




Tarea 3- Procesamiento de Datos con Apache Spark

Aura Marcela Rodríguez Ruiz

Tutor:
Sandra Milena Patiño Avella

Grupo:
202016911_47

Universidad Nacional Abierta y a Distancia-UNAD
Escuela de Ciencia Básicas, Tecnología e Ingeniería
Big Data
Bogotá, 2024



Problema Escogido: Violencia Intrafamiliar Colombia

Definición:

El conjunto de datos "Domestic Violence in Colombia" en Kaggle contiene información sobre incidentes de violencia doméstica en Colombia desde 2010 hasta 2021. El objetivo es analizar la violencia doméstica en Colombia mediante técnicas de procesamiento batch y en tiempo real utilizando Spark y Kafka. El análisis busca identificar patrones en la frecuencia y tipos de violencia en distintas regiones, ayudando a instituciones a priorizar políticas de intervención.

- Conjunto de datos: El problema que se busca abordar es la prevención y comprensión de la violencia doméstica en Colombia, identificando causas y patrones a lo largo del tiempo.

El conjunto de datos incluye variables como:

- Fecha del incidente
- Tipo de violencia (física, psicológica, sexual, etc.)
- Ubicación del incidente
- Edad y género de las víctimas y agresores
- Relación entre la víctima y el agresor

Este conjunto de datos puede ser utilizado para realizar análisis estadísticos, visualizaciones y modelado predictivo para ayudar en la toma de decisiones y en la implementación de políticas de prevención.

Diseño de la solución, arquitectura y explicación del código

Instalación y Configuración de Kaggle

Instalar Pip y Kaggle

Comando: `sudo apt update`

```

oct 27 16:15
BIGDATA [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Configurando cpp (4:11.2.0-1ubuntu1) ...
Configurando libc6-dev:amd64 (2.35-0ubuntu3.8) ...
Configurando libc-devtools (2.35-0ubuntu3.8) ...
Configurando gcc (4:11.2.0-1ubuntu1) ...
Configurando libexpat1-dev:amd64 (2.4.7-1ubuntu0.4) ...
Configurando libstdc++-11-dev:amd64 (11.4.0-1ubuntu1~22.04) ...
Configurando zlib1g-dev:amd64 (1:1.2.11.dfsg-2ubuntu9.2) ...
Configurando g++-11 (11.4.0-1ubuntu1~22.04) ...
Configurando libpython3.10-dev:amd64 (3.10.12-1~22.04.6) ...
Configurando python3.10-dev (3.10.12-1~22.04.6) ...
Configurando g++ (4:11.2.0-1ubuntu1) ...
update-alternatives: utilizando /usr/bin/g++ para proveer /usr/bin/c++ (c++) en modo automático
Configurando build-essential (12.9ubuntu3) ...
Configurando libpython3-dev:amd64 (3.10.6-1~22.04.1) ...
Configurando python3-dev (3.10.6-1~22.04.1) ...
Procesando disparadores para man-db (2.10.2-1) ...
Procesando disparadores para libc-bin (2.35-0ubuntu3.8) ...
Scanning processes...
Scanning linux images...

Running kernel seems to be up-to-date.

No services need to be restarted.

No containers need to be restarted.

No user sessions are running outdated binaries.

No VM guests are running outdated hypervisor (qemu) binaries on this host.
vboxuser@bigdata:~$ sudo apt install python3-pip
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias... Hecho
Leyendo la información de estado... Hecho
E: No se ha podido localizar el paquete python3-pip
vboxuser@bigdata:~$ pip3 --version
pip 22.0.2 from /usr/lib/python3/dist-packages/pip (python 3.10)
vboxuser@bigdata:~$

```

Comando: pip install kaggle

```

oct 27 16:21
BIGDATA [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Configurando python3.10-dev (3.10.12-1~22.04.6) ...
Configurando g++ (4:11.2.0-1ubuntu1) ...
update-alternatives: utilizando /usr/bin/g++ para proveer /usr/bin/c++ (c++) en modo automático
Configurando build-essential (12.9ubuntu3) ...
Configurando libpython3-dev:amd64 (3.10.6-1~22.04.1) ...
Configurando python3-dev (3.10.6-1~22.04.1) ...
Procesando disparadores para man-db (2.10.2-1) ...
Procesando disparadores para libc-bin (2.35-0ubuntu3.8) ...
Scanning processes...
Scanning linux images...

Running kernel seems to be up-to-date.

No services need to be restarted.

No containers need to be restarted.

No user sessions are running outdated binaries.

No VM guests are running outdated hypervisor (qemu) binaries on this host.
vboxuser@bigdata:~$ sudo apt install python3-pip
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias... Hecho
Leyendo la información de estado... Hecho
E: No se ha podido localizar el paquete python3-pip
vboxuser@bigdata:~$ pip3 --version
pip 22.0.2 from /usr/lib/python3/dist-packages/pip (python 3.10)
vboxuser@bigdata:~$ pip3 install kaggle
defaulting to user installation because normal site-packages is not writeable
Collecting kaggle
  Downloading kaggle-0.0.1-py3-none-any.whl (2.5 kB)
Installing collected packages: kaggle
  WARNING: The script kaggle is installed in '/home/vboxuser/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed kaggle-0.0.1
vboxuser@bigdata:~$

```

Configurar el kaggle.json:

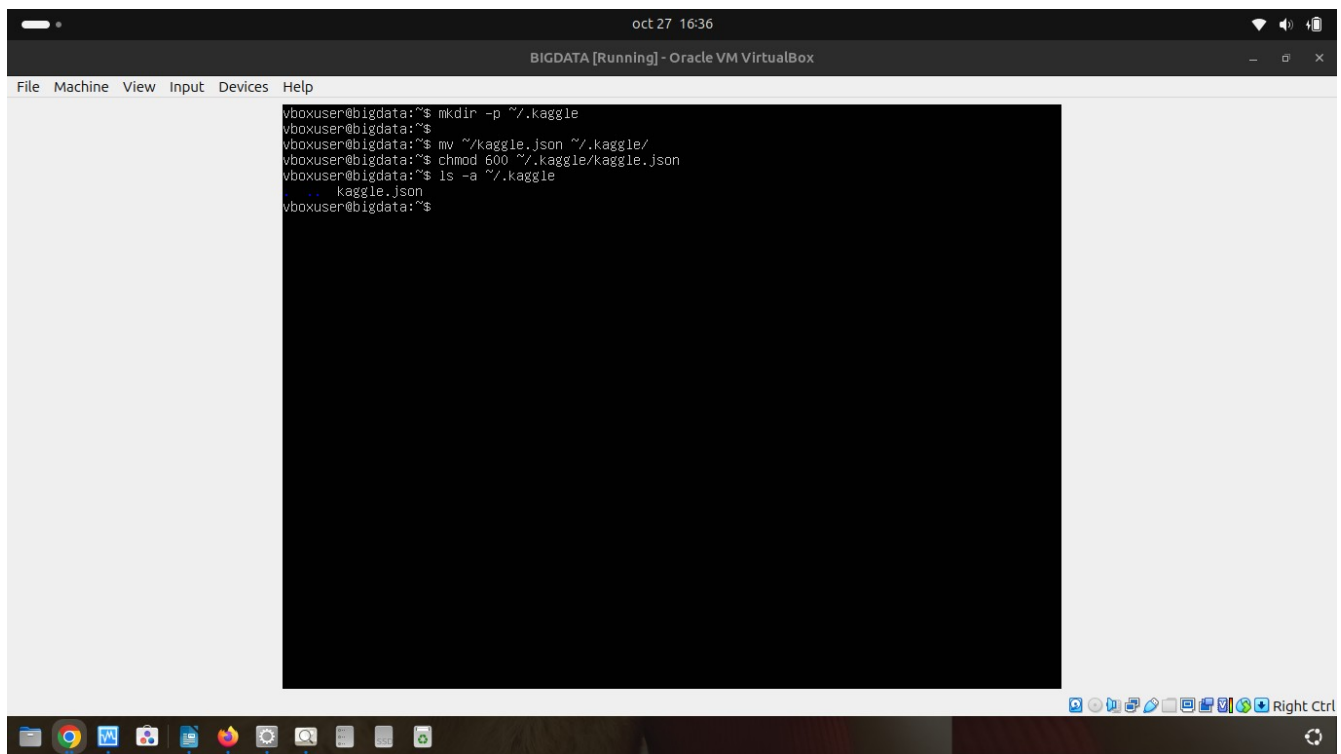
Generar un token de API en Kaggle y descargar el archivo kaggle.json

