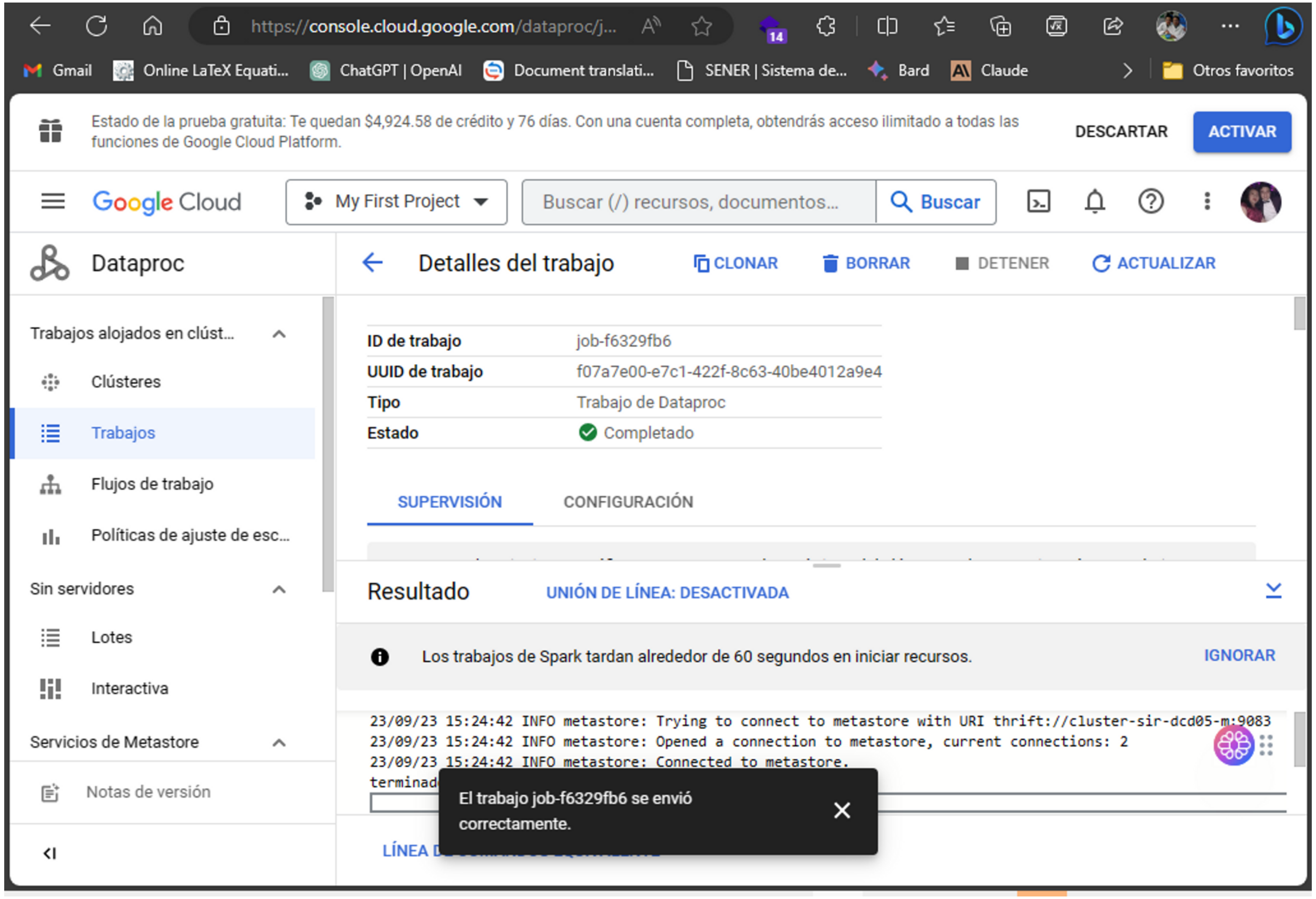
Diplomado en Ciencia de Datos UNAM
Módulo 13 Datos Masivos
Sesión 3 del 23 Septiembre de 2023

Sergio Ibarra Ramírez

Pantalla de Job de pysparkSQL ejecutado



Estado de la prueba gratuita: Te quedan $4,924.58 de crédito y 76 días. Con una cuenta completa, obtendrás acceso ilimitado a todas las funciones de Google Cloud Platform.

