

Máster en Big Data

Tecnologías de Almacenamiento

9. Hands-On: Spark

PRESENTADO:

JOSE DAVID ANGULO GARCIA

ALBERT RIPOLL

Índice

1. Introducción.....	3
2. Entorno	3
3. Inspección de los datos locales	3

1. Introducción

El objetivo de este Hands-On es el de familiarizarse con una de las librerías más populares del framework de Spark como es SparkSQL

2. Entorno

Para la realización de los ejercicios se va a utilizar *spark-shell* en scala o python ya que nos proporciona un entorno muy dinámico para la introducción de funciones y nos permite recibir una respuesta inmediata.

Para ello, nos descargaremos Spark y levantaremos una *spark-shell* en local.

Sigue las siguientes instrucciones si es necesario:

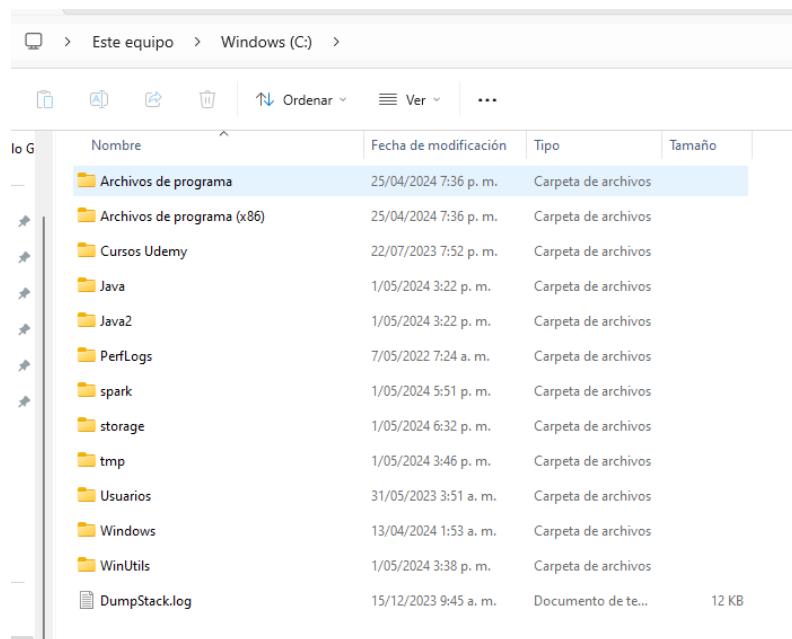
<https://sparkbyexamples.com/spark/install-apache-spark-on-mac/>

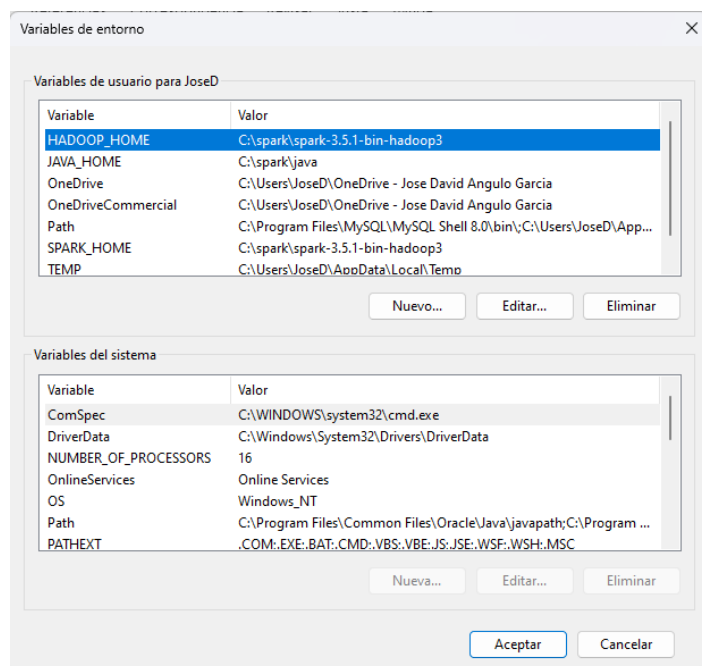
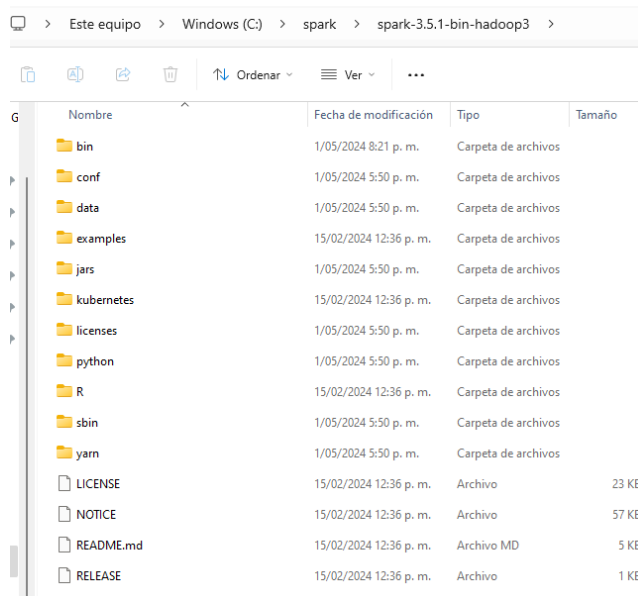
<https://sparkbyexamples.com/spark/apache-spark-installation-on-windows/>