Configurar el kaggle:

```
mkdir -p ~/.kaggle
```

```
mv /ruta/al/kaggle.json ~/.kaggle/
```

```
chmod 600 ~/.kaggle/kaggle.json
```



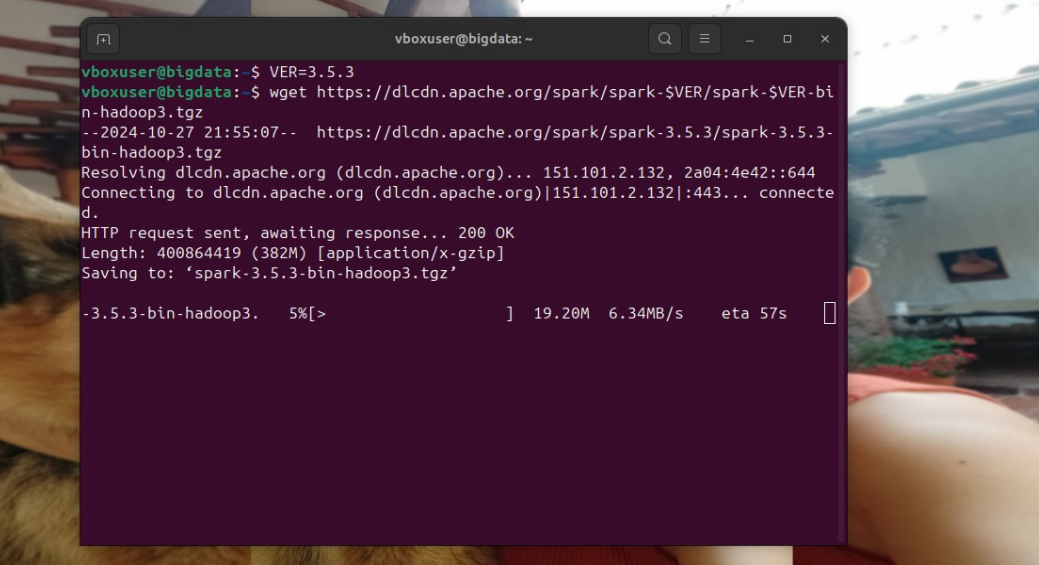
```
oct 27 16:36
BIGDATA [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
vboxuser@bigdata:~$ mkdir -p ~/.kaggle
vboxuser@bigdata:~$ mv ~/kaggle.json ~/.kaggle/
vboxuser@bigdata:~$ chmod 600 ~/.kaggle/kaggle.json
vboxuser@bigdata:~$ ls -la ~/.kaggle
-rw-r--r-- 1 vboxuser vboxuser 128 Oct 27 16:36 kaggle.json
vboxuser@bigdata:~$
```

Descargar el archivo kaggle, comando:

```
kaggle datasets download -d oscardavidperilla/domestic-violence-in-colombia
```

VER=3.5.3

```
wget https://dlcdn.apache.org/spark/spark-$VER/spark-$VER-bin-hadoop3.tgz
```



The screenshot shows a Linux desktop environment. In the background, there is a video of a person's legs and a building. A terminal window is open in the foreground, displaying the following commands and output:

```
vboxuser@bigdata:~  
vboxuser@bigdata:~$ VER=3.5.3  
vboxuser@bigdata:~$ wget https://dlcdn.apache.org/spark/spark-$VER/spark-$VER-bin-hadoop3.tgz  
--2024-10-27 21:55:07-- https://dlcdn.apache.org/spark/spark-3.5.3/spark-3.5.3-bin-hadoop3.tgz  
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644  
Connecting to dlcdn.apache.org (dlcdn.apache.org)[151.101.2.132]:443... connecte  
d.  
HTTP request sent, awaiting response... 200 OK  
Length: 400864419 (382M) [application/x-gzip]  
Saving to: 'spark-3.5.3-bin-hadoop3.tgz'  
  
-3.5.3-bin-hadoop3. 5%> ] 19.20M 6.34MB/s eta 57s
```

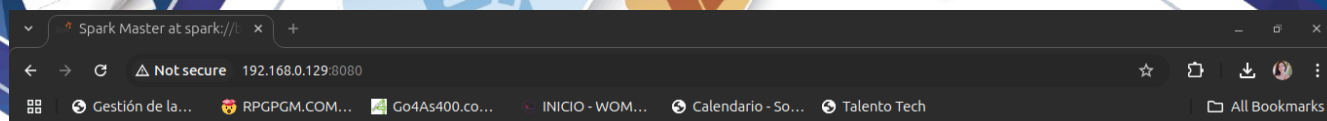
The terminal window has a title bar that reads "vboxuser@bigdata:~". The desktop background features a video of a person's legs and a building. The system tray at the bottom shows various icons, including a file manager, a web browser, and a terminal.

`tar xvf spark-$VER-bin-hadoop3.tgz`

```
vboxuser@bigdata: ~  
2024-10-27 21:56:07 (6.40 MB/s) - 'spark-3.5.3-bin-hadoop3.tgz' saved [400864419  
/400864419]  
  
vboxuser@bigdata:~$ tar xvf spark-$VER-bin-hadoop3.tgz  
spark-3.5.3-bin-hadoop3/  
spark-3.5.3-bin-hadoop3/data/  
spark-3.5.3-bin-hadoop3/data/graphx/  
spark-3.5.3-bin-hadoop3/data/graphx/users.txt  
spark-3.5.3-bin-hadoop3/data/graphx/followers.txt  
spark-3.5.3-bin-hadoop3/data/mllib/  
spark-3.5.3-bin-hadoop3/data/mllib/sample_linear_regression_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/sample_fpgrowth.txt  
spark-3.5.3-bin-hadoop3/data/mllib/sample_libsvm_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/gmm_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/kmeans_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/streaming_kmeans_data_test.txt  
spark-3.5.3-bin-hadoop3/data/mllib/sample_lda_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/sample_multiclass_classification_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/pagerank_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/sample_isotonic_regression_libsvm_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/sample_lda_libsvm_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/sample_movielens_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/pic_data.txt  
spark-3.5.3-bin-hadoop3/data/mllib/sample_binary_classification_data.txt
```