Detalles del trabajo

ID de trabajo	job-f6329fb6
UUID de trabajo	f07a7e00-e7c1-422f-8c63-40be4012a9e4
Tipo	Trabajo de Dataproc
Estado	✓ Completado

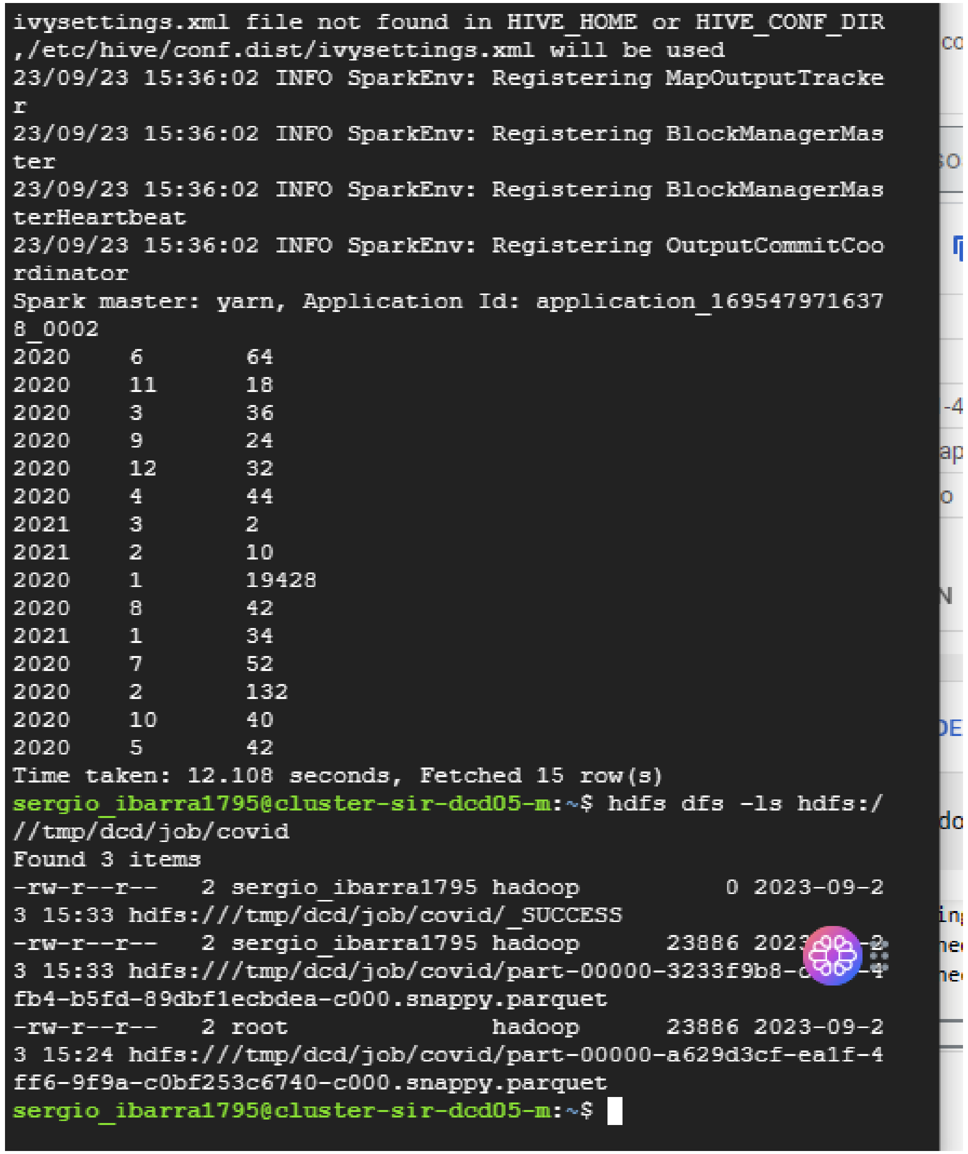
Resultado

Los trabajos de Spark tardan alrededor de 60 segundos en iniciar recursos.

```
23/09/23 15:24:42 INFO metastore: Trying to connect to metastore with URI thrift://cluster-sir-dcd05-m:9083
23/09/23 15:24:42 INFO metastore: Opened a connection to metastore, current connections: 2
23/09/23 15:24:42 INFO metastore: Connected to metastore.
terminad
```

