/2022, 09:18:49 on ejemplo\_kafka

Cmd 7

### 4. Inicialización de stream de datos

Cmd 8

```
1 promediosStreamingDF = promediarValores(parsedDF)
2 salida = promediosStreamingDF\
3     .writeStream\
4     .queryName("AgregacionPromedios")\
5     .outputMode("complete")\
6     .format("memory")\
7     .start()
```

Cancel

▶ (1) Spark Jobs

▶ AgregacionPromedios (jid: 86228607-ba35-4049-bcc8-bf3cb91060e6) Last updated: About now

Cmd 9

```
Cmd 9

5. Mostrar resultados

Cmd 10

1 promediosDF = spark.sql("select * from AgregacionPromedios")
2 promediosDF.show()

+---+-----+
|tipo|promedio|
+---+-----+
+---+-----+

Command complete

Cmd 11
```

9. Ejecutar el paso 4 del notebook

10. Dentro de la Web Terminal ingresar los siguientes datos

```
{"tipo": "gasto", "total": 3.5}
{"tipo": "ingreso", "total": 7.0}
{"tipo": "ingreso", "total": 6.5}
{"tipo": "ingreso", "total": 4.0}
{"tipo": "gasto", "total": 2.5}
```

```
root@1122-204325-uab19ybs-10-172-212-147:/databricks/driver/kafka_2.12-2.8.2# bin/kafka
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/databricks/jars/----workspace_spark_3_2--maven-tree/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/databricks/driver/kafka_2.12-2.8.2/libs/slf4j-log4j12.jar:/]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
>{"tipo": "gasto", "total": 3.5}
>{"tipo": "ingreso", "total": 7.0}
>{"tipo": "ingreso", "total": 6.5}
>{"tipo": "ingreso", "total": 4.0}
>{"tipo": "gasto", "total": 2.5}
>
```

11. Ejecutar el paso 5 del notebook hasta que se muestren los resultados

## 5. Mostrar resultados

Cmd 10

```
1 promediosDF = spark.sql("select * from AgregacionPromedios")
2 promediosDF.show()
```

► (2) Spark Jobs

```
+-----+-----+
| tipo|      promedio|
+-----+-----+
| ingreso|5.833333333333333|
| gasto|          3.0|
+-----+-----+
```

Command took 0.18 seconds -- by e.fmaceda@rotoplas.com at 22/11/2022, 14:56:36 on ejemplo\_kafka