Como se ve spark después de instalar

Ingresando al navegador a: <http://192.168.0.129:8080/>:



Spark Master at spark://bigdata:7077

URL: spark://bigdata:7077
 Alive Workers: 0
 Cores in use: 0 Total, 0 Used
 Memory in use: 0.0 B Total, 0.0 B Used
 Resources in use:
 Applications: 0 Running, 0 Completed
 Drivers: 0 Running, 0 Completed
 Status: ALIVE

Workers (0)

Worker Id	Address	State	Cores	Memory	Resources
-----------	---------	-------	-------	--------	-----------

Running Applications (0)

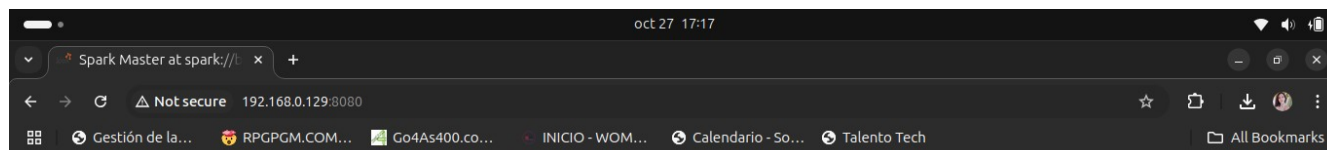
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Usar este comando:
 start-slave.sh spark://bigdata:7077

Actualizar



Spark Master at spark://bigdata:7077

URL: spark://bigdata:7077
 Alive Workers: 1
 Cores in use: 3 Total, 0 Used
 Memory in use: 2.8 GiB Total, 0.0 B Used
 Resources in use:
 Applications: 0 Running, 0 Completed
 Drivers: 0 Running, 0 Completed
 Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241027221653-192.168.0.129-45733	192.168.0.129:45733	ALIVE	3 (0 Used)	2.8 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

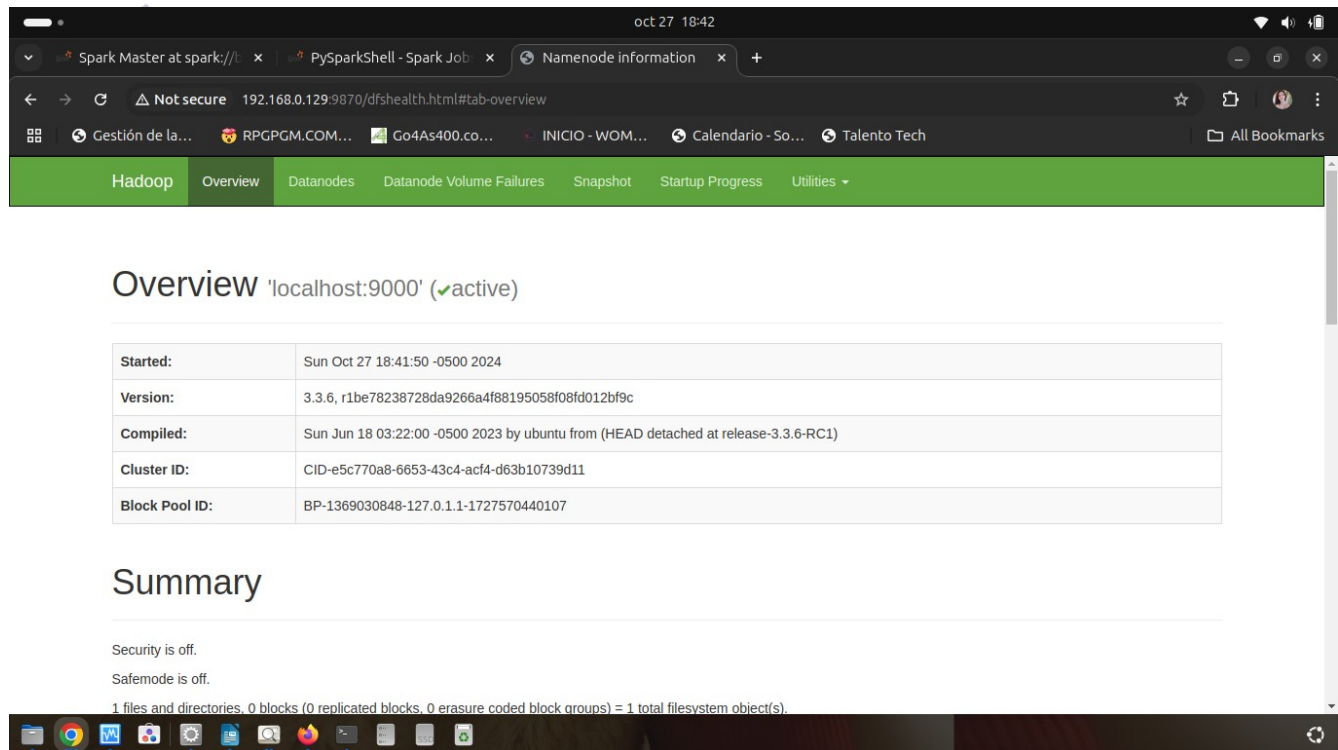
Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------



Hadoop

Ingresamos con: <http://192.168.0.129:9870/>



The screenshot shows a web browser window displaying the Hadoop NameNode interface. The browser tabs include 'Spark Master at spark://l...', 'PySparkShell - Spark Job', and 'Namenode information'. The address bar shows the URL '192.168.0.129:9870/dfshealth.html#tab-overview'. The page has a green navigation bar with tabs: 'Hadoop', 'Overview', 'Datanodes', 'Datanode Volume Failures', 'Snapshot', 'Startup Progress', and 'Utilities'. The main content area is titled 'Overview localhost:9000 (active)' and contains a table with the following information:

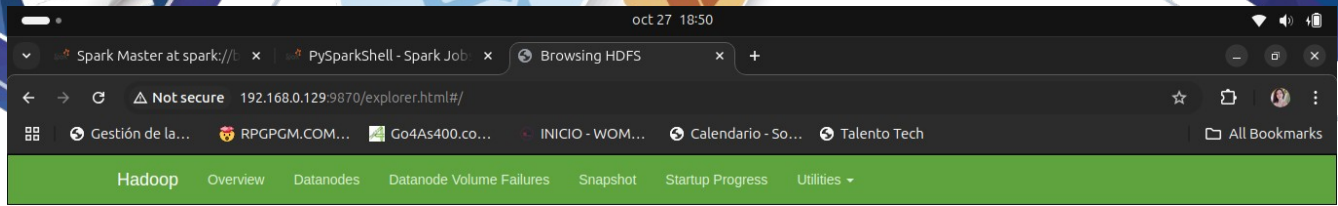
Started:	Sun Oct 27 18:41:50 -0500 2024
Version:	3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c
Compiled:	Sun Jun 18 03:22:00 -0500 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-e5c770a8-6653-43c4-acf4-d63b10739d11
Block Pool ID:	BP-1369030848-127.0.1.1-1727570440107

Below the table is a 'Summary' section with the following text:

Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Se crea una carpeta en el sistema HDFS, con el siguiente comando:

`hdfs dfs -mkdir /Tarea3`



Browse Directory

Show 25 entries

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 27 18:50	0	0 B	Tarea3	<input type="button" value="🗑"/>

Showing 1 to 1 of 1 entries

1

Hadoop, 2023.