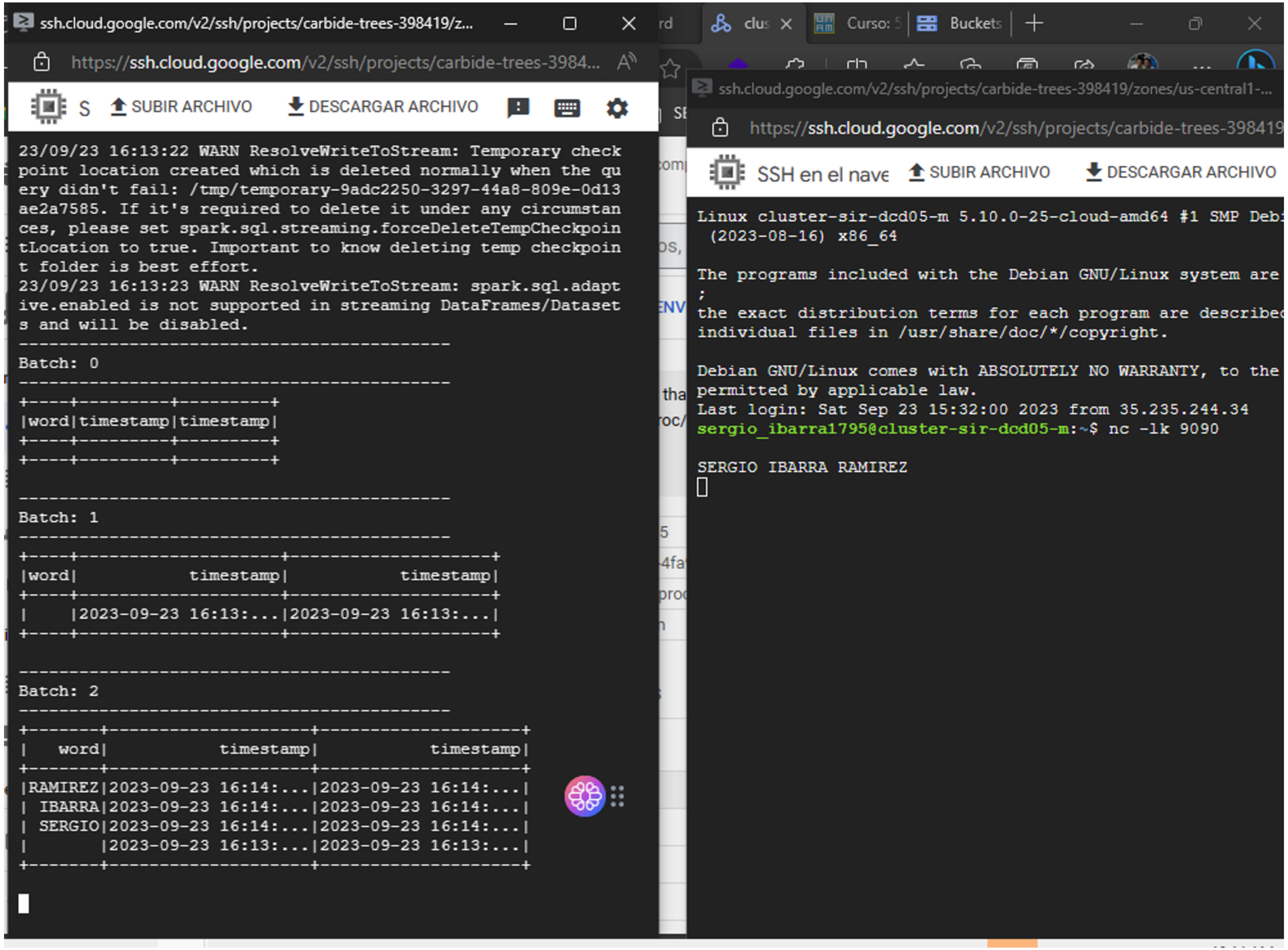
El trabajo job-f6329fb6 se envió correctamente.

Pantalla 2



```
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR
,/etc/hive/conf.dist/ivysettings.xml will be used
23/09/23 15:36:02 INFO SparkEnv: Registering MapOutputTracke
r
23/09/23 15:36:02 INFO SparkEnv: Registering BlockManagerMas
ter
23/09/23 15:36:02 INFO SparkEnv: Registering BlockManagerMas
terHeartbeat
23/09/23 15:36:02 INFO SparkEnv: Registering OutputCommitCoo
rdinator
Spark master: yarn, Application Id: application_169547971637
8_0002
2020    6       64
2020    11      18
2020    3       36
2020    9       24
2020    12      32
2020    4       44
2021    3       2
2021    2       10
2020    1       19428
2020    8       42
2021    1       34
2020    7       52
2020    2       132
2020    10      40
2020    5       42
Time taken: 12.108 seconds, Fetched 15 row(s)
sergio_ibarra1795@cluster-sir-dcd05-m:~$ hdfs dfs -ls hdfs:/
//tmp/dcd/job/covid
Found 3 items
-rw-r--r--   2 sergio_ibarra1795 hadoop          0 2023-09-2
3 15:33 hdfs:///tmp/dcd/job/covid/_SUCCESS
-rw-r--r--   2 sergio_ibarra1795 hadoop      23886 2023-09-2
3 15:33 hdfs:///tmp/dcd/job/covid/part-00000-3233f9b8-c[illegible]-4
fb4-b5fd-89dbf1ecbdea-c000.snappy.parquet
-rw-r--r--   2 root              hadoop      23886 2023-09-2
3 15:24 hdfs:///tmp/dcd/job/covid/part-00000-a629d3cf-ea1f-4
ff6-9f9a-c0bf253c6740-c000.snappy.parquet
sergio_ibarra1795@cluster-sir-dcd05-m:~$
```