<https://sparkbyexamples.com/spark/spark-installation-on-linux-ubuntu/>

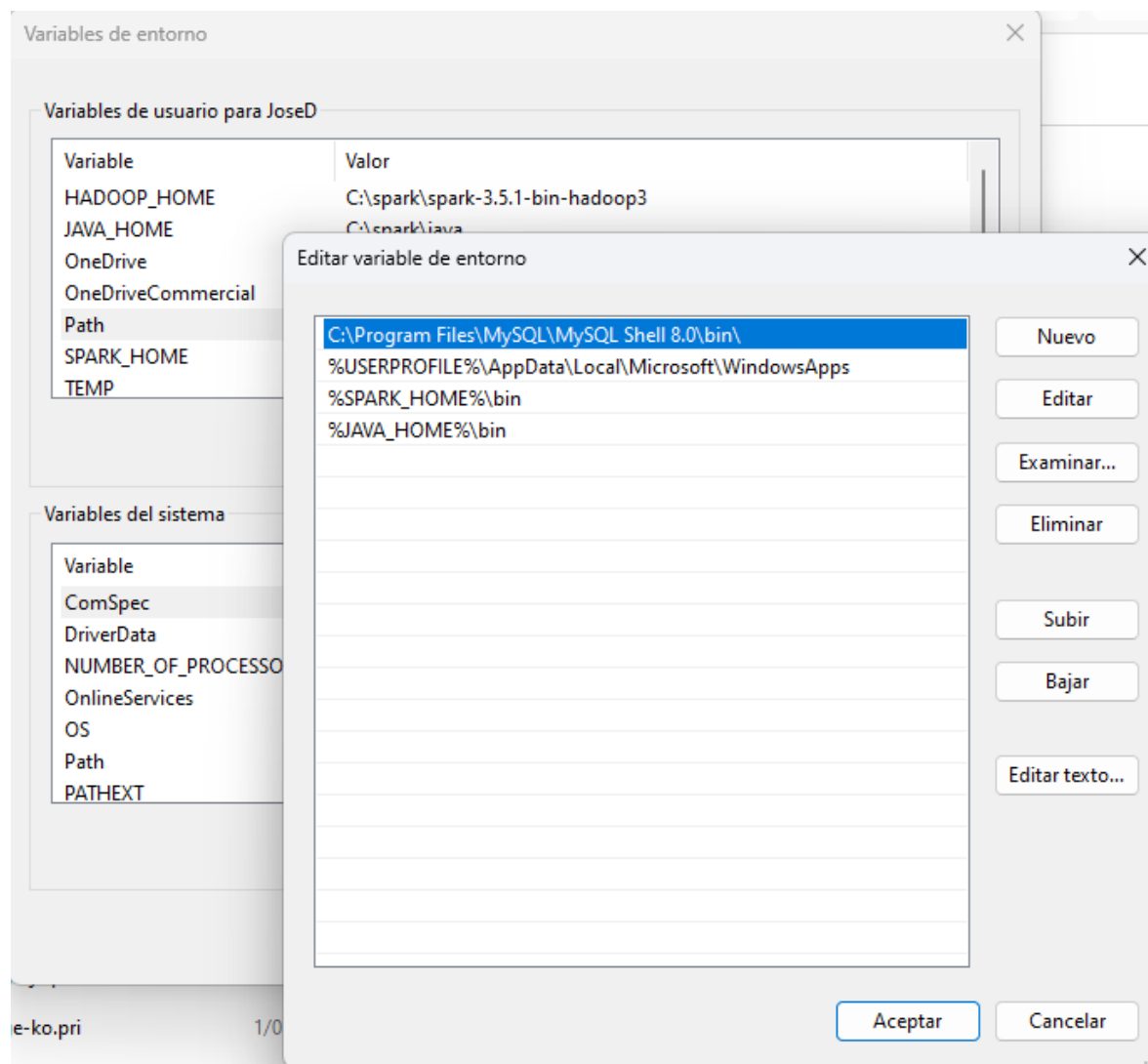
Configuración del ENTORNO

Creación de las carpetas donde se alojarán los instaladores y ficheros

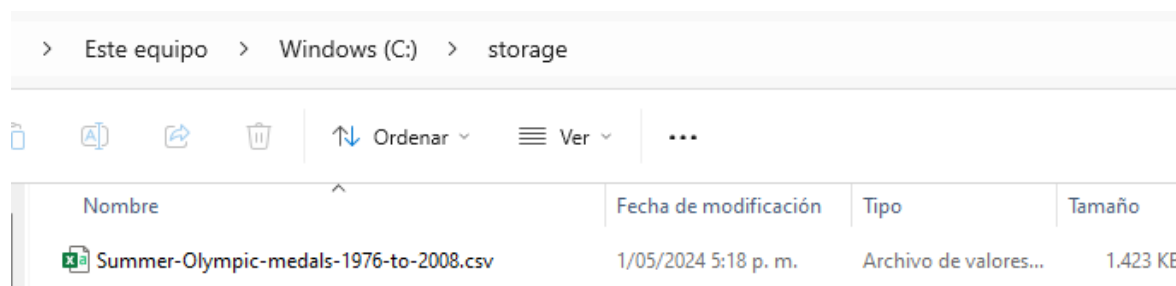




Creación de los path



Guardamos el dataset a utilizar

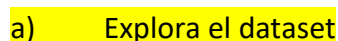


El dataset que utilizaremos se llama *Summer-Olympic-medals-1976-to-2008.csv*.



```
val dataset = spark.read.option("header", "true").option("inferSchema",
"true").csv("/path/to/Summer-Olympic-medals-1976-to-2008.csv")
```

```
val olym_dataset =
  spark.read.option("header","true").option("inferSchema","true").csv("C:/storage/Summer-
  Olympic-medals-1976-to-2008.csv")
```



```
scala> val olym_dataset = spark.read.option("header", "true").option("inferSchema", "true").csv("C:/storage/Summer-Olympic-medals-1976")
olym_dataset: org.apache.spark.sql.DataFrame = [City: string, Year: int ... 9 more fields]

scala> olym_dataset.show()
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| City|Year| Sport|Discipline| Event| Athlete|Gender|Country_Code| Country|Event_gender| Medal|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Montreal|1976|Aquatics| Diving| 3m springboard| K?HLER, Christa| Women| GDR| East Germany| W| Silver|
|Montreal|1976|Aquatics| Diving| 3m springboard| KOSENKOV, Aleksandr| Men| URS| Soviet Union| M| Bronze|