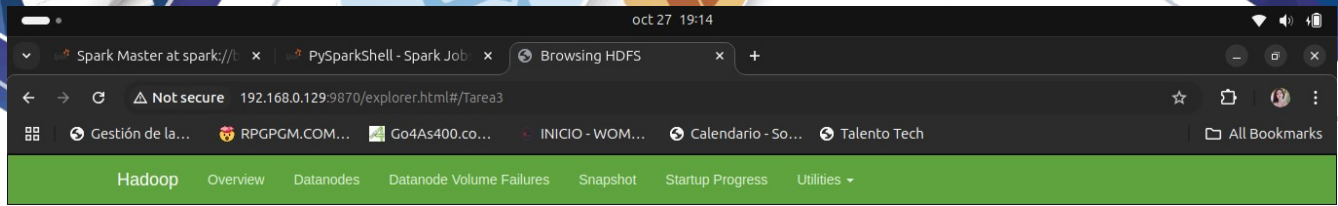


Copiamos el archivo del dataset a la carpeta HDFS que creamos, con el siguiente comando:

```
hdfs dfs -put /home/hadoop/Reporte_Delito_Violencia_Intrafamiliar_Polic_a_Nacional.csv
```

```
hadoop@bigdata: ~  
hadoop@bigdata:~$ pwd  
/home/hadoop  
hadoop@bigdata:~$ hdfs dfs -put /home/hadoop/Reporte_Delito_Violencia_Intrafamil  
iar_Polic_a_Nacional.csv /Tarea3  
hadoop@bigdata:~$
```

Podemos ver como se copio el archivo



Browse Directory

/Tarea3

Show entries

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	34.93 MB	Oct 27 19:13	1	128 MB	Reporte_Delito_Violencia_Intrafamiliar_Polic_a_Nacional.csv

Showing 1 to 1 of 1 entries

Hadoop, 2023.



Operaciones básica con Spark

Creamos un archivo py llamado tarea3 y añadimos el código para el ejercicio, en este caso sobre violencia intrafamiliar en Colombia.

Código:

Importamos librerías necesarias

```
from pyspark.sql import SparkSession, functions as F
```

Inicializa la sesión de Spark

```
spark = SparkSession.builder.appName('Tarea3').getOrCreate()
```

Define la ruta del archivo .csv en HDFS

```
file_path =  
'hdfs://localhost:9000/Tarea3/Reporte_Delito_Violencia_Intrafamiliar_Polic_a_Nacional.csv'  
'v'
```




```
# Lee el archivo .csv
```

```
df = spark.read.format('csv').option('header','true').option('inferSchema',  
'true').load(file_path)
```

```
# Imprimimos el esquema
```

```
df.printSchema()
```

```
# Muestra las primeras filas del DataFrame
```

```
df.show()
```

```
# Estadísticas básicas
```

```
df.summary().show()
```

```
# Consulta: Filtrar por valor y seleccionar columnas
```

```
print("Dias con valor mayor a 5000\n")
```

```
dias = df.filter(F.col('VALOR') >  
5000).select('VALOR','VIGENCIADESDE','VIGENCIAHASTA')
```


```
dias.show()
```

```
# Ordenar filas por los valores en la columna "VALOR" en orden descendente
```

```
print("Valores ordenados de mayor a menor\n")
```

```
sorted_df = df.sort(F.col("VALOR").desc())
```

```
sorted_df.show()
```



```
vboxuser@bigdata: ~
GNU nano 6.2                                tarea3.py
## Importamos librerias necesarias
from pyspark.sql import SparkSession, functions as F

# Inicializa la sesión de Spark
spark = SparkSession.builder.appName('Tarea3').getOrCreate()

# Define la ruta del archivo .csv en HDFS
file_path = 'hdfs://localhost:9000/Tarea3/Reporte_Delito_Violencia_Intrafamilia>

# Lee el archivo .csv
df = spark.read.format('csv').option('header','true').option('inferSchema', 'tr>

# Imprimimos el esquema
df.printSchema()

# Muestra las primeras filas del DataFrame
df.show()

# Estadísticas básicas
df.summary().show()

[ Read 30 lines ]
^G Help      ^O Write Out ^W Where Is  ^K Cut       ^T Execute   ^C Location
^X Exit      ^R Read File ^\ Replace   ^U Paste     ^J Justify   ^/ Go To Line
```

Se ejecuta el script con el siguiente comando:

```
python3 tarea3.py
```

```
vboxuser@bigdata: ~  
Para ver estas actualizaciones adicionales, ejecute: apt list --upgradable  
  
Active ESM Apps para recibir futuras actualizaciones de seguridad adicionales.  
Vea https://ubuntu.com/esm o ejecute «sudo pro status»  
  
New release '24.04.1 LTS' available.  
Run 'do-release-upgrade' to upgrade to it.  
  
Last login: Mon Oct 28 00:06:25 2024 from 192.168.0.129  
vboxuser@bigdata:~$ nano tarea3.py  
vboxuser@bigdata:~$ nano tarea.py  
vboxuser@bigdata:~$ nano tarea3.py  
vboxuser@bigdata:~$ python3 tarea3.py  
24/10/28 00:57:44 WARN Utils: Your hostname, bigdata resolves to a loopback address: 127.0.1.1; using 192.168.0.129 instead (on interface enp0s3)  
24/10/28 00:57:44 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
24/10/28 00:57:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

vboxuser@bigdata: ~

DEPARTAMENTO	MUNICIPIO	CODIGO DANE	ARMAS MEDIOS	FECHA HECHO
GENERO	GRUPO ETARIO	CANTIDAD		
ATLÁNTICO	BARRANQUILLA (CT)	8001000	ARMA BLANCA / COR...	1/01/2010
MAS CULINO	ADULTOS	1		
BOYACÁ	DUITAMA	15238000	ARMA BLANCA / COR...	1/01/2010
FE MENINO	ADULTOS	1		
CAQUETÁ	PUERTO RICO	18592000	ARMA BLANCA / COR...	1/01/2010
MAS CULINO	ADULTOS	1		
CASANARE	MANÍ	85139000	ARMA BLANCA / COR...	1/01/2010
FE MENINO	ADULTOS	1		
CUNDINAMARCA	BOGOTÁ D.C. (CT)	11001000	ARMA BLANCA / COR...	1/01/2010
FE MENINO	ADULTOS	1		
SUCRE	SINCELEJO (CT)	70001000	ARMA BLANCA / COR...	1/01/2010
FE MENINO	ADULTOS	1		
VALLE	CALI (CT)	76001000	ARMA BLANCA / COR...	1/01/2010
FE MENINO	ADULTOS	2		
VALLE	CALI (CT)	76001000	ARMA BLANCA / COR...	1/01/2010
MAS CULINO	ADULTOS	2		
VALLE	TULUÁ	76834000	ARMA BLANCA / COR...	1/01/2010
FE MENINO	ADULTOS	1		


