



**UNIVERSIDAD MAYOR DE SAN SIMÓN**  
**FACULTAD DE CIENCIAS Y TECNOLOGÍA**

**DIRECCIÓN DE POSGRADO**



# **DIPLOMADO ESTADÍSTICA APLICADA A LA TOMA DE DECISIONES**

## **SEGUNDA VERSIÓN**

**Laboratorio Integración hive Spark**

**NOMBRE : HEBERT JUAN DE DIOS DELGADILLO FERNANDEZ**  
**CARLOS ALFREDO ORIHUELA BERRIOS**

**DOCENTE : DANNY LUIS HUANCA SEVILLA**

**Cochabamba – Bolivia**

**2023**

Instalando pyspark en COLAB, usando el comando `!pip install pyspark`

```
[2] !pip install pyspark
    #import findspark
    #findspark.init()

Collecting pyspark
  Downloading pyspark-3.4.1.tar.gz (310.8 MB)
    310.8/310.8 MB 2.3 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.4.1-py2.py3-none-any.whl size=311285387 sha256=c07154b711aa372423b25
  Stored in directory: /root/.cache/pip/wheels/0d/77/a3/ff2f74cc9ab41f8f594dabf0579c2a7c6de920d584206e0834
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.4.1
```

Importando la librería Spark, para poder interactuar con un clúster de Spark y realizar operaciones de procesamiento de datos distribuidas.

```
✓ [3] from pyspark.sql import SparkSession
0 s

✓ [4] spark = SparkSession.builder.appName("SesionDataframes").getOrCreate()
8 s
```

Verificamos las base datos existentes, las cuales solo hay una por defecto en Spark.

```
✓ [5] spark.sql("show databases").show()
19 s

+-----+
|namespace|
+-----+
| default|
+-----+
```

Seleccionamos la Base de datos **default**, luego revisamos las tablas que contiene. con el comando `spark.sql("show tables").show()`

```
✓ 0s [6] spark.sql("use default").show()

++
||
++
++
```

```
✓ 1s [7] spark.sql("show tables").show()

+-----+-----+-----+
|namespace|tableName|isTemporary|
+-----+-----+-----+
+-----+-----+-----+
```

Seguidamente creamos una nueva Base de Datos llamada **prueba\_olimpica**, con el comando `spark.sql("create database prueba_olimpica")`

```
✓ 0s [8] spark.sql("create database prueba_olimpica")

DataFrame[]
```

Verificamos si se creó exitosamente con el comando

`spark.sql("show databases").show()`

```
✓ 0s [9] spark.sql("show databases").show()

+-----+
|      namespace      |
+-----+
|      default        |
|prueba_olimpica      |
+-----+
```

Seleccionamos la Base de datos **prueba\_olimpica**, y creamos una tabla llamada **deportes**, con sus atributos o columnas.

```
✓ [10] spark.sql("use prueba_olimpica")
0 s
DataFrame[]

✓ [11] spark.sql("create table deportes \
1 s      (ID int,Name string,Sex string,Age int,Height int,Weight int ,Team string, NOC string, Games string,Year int,Season string, Cit
      row format delimited fields terminated by ',' \
      stored as textfile ")
DataFrame[]
```

Verificamos si la tabla de creo con éxito usando el comando

```
spark.sql("show tables").show()
```

```
✓ [12] spark.sql("show tables").show()
1 s

+-----+-----+-----+
| namespace | tableName | isTemporary |
+-----+-----+-----+
| prueba_olimpica | deportes | false |
+-----+-----+-----+
```

Revisamos sus atributos o columnas con el comando,

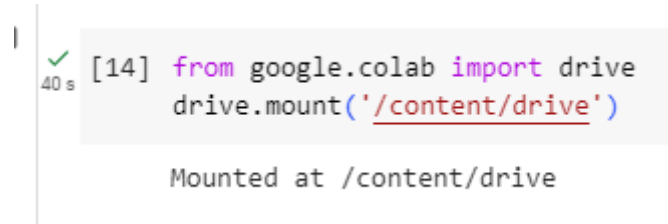
```
spark.sql("select * from deportes").show()
```

```
✓ [13] spark.sql("select * from deportes").show()
1 s

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| ID|Name|Sex|Age|Height|Weight|Team|NOC|Games|Year|Season|City|Sport|Event|Medal|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
```