|Montreal|1976|Aquatics| Diving| 3m springboard| BOGGS, Philip George| Men| USA| United States| M| Gold|
|Montreal|1976|Aquatics| Diving| 3m springboard| CAGNOTTO, Giorgio...| Men| ITA| Italy| M| Silver|
|Montreal|1976|Aquatics| Diving| 10m platform| WILSON, Deborah K...| Women| USA| United States| W| Bronze|
|Montreal|1976|Aquatics| Diving| 10m platform| LOUGANIS, Gregory| Men| USA| United States| M| Silver|
|Montreal|1976|Aquatics| Diving| 10m platform| VAYTSEKHOVSKAYA, ...| Women| URS| Soviet Union| W| Gold|
|Montreal|1976|Aquatics| Diving| 3m springboard| POTTER-MCINGVALE,...| Women| USA| United States| W| Bronze|
|Montreal|1976|Aquatics| Diving| 10m platform| DIBIASI, Klaus| Men| ITA| Italy| M| Gold|
|Montreal|1976|Aquatics| Diving| 10m platform| ALEINIK, Vladimir| Men| URS| Soviet Union| M| Bronze|
|Montreal|1976|Aquatics| Diving| 10m platform| KNAPE-LINDBERGH, ...| Women| SWE| Sweden| W| Silver|
|Montreal|1976|Aquatics| Diving| 3m springboard| CHANDLER, Jennife...| Women| USA| United States| W| Gold|
|Montreal|1976|Aquatics| Swimming| 4x100m freestyle ...| BABASHOFF, Shirle...| Women| USA| United States| W| Gold|
|Montreal|1976|Aquatics| Swimming| 400m freestyle| SHAW, Timothy Andrew| Men| USA| United States| W| Gold|
```

b) ¿Que han aportado las opciones “header” y “inferSchema”?