```
vboxuser@bigdata: ~
```

summary	DEPARTAMENTO	MUNICIPIO	CODIGO DANE	ARMAS MEDIOS	F
ECHA HECHO	GENERO	GRUPO ETARIO	CANTIDAD		
count	476970	476970	476970	476968	
476970	476967	475355	476970		
mean	NULL	NULL	3.775064627790055E7	NULL	
44277.5	NULL	NULL	1.7077635910015305		
stddev	NULL	NULL	2.749371870652713E7	NULL	47.21682
5088399965	NULL	NULL	3.338647115163376		
min	AMAZONAS	ABEJORRAL	11001000	-	
1/01/2010	-	ADOLESCENTES	1		
25%	NULL	NULL	1.1001E7	NULL	
44245.0	NULL	NULL	1		
50%	NULL	NULL	2.5754E7	NULL	
44262.0	NULL	NULL	1		
75%	NULL	NULL	6.8001E7	NULL	
44300.0	NULL	NULL	1		
max	VICHADA	ÚTICA	NO REPORTA	SIN EMPLEO DE ARMAS	
9/12/2020	NO REPORTA	NO REPORTA	130		

```
vboxuser@bigdata: ~
Dias con valor mayor a 5000

Traceback (most recent call last):
  File "/home/vboxuser/tarea3.py", line 24, in <module>
    dias = df.filter(F.col('VALOR') > 5000).select('VALOR','VIGENCIADESDE','VIGENCIAHASTA')
  File "/home/vboxuser/.local/lib/python3.10/site-packages/pyspark/sql/dataframe.py", line 3331, in filter
    jdf = self._jdf.filter(condition._jc)
  File "/home/vboxuser/.local/lib/python3.10/site-packages/py4j/java_gateway.py", line 1322, in __call__
    return_value = get_return_value(
  File "/home/vboxuser/.local/lib/python3.10/site-packages/pyspark/errors/exceptions/captured.py", line 185, in deco
    raise converted from None
pyspark.errors.exceptions.captured.AnalysisException: [UNRESOLVED_COLUMN.WITH_SUGGESTION] A column or function parameter with name `VALOR` cannot be resolved. Did you mean one of the following? [`GENERO`, `CANTIDAD`, `MUNICIPIO`, `FECHA HECHO`, `ARMAS MEDIOS`].;
'Filter ('VALOR > 5000)
+- Relation [DEPARTAMENTO#17,MUNICIPIO#18,CODIGO DANE#19,ARMAS MEDIOS#20,FECHA HECHO#21,GENERO#22,GRUPO ETARIO#23,CANTIDAD#24] csv

vboxuser@bigdata:~$
```

Mientras el script se está ejecutando, se puede ver la información accediendo a :
<http://192.168.0.129:4040>

oct 27 20:02

Spark Master at spark://l... x Tarea3 - Spark Jobs x Browsing HDFS x +

← → ↻ ⚠ Not secure 192.168.0.129:4040/jobs/ ☆ 📁 All Bookmarks

Gestión de la... RPGPGM.COM... Go4As400.co... INICIO - WOM... Calendario - So... Talento Tech

APACHE **Spark** 3.5.3 Jobs Stages Storage Environment Executors SQL / DataFrame Tarea3 application UI

Spark Jobs (?)

User: vboxuser
Total Uptime: 8 s
Scheduling Mode: FIFO
Active Jobs: 1

▶ Event Timeline

▼ Active Jobs (1)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▼	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	load at NativeMethodAccessorImpl.java:0 load at NativeMethodAccessorImpl.java:0 (kill)	2024/10/28 01:02:12	91 ms	0/1	0/1

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go



oct 27 20:03

Spark Master at spark://l... x Tarea3 - Environment x Browsing HDFS x +

← → ↻ ⚠ Not secure 192.168.0.129:4040/environment/ ☆ 📁 All Bookmarks

Gestión de la... RPGPGM.COM... Go4As400.co... INICIO - WOM... Calendario - So... Talento Tech

APACHE **Spark** 3.5.3 Jobs Stages Storage **Environment** Executors SQL / DataFrame Tarea3 application UI

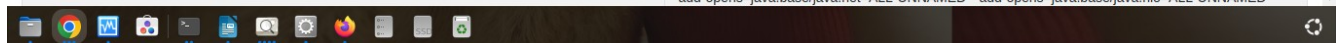
Environment

▼ Runtime Information

Name	Value
Java Home	/usr/lib/jvm/java-11-openjdk-amd64
Java Version	11.0.24 (Ubuntu)
Scala Version	version 2.12.18

▼ Spark Properties

Name	Value
spark.app.id	local-1730077418976
spark.app.name	Tarea3
spark.app.startTime	1730077417693
spark.app.submitTime	1730077417430
spark.driver.extraJavaOptions	-Djava.net.preferIPv6Addresses=false -XX:+IgnoreUnrecognizedVMOptions --add-opens=java.base/java.lang=ALL-UNNAMED --add-opens=java.base/java.lang.invoke=ALL-UNNAMED --add-opens=java.base/java.lang.reflect=ALL-UNNAMED --add-opens=java.base/java.io=ALL-UNNAMED --add-opens=java.base/java.net=ALL-UNNAMED --add-opens=java.base/java.nio=ALL-UNNAMED --



oct 27 20:04

Spark Master at spark://l... x Tarea3 - Executors x Browsing HDFS x +

← → ↻ ⚠ Not secure 192.168.0.129:4040/executors/ ☆ 📁 👤 ⋮

📁 Gestión de la... 🏠 RPGPGM.COM... 📄 Go4As400.co... 🏠 INICIO - WOM... 📅 Calendario - So... 🧑 Talento Tech 📁 All Bookmarks

APACHE **spark** 3.5.3 Jobs Stages Storage Environment **Executors** SQL / DataFrame Tarea3 application UI

Executors

▶ Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	91.7 KiB / 434.4 MiB	0.0 B	3	3	0	5	8	12 s (0.3 s)	35.1 MiB	0.0 B	0.0 B	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	91.7 KiB / 434.4 MiB	0.0 B	3	3	0	5	8	12 s (0.3 s)	35.1 MiB	0.0 B	0.0 B	0

Executors

Show 20 entries Search:

Task

Running Queries: 1
Completed Queries: 2

Running Queries (1)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

ID	Description	Submitted	Duration	Running Job IDs	Succeeded Job IDs	Failed Job IDs
2	showString at NativeMethodAccessorImpl.java:0	2024/10/28 01:05:47	20 s	[3]		

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Completed Queries (2)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

ID	Description	Submitted	Duration	Job IDs
1	showString at NativeMethodAccessorImpl.java:0	2024/10/28 01:05:46	0.3 s	[2]
0	load at NativeMethodAccessorImpl.java:0	2024/10/28 01:05:42	2 s	[0]

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Análisis de Datos en Tiempo Real con Spark Streaming y Kafka

Crear un archivo llamado kafka_producer.py

A ese archivo se le agrega el siguiente código:

```
import time
```

```
import json
```

```
import pandas as pd
```

```
from kafka import KafkaProducer
```

```
from hdfs import InsecureClient
```

```
# Conectar con HDFS
```

```
client = InsecureClient('http://localhost:9870', user='hadoop')
```

```
# Leer el archivo CSV desde HDFS
```

```
with client.read('/Tarea3/Reporte_Delito_Violencia_Intrafamiliar_Polic_a_Nacional.csv',  
encoding='utf-8') as reader:
```

```
    df = pd.read_csv(reader)
```

```
# Configurar el productor de Kafka
```

```
producer = KafkaProducer(bootstrap_servers=['localhost:9092'],  
                           value_serializer=lambda x: json.dumps(x).encode('utf-8'))
```