Montando drive en COLAB, para utilizar datos almacenados. usando los siguientes comandos..

```
from google.colab import drive
drive.mount('/content/drive')
```

A screenshot of a Jupyter Notebook cell. On the left, there is a green checkmark icon and the text '40 s'. The cell contains the following code: 

```
[14] from google.colab import drive
drive.mount('/content/drive')
```

 Below the code, the output text reads: 

```
Mounted at /content/drive
```

```
[14] from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive
```

Seleccionamos la cuenta de google Drive y presionamos **permitir** para finalizar con éxito montar la unidad de drive en nuestro notebook.

#### Confirma que confías en Google Drive for desktop

Puede que estés compartiendo información sensible con este sitio o esta aplicación. Puedes ver o retirar el acceso en cualquier momento en tu [cuenta de Google](#).

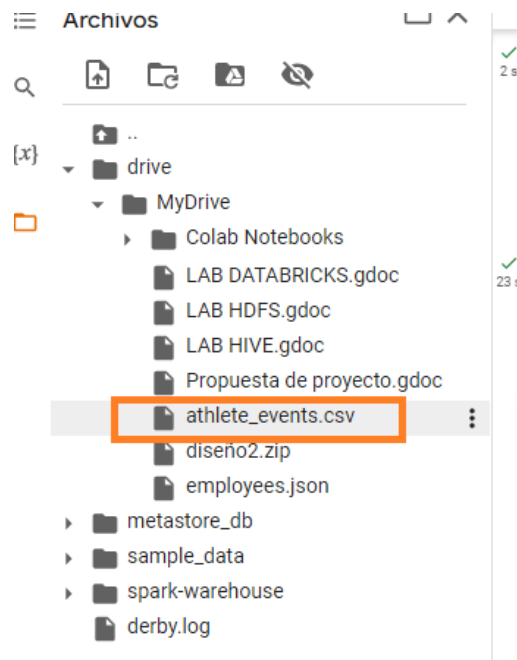
Descubre cómo te ayuda Google a [compartir datos de forma segura](#).

Consulta la [Política de Privacidad](#) y los [Términos del Servicio](#) de Google Drive for desktop.

Cancelar

Permitir

Verificamos que esté el archivo con los datos de **athlete\_events.csv** en la unidad virtual de COLAB



Cargamos los datos para poder trabajar desde la unidad de drive montado en COLAB, con el comando

```
spark.sql("load data local inpath  
'/content/drive/MyDrive/athlete_events.csv' overwrite into table  
deportes")
```

```
✓ [15] # desde google drive  
2 s spark.sql("load data local inpath '/content/drive/MyDrive/athlete_events.csv' overwrite into table deportes")  
  
DataFrame[]
```

Revisamos cuántos registros contiene, y luego mostramos los primeros 10 registros de la tabla deportes.

```
[31] spark.sql("select count(*) from deportes").show()
```

```
+-----+
|count(1)|
+-----+
|    271116|
+-----+
```

```
spark.sql("select * from deportes").show(10)
```

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport
1	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball
2	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer	London	Judo
3	Gunnar Nielsen Aabye	M	24	null	null	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football
4	Edgar Lindenau Aabye	M	34	null	null	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War
5	Christine Jacoba ...	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating
5	Christine Jacoba ...	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating
5	Christine Jacoba ...	F	25	185	82	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating
5	Christine Jacoba ...	F	25	185	82	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating
5	Christine Jacoba ...	F	27	185	82	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating
5	Christine Jacoba ...	F	27	185	82	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating

only showing top 10 rows

Mostramos la lista de todos los países y la cantidad de representantes de forma descendente con el comando

```
spark.sql("select NOC, count(*) from deportes group by NOC order by count(*) desc").show()
```

```
[33] spark.sql("select NOC, count(*) from deportes group by NOC order by count(*) desc").show()
```

```
+---+-----+
|NOC|count(1)|
+---+-----+
|USA|    17447|
|FRA|    12690|
|GBR|    12118|
|ITA|    10707|
|GER|     9718|
|CAN|     9641|
|JPN|     8444|
|SWE|     8284|
|AUS|     7598|
|HUN|     6503|
|SUI|     6145|
|POL|     6144|
|NED|     5811|
|URS|     5642|
|FIN|     5447|
|ESP|     5311|
|CHN|     5141|
|RUS|     5135|
|AUT|     5082|
|NOR|     4889|
+---+-----+
only showing top 20 rows
```

