Diplomado_DS_Mod13_Pyspark_sesion1

September 9, 2023

[1]:
```
spark
```

[1]:
```
<pyspark.sql.session.SparkSession at 0x7f9f63db8bb0>
```

[2]:
```
!pwd
```

```
/
```

[3]:
```
! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/03/2023-01.csv

! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/03/2023-02.csv

! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/04/2023-03.csv

! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/05/2023-04.csv

! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/06/2023-05.csv

! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/07/2023_06.csv

! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/08/2023_07.csv

! wget https://ecobici.cdmx.gob.mx/wp-content/uploads/2023/09/2023_08.csv
```

```
--2023-09-09 19:50:28--  https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/03/2023-01.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)… 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 41926344 (40M) [text/csv]
Saving to: ‘2023-01.csv’

2023-01.csv         100%[==================>]  39.98M  33.9MB/s    in 1.2s

2023-09-09 19:50:30 (33.9 MB/s) - ‘2023-01.csv’ saved [41926344/41926344]

--2023-09-09 19:50:30--  https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/03/2023-02.csv
```

```
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)… 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 47532884 (45M) [text/csv]
Saving to: ‘2023-02.csv’

2023-02.csv         100%[==================>]  45.33M  42.6MB/s    in 1.1s

2023-09-09 19:50:31 (42.6 MB/s) - ‘2023-02.csv’ saved [47532884/47532884]

--2023-09-09 19:50:31--  https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/04/2023-03.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)… 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 58449187 (56M) [text/csv]
Saving to: ‘2023-03.csv’

2023-03.csv         100%[==================>]  55.74M  34.4MB/s    in 1.6s

2023-09-09 19:50:33 (34.4 MB/s) - ‘2023-03.csv’ saved [58449187/58449187]

--2023-09-09 19:50:33--  https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/05/2023-04.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)… 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 58119473 (55M) [text/csv]
Saving to: ‘2023-04.csv’

2023-04.csv         100%[==================>]  55.43M  48.4MB/s    in 1.1s

2023-09-09 19:50:35 (48.4 MB/s) - ‘2023-04.csv’ saved [58119473/58119473]

--2023-09-09 19:50:35--  https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/06/2023-05.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)… 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 67532130 (64M) [text/csv]
Saving to: ‘2023-05.csv’

2023-05.csv         100%[==================>]  64.40M  48.6MB/s    in 1.3s
```

```
2023-09-09 19:50:36 (48.6 MB/s) - ‘2023-05.csv’ saved [67532130/67532130]

--2023-09-09 19:50:36--  https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/07/2023_06.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)… 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 72151312 (69M) [text/csv]
Saving to: ‘2023_06.csv’

2023_06.csv         100%[===================>]  68.81M  29.4MB/s    in 2.3s

2023-09-09 19:50:39 (29.4 MB/s) - ‘2023_06.csv’ saved [72151312/72151312]

--2023-09-09 19:50:39--  https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/08/2023_07.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)… 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 71416339 (68M) [text/csv]
Saving to: ‘2023_07.csv’

2023_07.csv         100%[===================>]  68.11M  52.8MB/s    in 1.3s

2023-09-09 19:50:41 (52.8 MB/s) - ‘2023_07.csv’ saved [71416339/71416339]

--2023-09-09 19:50:41--  https://ecobici.cdmx.gob.mx/wp-
content/uploads/2023/09/2023_08.csv
Resolving ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)… 157.230.72.27
Connecting to ecobici.cdmx.gob.mx (ecobici.cdmx.gob.mx)|157.230.72.27|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 79235355 (76M) [text/csv]
Saving to: ‘2023_08.csv’

2023_08.csv         100%[===================>]  75.56M  54.8MB/s    in 1.4s

2023-09-09 19:50:42 (54.8 MB/s) - ‘2023_08.csv’ saved [79235355/79235355]
```

[4]:
```
##Crear directorios HDFS, subir archivos y eliminar archivos locales
! hdfs dfs -mkdir -p /tmp/dcd/ecobici/
! hdfs dfs -put *.csv /tmp/dcd/ecobici/
! rm *.csv
```

[5]:
```
##Ver el contenido HDFS

! hdfs dfs -ls -R /tmp/dcd/ecobici/
```

```
-rw-r--r--   2 root hadoop   41926344 2023-09-09 19:52
/tmp/dcd/ecobici/2023-01.csv
-rw-r--r--   2 root hadoop   47532884 2023-09-09 19:52
/tmp/dcd/ecobici/2023-02.csv
-rw-r--r--   2 root hadoop   58449187 2023-09-09 19:52
/tmp/dcd/ecobici/2023-03.csv
-rw-r--r--   2 root hadoop   58119473 2023-09-09 19:52
/tmp/dcd/ecobici/2023-04.csv
-rw-r--r--   2 root hadoop   67532130 2023-09-09 19:52
/tmp/dcd/ecobici/2023-05.csv
-rw-r--r--   2 root hadoop   72151312 2023-09-09 19:52
/tmp/dcd/ecobici/2023_06.csv
-rw-r--r--   2 root hadoop   71416339 2023-09-09 19:52
/tmp/dcd/ecobici/2023_07.csv
-rw-r--r--   2 root hadoop   79235355 2023-09-09 19:52
/tmp/dcd/ecobici/2023_08.csv
```