`olym_dataset.head()`

Este comando permite ver todos los datos de la primera fila del dataset explorado sin el nombre de las columnas.

`olym_dataset.schema`

permite ver las características que tiene cada columna del dataset como el tipo de dato.

```
C:\Users\JoseD\Downloads\st x + -
|Montreal|1976|Aquatics| Swimming| 800m freestyle| TH?MER, Petra| Women| GDR| East Germany| W| Gold|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

scala> olym_dataset.head()
res1: org.apache.spark.sql.Row = [Montreal,1976,Aquatics,Diving,3m springboard,K?HLER, Christa,Women,GDR,East Germany,W, Silver]

scala> olym_dataset.schema
res2: org.apache.spark.sql.types.StructType = StructType(StructField(City,StringType,true),StructField(Year,IntegerType,true),StructField(Sport,StringType,true),StructField(Discipline,StringType,true),StructField(Event,StringType,true),StructField(Athlete,StringType,true),StructField(Gender,StringType,true),StructField(Country_Code,StringType,true),StructField(Country,StringType,true),StructField(Event_gender,StringType,true),StructField(Medal,StringType,true))

scala>
```

c) ¿Cómo harías para contar las medallas conseguidas por año y país?

Primero importar la librería de funciones de sql para que el spark pueda interpretar las sentencias que se le van a entregar a través de una variable, donde se agrupara por año y país, contando las medallas que tiene cada uno.

```
scala> import org.apache.spark.sql.functions._
import org.apache.spark.sql.functions._
```

```
scala> val totalmedallas = oym_dataset.groupBy("Year", "Country").agg(count("Medal"))
totalmedallas: org.apache.spark.sql.DataFrame = [Year: int, Country: string ... 1 more field]
```

```
scala> totalmedallas.show
+-----+-----+-----+
|Year|    Country|count(Medal)|
+-----+-----+-----+
|1996|    Romania|          38|
|2000|    Russia|         188|
|2004|    India|           1|
|2008|   Portugal|           2|
|2008|   Mauritius|          1|
|1980|    Finland|           9|
|1996|    Austria|           3|
|1988|   Switzerland|          8|
|1992|    China|          83|
|1996|    Ukraine|          34|
|2000|    Algeria|           5|
|2008|   Netherlands|          62|
|1980|    Guyana|           1|
|1988|   Soviet Union|         294|
|1976|    Pakistan|          16|
|1976|   Korea, South|          17|
|1976|    Jamaica|           2|
|1980|   East Germany|         260|
|1984|    Nigeria|           5|
|1992|   Czechoslovakia|          8|
+-----+-----+-----+
only showing top 20 rows

scala>
```

d) Usando SparkSQL muestra alguna métrica interesante.

Para poder ejecutar alguna sentencia select, se debe crear una vista de la variable que contiene el dataset. Y luego crear otra variable que contenga la consulta a realizar. Como sigue:

```
oym_dataset.createTempView("agrupacion")
```

```
val cantciudad = spark.sql("""SELECT City, Sport, COUNT(*) AS CantMedallas FROM agrupacion
GROUP BY City, Sport ORDER BY CantMedallas DESC""")
```



```
scala> olyn_dataset.createTempView("agrupacion")

scala> val cantciudad = spark.sql("""SELECT City,Sport,COUNT(*) AS CantMedallas FROM agrupacion GROUP BY City,Sport ORDE
R BY CantMedallas DESC""")
cantciudad: org.apache.spark.sql.DataFrame = [City: string, Sport: string ... 1 more field]

scala> cantciudad.show
+-----+-----+-----+
| City | Sport | CantMedallas |
+-----+-----+-----+
| Beijing | Aquatics | 347 |
| Athens | Aquatics | 332 |
| Sydney | Aquatics | 329 |
| Atlanta | Aquatics | 262 |
| Barcelona | Aquatics | 228 |
| Seoul | Aquatics | 202 |
| Los Angeles | Aquatics | 192 |
| Sydney | Athletics | 184 |
| Athens | Athletics | 183 |
| Atlanta | Athletics | 180 |
| Barcelona | Athletics | 178 |
| Beijing | Athletics | 177 |
| Seoul | Athletics | 163 |
| Moscow | Rowing | 162 |
| Los Angeles | Rowing | 162 |
| Montreal | Rowing | 162 |
| Los Angeles | Athletics | 161 |
| Seoul | Rowing | 159 |
| Moscow | Aquatics | 159 |
| Montreal | Aquatics | 159 |
+-----+-----+-----+
only showing top 20 rows
```

e) ¿Que muestra la SparkUI: <http://localhost:4040> ?