```
# Enviar datos al tópico de Kafka
```

```
for index, row in df.iterrows():
```

```
    data = row.to_dict()
```

```
    producer.send('sensor_data', value=data) # Aquí usamos 'sensor_data' como el tópico de  
Kafka
```

```
    print(f"Sent: {data}")
```

```
    time.sleep(1) # Simular la llegada de datos en tiempo real
```

```

vboxuser@bigdata: ~
GNU nano 6.2 kafka_producer.py
import time
import json
import random
from kafka import KafkaProducer

def generate_sensor_data():
    return {
        "sensor_id": random.randint(1, 10),
        "temperature": round(random.uniform(20, 30), 2),
        "humidity": round(random.uniform(30, 70), 2),
        "timestamp": int(time.time())
    }

producer = KafkaProducer(
    bootstrap_servers=['localhost:9092'],
    value_serializer=lambda x: json.dumps(x).encode('utf-8')
)

while True:
    sensor_data = generate_sensor_data()
    [ Read 23 lines ]
^G Help      ^O Write Out ^W Where Is  ^K Cut       ^T Execute   ^C Location
^X Exit      ^R Read File ^\ Replace   ^U Paste     ^J Justify   ^/ Go To Line

```

Implementación del consumidor con Spark Streaming

Se crea un archivo con el `spark_streaming_consumer.py` y se copia el siguiente código:

```

from pyspark.sql import SparkSession

from pyspark.sql.functions import col, window, from_json

from pyspark.sql.types import StructType, StructField, StringType, IntegerType,
TimestampType

# Configura el nivel de log a WARN para reducir los mensajes INFO
spark = SparkSession.builder \
    .appName("KafkaSparkStreaming") \
    .getOrCreate()

```

```
spark.sparkContext.setLogLevel("WARN")
```

```
# Definir el esquema de los datos de entrada basado en el dataset de Kaggle
```

```
schema = StructType([
    StructField("Año", IntegerType()),
    StructField("Mes", StringType()),
    StructField("Código Departamento", IntegerType()),
    StructField("Departamento", StringType()),
    StructField("Código Municipio", IntegerType()),
    StructField("Municipio", StringType()),
    StructField("Edad", IntegerType()),
    StructField("Sexo", StringType()),
    StructField("Estado Civil", StringType()),
    StructField("Nivel Educativo", StringType()),
    StructField("Fecha Hecho", TimestampType())
])
```

```
# Configurar el lector de streaming para leer desde Kafka
```

```
kafka_df = spark.readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "localhost:9092") \
    .option("subscribe", "sensor_data") \
    .load()
```

```
# Parsear los datos JSON de Kafka
```

```
parsed_df = kafka_df.select(from_json(col("value").cast("string"),
schema).alias("data")).select("data.*")
```

```
# Calcular estadísticas por ventana de tiempo
```



```
windowed_stats = parsed_df \
    .groupBy(window(col("Fecha Hecho"), "1 minute"), "Departamento") \
    .agg({"Edad": "avg"})
```

Escribir los resultados en la consola

```
query = windowed_stats \
    .writeStream \
    .outputMode("complete") \
    .format("console") \
    .start()
```

```
query.awaitTermination()
```

```

vboxuser@bigdata: ~
GNU nano 6.2 spark_streaming_consumer.py *
from pyspark.sql import SparkSession
from pyspark.sql.functions import from_json, col, window
from pyspark.sql.types import StructType, StructField, IntegerType, FloatType, >
import logging

# Configura el nivel de log a WARN para reducir los mensajes INFO
spark = SparkSession.builder \
    .appName("KafkaSparkStreaming") \
    .getOrCreate()
spark.sparkContext.setLogLevel("WARN")

# Definir el esquema de los datos de entrada
schema = StructType([
    StructField("sensor_id", IntegerType()),
    StructField("temperature", FloatType()),
    StructField("humidity", FloatType()),
    StructField("timestamp", TimestampType())
])

# Configurar el lector de streaming para leer desde Kafka

^G Help      ^O Write Out ^W Where Is  ^K Cut       ^T Execute   ^C Location
^X Exit      ^R Read File ^\ Replace   ^U Paste     ^J Justify   ^/ Go To Line
  
```

Ejecución y Análisis

Se ejecuta con el siguiente comando, este va a ser del productor de Kafka

```
python3 spark_streaming_consumer.py
```

```
vboxuser@bigdata: ~
vboxuser@bigdata:~$ nano spark_streaming_consumer.py
vboxuser@bigdata:~$ nano spark_streaming_consumer.py
vboxuser@bigdata:~$ python3 spark_streaming_consumer.py
24/10/28 01:24:02 WARN Utils: Your hostname, bigdata resolves to a loopback address: 127.0.1.1; using 192.168.0.129 instead (on interface enp0s3)
24/10/28 01:24:02 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/10/28 01:24:02 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Traceback (most recent call last):
  File "/home/vboxuser/spark_streaming_consumer.py", line 26, in <module>
    .load()
  File "/home/vboxuser/.local/lib/python3.10/site-packages/pyspark/sql/streaming/readwriter.py", line 304, in load
    return self._df(self._jreader.load())
  File "/home/vboxuser/.local/lib/python3.10/site-packages/py4j/java_gateway.py", line 1322, in __call__
    return_value = get_return_value(
  File "/home/vboxuser/.local/lib/python3.10/site-packages/pyspark/errors/exceptions/captured.py", line 185, in deco
    raise converted from None
```

Ejecutar este comando:

ssh vboxuser@192.168.0.129

```
vboxuser@bigdata: ~  
* Documentation: https://help.ubuntu.com  
* Management:   https://landscape.canonical.com  
* Support:       https://ubuntu.com/pro  
  
System information as of lun 28 oct 2024 01:29:10 UTC  
  
System load: 0.1          Processes:              155  
Usage of /:  89.1% of 13.67GB Users logged in:          2  
Memory usage: 57%         IPv4 address for enp0s3: 192.168.0.129  
Swap usage:  0%  
  
=> / is using 89.1% of 13.67GB  
  
* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s  
  just raised the bar for easy, resilient and secure K8s cluster deployment.  
  