[6]:
```
##Crear un DF con spark
eb_df = spark.read.csv('hdfs:///tmp/dcd/ecobici/', header=True,␣
 ↪inferSchema=True)
```

[7]:
```
##Ver los 10 primeros renglones
eb_df.show(10)
```

```
+--------------+------------+-------+--------------------+------------+--------
------------+--------------------+------------+--------------------+
|Genero_Usuario|Edad_Usuario|   Bici|Ciclo_Estacion_Retiro|Fecha_Retiro|
Hora_Retiro|Ciclo_EstacionArribo|Fecha Arribo|         Hora_Arribo|
+--------------+------------+-------+--------------------+------------+--------
------------+--------------------+------------+--------------------+
|             M|          24|2252039|                 007|
31/07/2023|2023-09-09 23:52:…|                064|  01/08/2023|2023-09-09
01:00:…|
|             M|          33|8626897|                 206|
31/07/2023|2023-09-09 23:48:…|                212|  01/08/2023|2023-09-09
01:00:…|
|             M|          34|4940557|                 215|
31/07/2023|2023-09-09 23:55:…|                212|  01/08/2023|2023-09-09
01:00:…|
|             F|          30|2036523|                 291|
31/07/2023|2023-09-09 23:32:…|                082|  01/08/2023|2023-09-09
01:01:…|
|             F|          23|8079220|                 546|
```

```
31/07/2023|2023-09-09 23:51:…|                 498|  01/08/2023|2023-09-09
01:01:…|
|             M|          20|7019979|                   306|
31/07/2023|2023-09-09 23:47:…|                 375|  01/08/2023|2023-09-09
01:01:…|
|             F|          30|2998797|                   065|
31/07/2023|2023-09-09 23:42:…|                 016|  01/08/2023|2023-09-09
01:01:…|
|             M|          33|2085892|                   221|
31/07/2023|2023-09-09 23:37:…|                 029|  01/08/2023|2023-09-09
01:01:…|
|             M|          30|5081708|                   394|
31/07/2023|2023-09-09 23:58:…|                 372|  01/08/2023|2023-09-09
01:01:…|
|             M|          29|4998803|                   375|
31/07/2023|2023-09-09 23:57:…|                 398|  01/08/2023|2023-09-09
01:01:…|
+--------------+------------+-------+--------------------+-----------+--------
------------+--------------------+-----------+--------------------+
only showing top 10 rows
```

[8]:
```
#Ver el numero de renglones, numero de columnas y las columnas del DF
print(f"Renglones {eb_df.count()}")
print(f"Columnas {len(eb_df.columns)}")
print(f"Columnas: {eb_df.columns}")
```

```
[Stage 3:================================================>         (5 + 1) / 6]

Renglones 6947149
Columnas 9
Columnas: ['Genero_Usuario', 'Edad_Usuario', 'Bici', 'Ciclo_Estacion_Retiro',
'Fecha_Retiro', 'Hora_Retiro', 'Ciclo_EstacionArribo', 'Fecha Arribo',
'Hora_Arribo']
```

[9]:
```
##Crea un nuevo DF con solo 4 columnas
sub_eb_df = eb_df.select(['Genero_Usuario', 'Edad_Usuario', 'Bici',␣
 ↪'Fecha_Retiro'])
sub_eb_df.show(10)
print(f"Renglones {sub_eb_df.count()}")
```

```
+--------------+------------+-------+------------+
|Genero_Usuario|Edad_Usuario|   Bici|Fecha_Retiro|
+--------------+------------+-------+------------+
|             M|          24|2252039|  31/07/2023|
|             M|          33|8626897|  31/07/2023|
|             M|          34|4940557|  31/07/2023|
|             F|          30|2036523|  31/07/2023|
```

```
|             F|          23|8079220|  31/07/2023|
|             M|          20|7019979|  31/07/2023|
|             F|          30|2998797|  31/07/2023|
|             M|          33|2085892|  31/07/2023|
|             M|          30|5081708|  31/07/2023|
|             M|          29|4998803|  31/07/2023|
+--------------+------------+-------+------------+
only showing top 10 rows


[Stage 7:================================================>         (5 + 1) / 6]

Renglones 6947149
```

[10]:
```
##Generar un nuevo DF quitando duplicados
sub_nd_eb_df = sub_eb_df.dropDuplicates()
print(f"Renglones {sub_nd_eb_df.count()}")
```

```
[Stage 12:=============================================>             (3 + 1) / 4]

Renglones 6010044
```

[11]:
```
##Ver estadisticos basicos
sub_nd_eb_df.describe().show()
```

```
[Stage 18:=============================================>             (3 + 1) / 4]

+-------+--------------+------------------+------------------+------------+
|summary|Genero_Usuario|      Edad_Usuario|              Bici|Fecha_Retiro|
+-------+--------------+------------------+------------------+------------+
|  count|       6010044|           6010044|           6010044|     6010044|
|   mean|          null| 34.62895452592612| 5471698.783170473|        null|
| stddev|          null|10.010392880603918|2026433.8338036423|        null|
|    min|             ?|               100|           2001188|  01/01/2023|
|    max|             0|              NULL|           8998171|  31/12/2022|
+-------+--------------+------------------+------------------+------------+
```

[12]:
```
sub_nd_eb_df.write.save("gs://dcd05-sir-bucket/dcd/pyspark/ecobici/",␣
↪format='csv', header=True)
```

[ ]:
```

```