SPARK_UI LOCAL: <http://joseangulo:4040/jobs/>

Observar que cada imagen es la vista de las opciones de arriba de esta pagina: JOBS, STAGES, ENVIROMENT, EXECUTORS, SQL

VISTA JOBS: Los Trabajos que se han ejecutado con sus características principales de performance

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
7	show at <console>-27 show at <console>-27	2024/05/01 20:28:38	0.1 s	1/1 (1 skipped)	1/1 (1 skipped)
6	show at <console>-27 show at <console>-27	2024/05/01 20:28:37	1 s	1/1	1/1
5	show at <console>-27 show at <console>-27	2024/05/01 20:03:47	0.1 s	1/1 (1 skipped)	1/1 (1 skipped)
4	show at <console>-27 show at <console>-27	2024/05/01 20:03:45	2 s	1/1	1/1
3	head at <console>-24 head at <console>-24	2024/05/01 19:48:07	27 ms	1/1	1/1
2	show at <console>-24 show at <console>-24	2024/05/01 19:36:12	0.2 s	1/1	1/1
1	csv at <console>-22 csv at <console>-22	2024/05/01 19:34:18	0.3 s	1/1	1/1
0	csv at <console>-22 csv at <console>-22	2024/05/01 19:34:17	0.5 s	1/1	1/1

VISTA DE STAGES:

Spark shell - Stages for All Jobs

Completed Stages: 8
Skipped Stages: 2

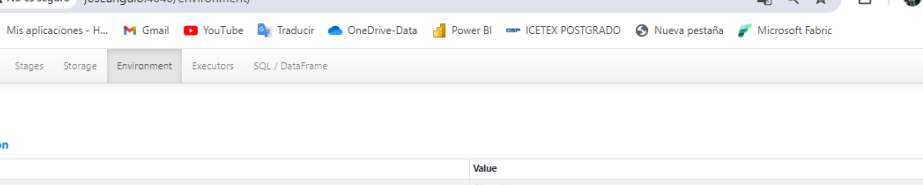
Completed Stages (8)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
9	show at <console>:27	2024/05/01 20:28:38	0,1 s	1/1			15,7 KiB	
7	show at <console>:27	2024/05/01 20:28:37	1 s	1/1	1422,9 KiB			15,7 KiB
6	show at <console>:27	2024/05/01 20:03:47	0,1 s	1/1			24,7 KiB	
4	show at <console>:27	2024/05/01 20:03:45	2 s	1/1	1422,9 KiB			24,7 KiB
3	head at <console>:24	2024/05/01 19:48:07	27 ms	1/1	64,0 KiB			
2	show at <console>:24	2024/05/01 19:36:12	0,1 s	1/1	64,0 KiB			
1	csv at <console>:22	2024/05/01 19:34:18	0,3 s	1/1	1422,9 KiB			
0	csv at <console>:22	2024/05/01 19:34:18	0,4 s	1/1	64,0 KiB			

Skipped Stages (2)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
8	show at <console>:27	Unknown	Unknown	0/1				
5	show at <console>:27	Unknown	Unknown	0/1				

VISTA DE LOS ENVIRONMENT:



Spark shell - Environment

No es seguro joseangulo:4040/environment/

Udemy Cursos Mis aplicaciones - H... Gmail YouTube Traducir OneDrive-Data Power BI ICETEX POSTGRADO Nueva pestaña Microsoft Fabric Todos los marcados

Spark shell application UI

Environment

Runtime Information

Name	Value
Java Home	C:\spark\java
Java Version	22.0.1 (Oracle Corporation)
Scala Version	version 2.12.18