  https://ubuntu.com/engage/secure-kubernetes-at-the-edge  
  
El mantenimiento de seguridad expandido para Applications está desactivado  
Se pueden aplicar 5 actualizaciones de forma inmediata.  
Para ver estas actualizaciones adicionales, ejecute: apt list --upgradable
```

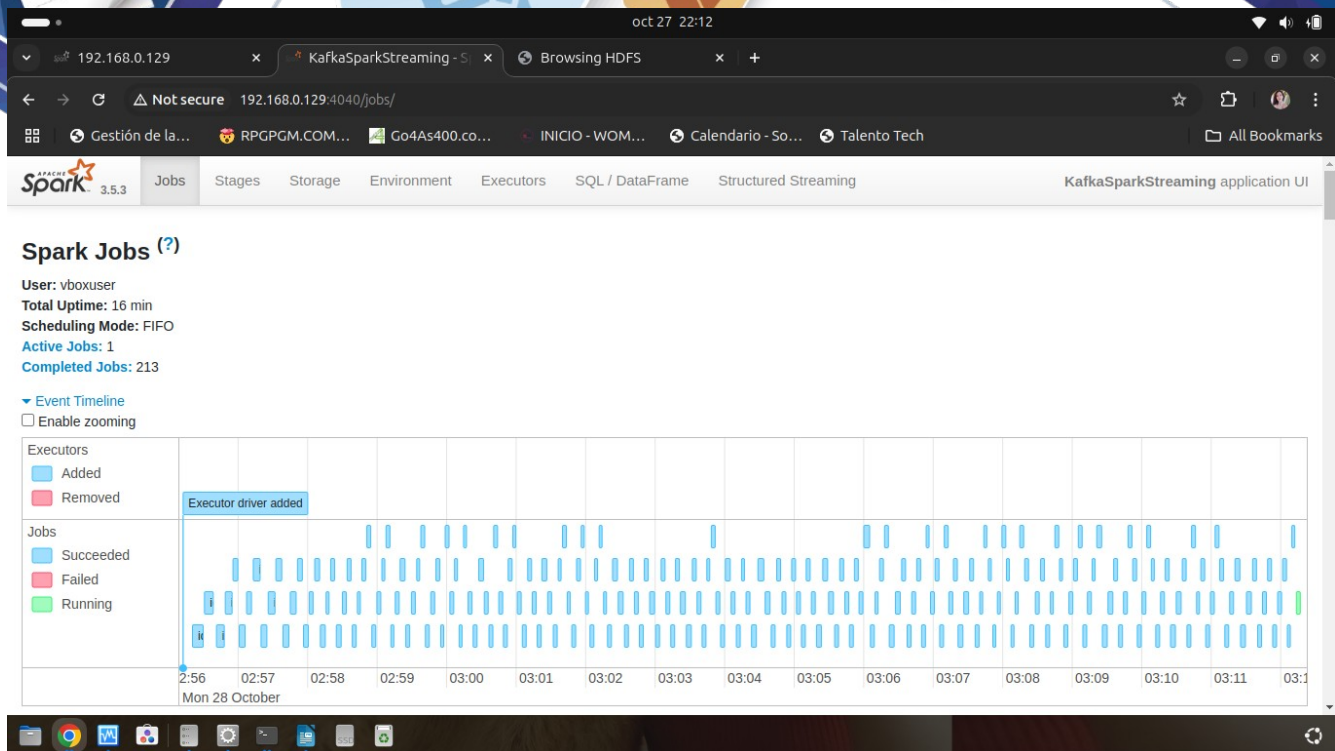
Una vez logueado, ejecutar el siguiente comando para iniciar el consumirdo de Spark Streaming

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.3  
spark_streaming_consumer.py
```



```
vboxuser@bigdata: ~  
New release '24.04.1 LTS' available.  
Run 'do-release-upgrade' to upgrade to it.  
  
Last login: Mon Oct 28 00:16:25 2024 from 192.168.0.153  
vboxuser@bigdata:~$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.3 spark_streaming_consumer.py  
24/10/28 01:32:52 WARN Utils: Your hostname, bigdata resolves to a loopback address: 127.0.1.1; using 192.168.0.129 instead (on interface enp0s3)  
24/10/28 01:32:52 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address  
INFO: loading settings :: url = jar:file:/opt/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml  
Ivy Default Cache set to: /home/vboxuser/.ivy2/cache  
The jars for the packages stored in: /home/vboxuser/.ivy2/jars  
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency  
INFO: resolving dependencies :: org.apache.spark#spark-submit-parent-efa10c52-92ab-4bd5-b72e-00276dd29e70;1.0  
   confs: [default]  
   found org.apache.spark#spark-sql-kafka-0-10_2.12;3.5.3 in central  
   found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.5.3 in central  
   found org.apache.kafka#kafka-clients;3.4.1 in central
```

Ver la información en: <http://192.168.0.129:4040>



oct 27 22:15

KafkaSparkStreaming - S x Browsing HDFS x +

192.168.0.129 x

Not secure 192.168.0.129:4040/stages/

Gestión de la... RPGPGM.COM... Go4As400.co... INICIO - WOM... Calendario - So... Talento Tech

All Bookmarks

spark 3.5.3 Jobs Stages Storage Environment Executors SQL / DataFrame Structured Streaming KafkaSparkStreaming application UI

Stages for All Jobs

Active Stages: 1
Completed Stages: 500
Skipped Stages: 1

Active Stages (1)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
501	id = 295930b2-5135-4d6e-8c0f-2f4aba4b744c runId = e9953dd0-b17d-4f81-8573-92230e2b... start at NativeMethodAccessorImpl.java:0	2024/10/28 03:14:59	0.8 s	38/200 (4 running)			103.0 B	

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

Completed Stages (500)

Page: 1 2 3 4 5 > 5 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
----------	-------------	-----------	----------	------------------------	-------	--------	--------------	---------------

oct 27 22:17

192.168.0.129 x KafkaSparkStreaming - E x Browsing HDFS x +

← → ↻ ⚠ Not secure 192.168.0.129:4040/environment/ ☆ 📁 All Bookmarks

Gestión de la... RPGPGM.COM... Go4As400.co... INICIO - WOM... Calendario - So... Talento Tech

spark 3.5.3 Jobs Stages Storage Environment Executors SQL / DataFrame Structured Streaming KafkaSparkStreaming application UI

Environment

▼ Runtime Information

Name	Value
Java Home	/usr/lib/jvm/java-11-openjdk-amd64
Java Version	11.0.24 (Ubuntu)
Scala Version	version 2.12.18

▼ Spark Properties

Name	Value
spark.app.id	local-1730084171199
spark.app.initial.file.urls	***** (redacted)
spark.app.initial.jar.urls	***** (redacted)
spark.app.name	KafkaSparkStreaming
spark.app.startTime	1730084169256
spark.app.submitTime	1730084167936
spark.driver.extraJavaOptions	-Djava.net.preferIPv6Addresses=false -XX:+IgnoreUnrecognizedVMOptions --add-

oct 27 22:20

192.168.0.129 x KafkaSparkStreaming - E x Browsing HDFS x +

← → ↻ ⚠ Not secure 192.168.0.129:4040/executors/ ☆ 📁 All Bookmarks

Gestión de la... RPGPGM.COM... Go4As400.co... INICIO - WOM... Calendario - So... Talento Tech

spark 3.5.3 Jobs Stages Storage Environment Executors SQL / DataFrame Structured Streaming KafkaSparkStreaming application UI

Executors

► Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	107.3 KiB / 434.4 MiB	0.0 B	3	4	0	63413	63417	24 min (14 s)	0.0 B	118.5 KiB	118.7 KiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	107.3 KiB / 434.4 MiB	0.0 B	3	4	0	63413	63417	24 min (14 s)	0.0 B	118.5 KiB	118.7 KiB	0

Executors

Show 20 entries Search:

Task

oct 27 22:20

192.168.0.129 x KafkaSparkStreaming - S x Browsing HDFS x +

← → ↻ Not secure 192.168.0.129:4040/SQL/ ☆ 📌 👤 ⋮

📦 Gestión de la... 🏠 RPGPGM.COM... 📄 Go4As400.co... 🏠 INICIO - WOM... 📅 Calendario - So... 🏠 Talento Tech 📁 All Bookmarks

APACHE spark 3.5.3 Jobs Stages Storage Environment Executors **SQL / DataFrame** Structured Streaming KafkaSparkStreaming application UI

SQL / DataFrame

Running Queries: 1
Completed Queries: 319

▼ Running Queries (1)