Pantalla 3


ssh.cloud.google.com/v2/ssh/projects/carbide-trees-398419/z...
https://ssh.cloud.google.com/v2/ssh/projects/carbide-trees-3984...
S
SUBIR ARCHIVO
DESCARGAR ARCHIVO
clus
Curso: 5
Buckets
ssh.cloud.google.com/v2/ssh/projects/carbide-trees-398419/zones/us-central1-...
https://ssh.cloud.google.com/v2/ssh/projects/carbide-trees-398419
SSH en el nave
SUBIR ARCHIVO
DESCARGAR ARCHIVO

```
23/09/23 16:13:22 WARN ResolveWriteToStream: Temporary check
point location created which is deleted normally when the qu
ery didn't fail: /tmp/temporary-9adc2250-3297-44a8-809e-0d13
ae2a7585. If it's required to delete it under any circumstan
ces, please set spark.sql.streaming.forceDeleteTempCheckpoin
tLocation to true. Important to know deleting temp checkpoin
t folder is best effort.
23/09/23 16:13:23 WARN ResolveWriteToStream: spark.sql.adapt
ive.enabled is not supported in streaming DataFrames/Dataset
s and will be disabled.
-------------------------------------------
Batch: 0
-------------------------------------------
+----+---------+---------+
|word|timestamp|timestamp|
+----+---------+---------+
+----+---------+---------+

-------------------------------------------
Batch: 1
-------------------------------------------
+----+--------------------+--------------------+
|word|           timestamp|           timestamp|
+----+--------------------+--------------------+
|    |2023-09-23 16:13:...|2023-09-23 16:13:...|
+----+--------------------+--------------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+-------+--------------------+--------------------+
|   word|           timestamp|           timestamp|
+-------+--------------------+--------------------+
|RAMIREZ|2023-09-23 16:14:...|2023-09-23 16:14:...|
| IBARRA|2023-09-23 16:14:...|2023-09-23 16:14:...|
| SERGIO|2023-09-23 16:14:...|2023-09-23 16:14:...|
|       |2023-09-23 16:13:...|2023-09-23 16:13:...|
+-------+--------------------+--------------------+
```