Spark Properties

Name	Value
spark.app.id	local-1714584199122
spark.app.name	Spark shell
spark.app.startTime	1714584197562
spark.app.submitTime	1714584191342
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 --add-opens=java.base/java.util=ALL-UNNAMED --add-opens=java.base/java.util.concurrent=ALL-UNNAMED --add-opens=java.base/java.util.concurrent.atomic=ALL-UNNAMED --add-opens=java.base/jdk.internal.ref=ALL-UNNAMED --add-opens=java.base/sun.nio.ch=ALL-UNNAMED --add-opens=java.base/sun.nio.cs=ALL-UNNAMED --add-opens=java.base/sun.security.action=ALL-UNNAMED --add-opens=java.base/sun.util.calendar=ALL-UNNAMED --add-opens=java.security.jgss/sun.security.krb5=ALL-UNNAMED -Djdk.reflect.useDirectMethodHandle=false
spark.driver.host	JOSEANGULO
spark.driver.port	49918
spark.executor.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 --add-opens=java.base/java.util=ALL-UNNAMED --add-opens=java.base/java.util.concurrent=ALL-UNNAMED --add-opens=java.base/java.util.concurrent.atomic=ALL-UNNAMED --add-opens=java.base/jdk.internal.ref=ALL-UNNAMED --add-opens=java.base/sun.nio.ch=ALL-UNNAMED --add-opens=java.base/sun.nio.cs=ALL-UNNAMED --add-opens=java.base/sun.security.action=ALL-UNNAMED --add-opens=java.base/sun.util.calendar=ALL-UNNAMED --add-opens=java.security.jgss/sun.security.krb5=ALL-UNNAMED -Djdk.reflect.useDirectMethodHandle=false

VISTA DE LOS EXECUTORS:

Spark shell - Executors

No es seguro joseangulo4040/executors/

Udemy Cursos

Mis aplicaciones - H...

Gmail

YouTube

Traducir

OneDrive-Data

Power BI

ICETEX POSTGRADO

Nueva pestaña

Microsoft Fabric

Reinicia para actualizar

Todos los marcadores

Spark

3.5.1

JobsStagesStorageEnvironmentExecutorsSQL / DataFrame

Spark shell application l

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	20.4 KB / 434.4 MiB	0.0 B	16	0	0	8	8	1.2 h (0.5 s)	4.4 MiB	40.4 KiB	40.4 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	20.4 KB / 434.4 MiB	0.0 B	16	0	0	8	8	1.2 h (0.5 s)	4.4 MiB	40.4 KiB	40.4 KiB	0

Executors

Show 20 entries

Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump	Heap Histogram	Add Time	Remove Time
driver	JOSEANGULO@9919	Active	0	20.4 KB / 434.4 MiB	0.0 B	16	0	0	8	8	1.2 h (0.5 s)	4.4 MiB	40.4 KiB	40.4 KiB	Thread Dump	Heap Histogram	2024-05-01 19:23:19	-

Showing 1 to 1 of 1 entries

Previous 1 Next

VISTA DE SQL:

SQL / DataFrame

Completed Queries: 6

- Completed Queries (6)

ID	Description	Submitted	Duration	Job IDs
5	show at <console>:27	2024/05/01 20:28:37	2 s	[6][7]
4	createTempView at <console>:27	2024/05/01 20:26:05	10 ms	
3	show at <console>:27	2024/05/01 20:03:45	2 s	[4][5]
2	head at <console>:24	2024/05/01 19:48:07	98 ms	[3]
1	show at <console>:24	2024/05/01 19:36:12	0.3 s	[2]
0	csv at <console>:22 csv at <console>:22 <pre>org.apache.spark.sql.DataFrameReader.csv(DataFrameReader.scala:444) \$line14.\$read\$\$\$iwid\$\$iwid\$\$iwid\$\$iwid\$.init(<console>:22) \$line14.\$read\$\$\$iwid\$\$iwid\$\$iwid\$\$iwid\$.init(<console>:26) \$line14.\$read\$\$\$iwid\$\$iwid\$\$iwid\$\$iwid\$.init(<console>:28) \$line14.\$read\$\$\$iwid\$\$iwid\$\$iwid\$\$iwid\$.init(<console>:30) \$line14.\$read\$\$\$iwid\$\$iwid\$.init(<console>:32) \$line14.\$read\$\$\$iwid\$\$iwid\$.init(<console>:34) \$line14.\$read\$\$\$iwid\$.init(<console>:36) \$line14.\$read\$\$\$iwid\$.init(<console>:38) \$line14.\$read\$.init(<console>:40) \$line14.\$read\$.init(<console>:44)</pre>	2024/05/01 19:34:17	1 s	[0]