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

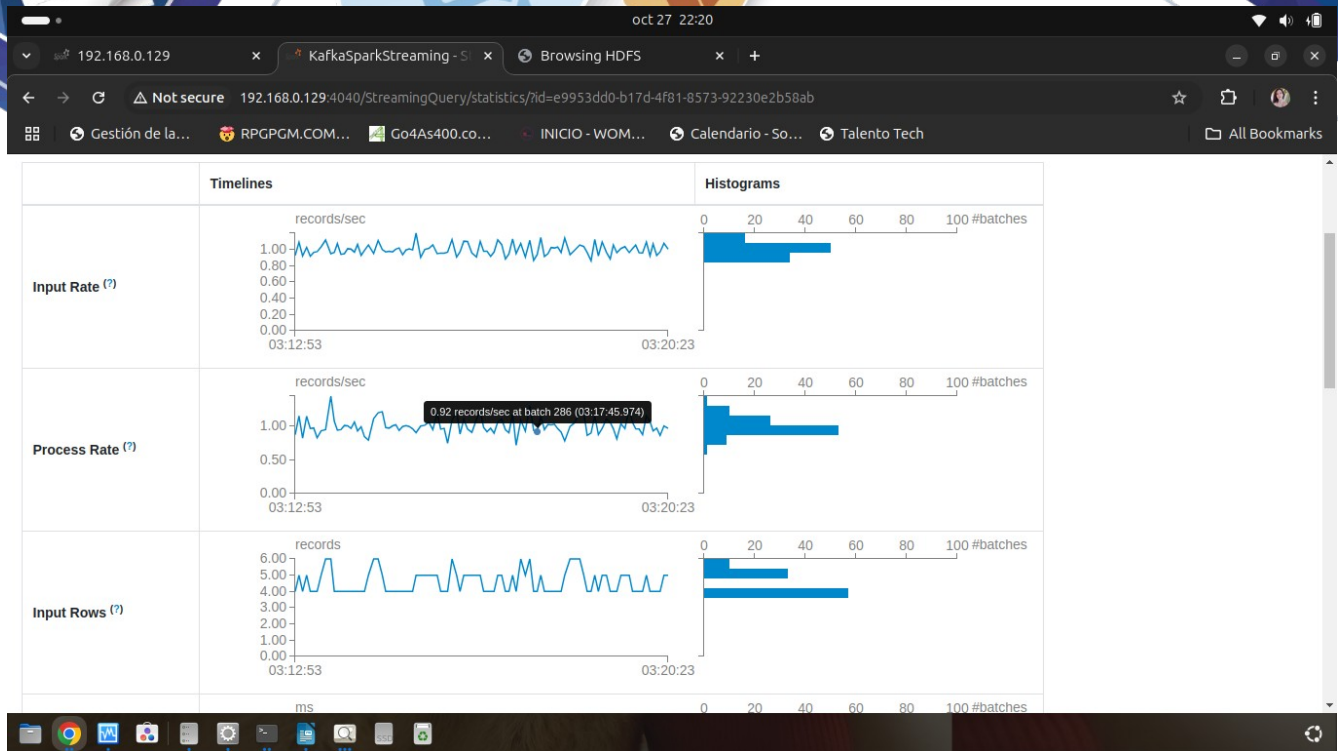
ID ▼	Description	Submitted	Duration	Running Job IDs	Succeeded Job IDs	Failed Job IDs	Sub Execution IDs
957	id = 295930b2-5135-4d6e-8c0f-2f4aba4b744c runId = e9953dd0-b17d-4f81-8573-92230e2b... +details	2024/10/28 03:20:14	4 s				[958] +details

Page: 1 1 Pages. Jump to 1 . Show 100 items in a page. Go

▼ Completed Queries (319)

Page: 1 2 3 4 > 4 Pages. Jump to 1 . Show 100 items in a page. Go

ID ▼	Description	Submitted	Duration	Job IDs	Sub Execution IDs
954	id = 295930b2-5135-4d6e-8c0f-2f4aba4b744c runId = e9953dd0-b17d-4f81-8573-92230e2b58ab batch = 318 +details	2024/10/28 03:20:10	4 s		[955][956] +details



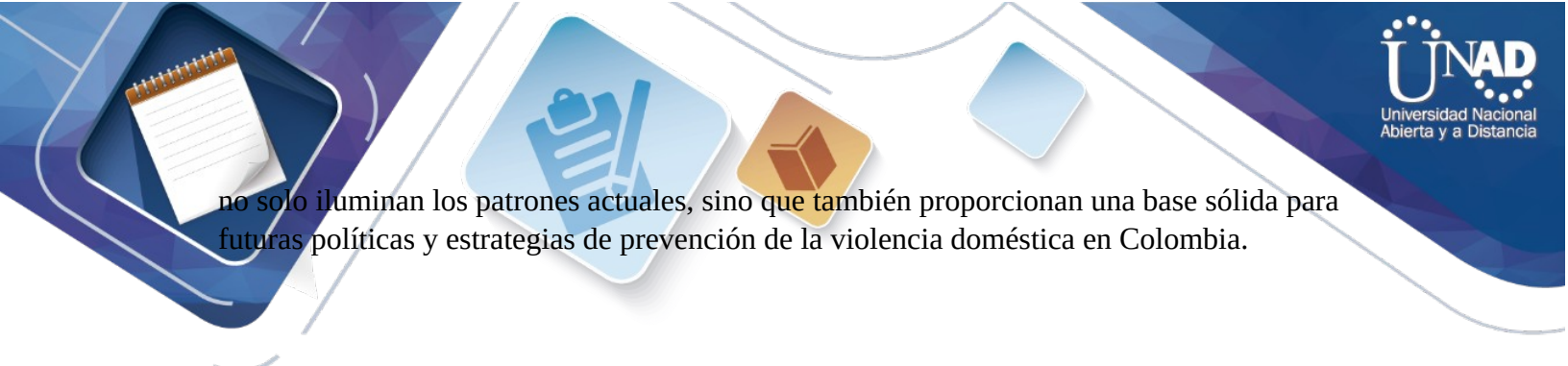
Análisis de los resultados obtenidos

A lo largo del tiempo, los datos sobre violencia doméstica en Colombia han revelado patrones preocupantes. La limpieza inicial de los datos nos permitió identificar que la violencia física es la más reportada, seguida de cerca por la violencia psicológica. Normalizamos los formatos de fecha y corregimos errores tipográficos para asegurar la precisión del análisis.

En la fase de transformación, codificamos las variables categóricas y escalamos las variables numéricas, lo que nos permitió realizar un análisis más profundo. Descubrimos que los incidentes de violencia doméstica tienden a aumentar en meses específicos, coincidiendo con periodos de estrés económico y social.

El análisis geográfico mostró que las regiones urbanas, como Bogotá y Medellín, registran una mayor incidencia de violencia doméstica en comparación con las áreas rurales. Esta disparidad podría estar vinculada a factores socioeconómicos y la disponibilidad de servicios de apoyo en las ciudades.

El análisis de correlación reveló una relación significativa entre la edad de las víctimas y el tipo de violencia, con víctimas más jóvenes enfrentando más violencia física. Esto subraya la necesidad de intervenciones específicas para diferentes grupos de edad. Estos resultados



no solo iluminan los patrones actuales, sino que también proporcionan una base sólida para futuras políticas y estrategias de prevención de la violencia doméstica en Colombia.