Listar los nombres de jugadores hombres de los Estados Unidos que juegan basketball y su edad es mayor a 28

```
spark.sql("select name from deportes where NOC = 'USA' and sex = 'M' and age > 28 and Sport = 'Basketball').show()
```

name
Carmelo Kyan Anthony
Charles Wade Barkley
Charles Wade Barkley
Larry Joe Bird
Kobe Bean Bryant
Kobe Bean Bryant
Tyson Cleotis Cha...
Clyde Austin Drexler
Patrick Aloysius ...
"Timothy Duane ""...
Allan Wade Houston
Allen Ezail Iverson
Michael Jeffrey J...
Jason Frederick Kidd
Kyle Lowry
Karl Malone
Karl Malone
"Reginald Wayne "...
Alonzo Harding Mo...
Hakeem Abdul Olaj...

only showing top 20 rows

Encontrar la lista de los jugadores que han jugado las olimpiadas de invierno después de 1952 y el Athletics Women's High Jump

```
k.sql("select year,Season,Event from deportes where Event='Athletics Womens High Jump' and year>1952 order by year,Season Desc")
```

year	Season	Event
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump
1956	Summer	Athletics Womens High Jump

Ese evento se realiza solo en verano por lo que el filtro de invierno no es necesario.

Encontrar el name, age, team de todos los jugadores de Denmark quienes hay jugado las Summer Olympics en Río de Janeiro



```

0s spark.sql("select name, age, Team from deportes where Team = 'Denmark' and Season = 'Summer' and City = 'Rio de Janeiro' ").show()

```

name	age	Team
Anne Dsane Andersen	23	Denmark
Simon Andreassen	18	Denmark
Viktor Axelsen	22	Denmark
Jacob Jepsen Barse	27	Denmark
Anne Sofie Holm B...	28	Denmark
Jakob Blbjerg Mat...	21	Denmark
Pernille Blume	22	Denmark
Pernille Blume	22	Denmark
Pernille Blume	22	Denmark
Pernille Blume	22	Denmark
Mathias Boe	36	Denmark
Frederik Lindbg B...	21	Denmark
Sarah Bro	20	Denmark
Nicolai Brock-Madsen	23	Denmark
Viktor Bregner Br...	23	Denmark
Andreas Hjartbro ...	29	Denmark
Lrke Buhl-Hansen	24	Denmark
Simone Tetsche Ch...	22	Denmark
Mads Christiansen	30	Denmark
Anders Dahl	40	Denmark

only showing top 20 rows

Encontrar el name y la age de todos los jugadores quienes han jugado Football como un deporte de los United States, France y Uruguay

```

0s [69] spark.sql("select name, age from deportes where Sport = 'Football' and NOC = 'USA' or NOC = 'FRA' or NOC = 'URU' ").show()

```

name	age
Jamale (Djamel-) ...	30
Patrick Abada	22
Ren Abadie	21
Ren Abadie	21
Luc Abalo	23
Luc Abalo	27
Luc Abalo	31
Jol Marc Abati	34
Jol Marc Abati	38
Ould Lamine Abdallah	null
Ben Ahmed Abdelkrim	20
Ramn Neptuno Abel...	26
David Abibssira	21
Camille Anne Fran...	27
Camille Anne Fran...	31
Sarah Abitbol	22
Stphan Abrahamian...	22
David Abrard	19
Franck Abrial	24
Julien Absalon	23

only showing top 20 rows

Encontrar el name y el NOC de jugadores quienes han jugado el **2012 Summer** y las olimpiadas de invierno **2006 Winter Olympics**

```

2s [98] spark.sql(" SELECT DISTINCT name,NOC from deportes where Games = '2006 Winter' \
AND name IN (SELECT name from deportes where Games = '2012 Summer') ").show()

```

name	NOC
Clara Hughes	CAN