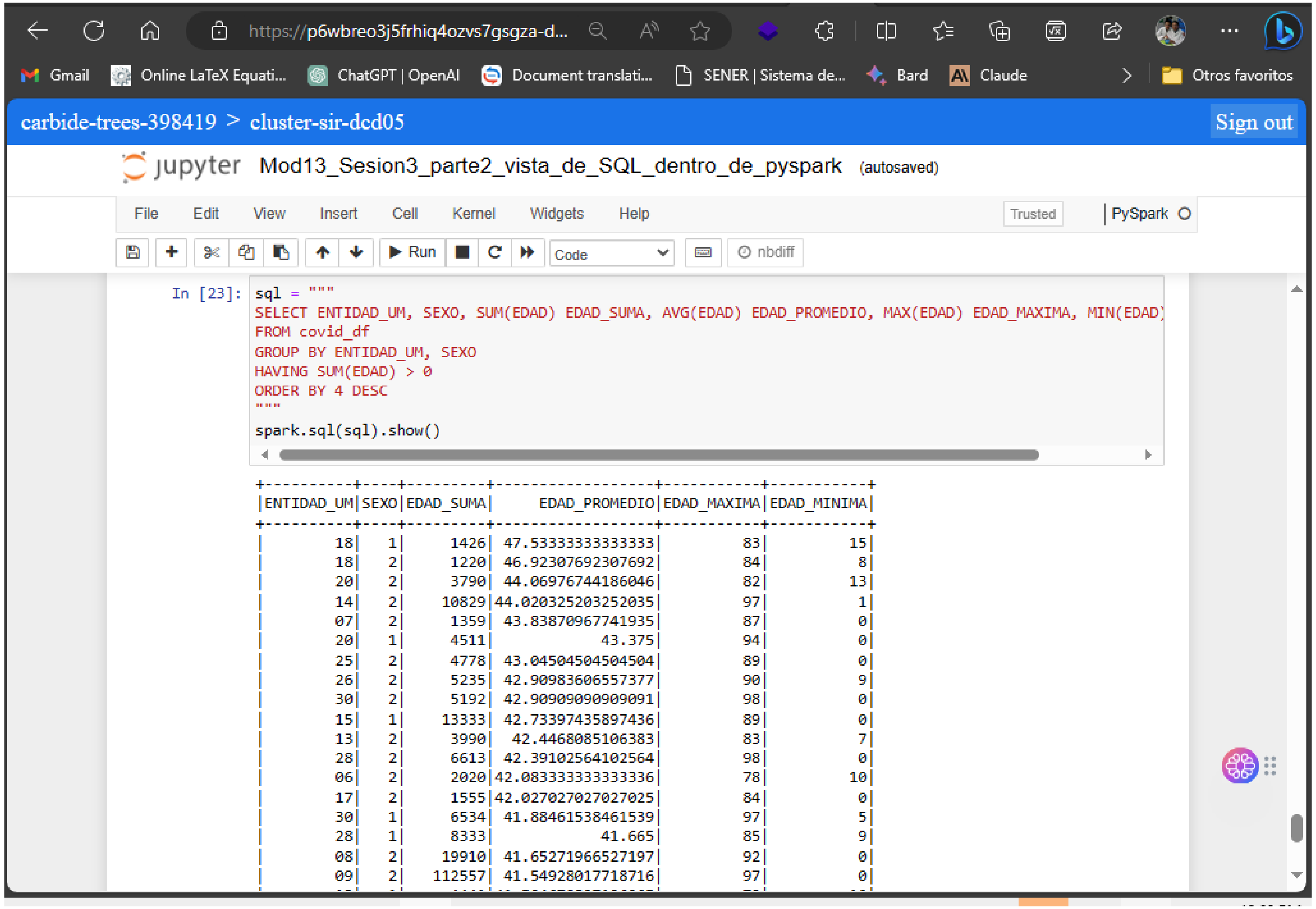
```
Linux cluster-sir-dcd05-m 5.10.0-25-cloud-amd64 #1 SMP Deb
 (2023-08-16) x86_64

The programs included with the Debian GNU/Linux system are
;
the exact distribution terms for each program are describe
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the
permitted by applicable law.
Last login: Sat Sep 23 15:32:00 2023 from 35.235.244.34
sergio_ibarra1795@cluster-sir-dcd05-m:~$ nc -lk 9090

SERGIO IBARRA RAMIREZ
```

Pantalla 4



```
sql = """
SELECT ENTIDAD_UM, SEXO, SUM(EDAD) EDAD_SUMA, AVG(EDAD) EDAD_PROMEDIO, MAX(EDAD) EDAD_MAXIMA, MIN(EDAD)
FROM covid_df
GROUP BY ENTIDAD_UM, SEXO
HAVING SUM(EDAD) > 0
ORDER BY 4 DESC
"""

spark.sql(sql).show()
```

```
+----------+----+---------+------------------+-----------+-----------+
|ENTIDAD_UM|SEXO|EDAD_SUMA|     EDAD_PROMEDIO|EDAD_MAXIMA|EDAD_MINIMA|
+----------+----+---------+------------------+-----------+-----------+
|        18|   1|     1426| 47.53333333333333|         83|         15|
|        18|   2|     1220| 46.92307692307692|         84|          8|
|        20|   2|     3790| 44.06976744186046|         82|         13|
|        14|   2|    10829|44.020325203252035|         97|          1|
|        07|   2|     1359| 43.83870967741935|         87|          0|
|        20|   1|     4511|            43.375|         94|          0|
|        25|   2|     4778| 43.04504504504504|         89|          0|
|        26|   2|     5235| 42.90983606557377|         90|          9|
|        30|   2|     5192| 42.90909090909091|         98|          0|
|        15|   1|    13333| 42.73397435897436|         89|          0|
|        13|   2|     3990|  42.4468085106383|         83|          7|
|        28|   2|     6613| 42.39102564102564|         98|          0|
|        06|   2|     2020|42.083333333333336|         78|         10|
|        17|   2|     1555|42.027027027027025|         84|          0|
|        30|   1|     6534| 41.88461538461539|         97|          5|
|        28|   1|     8333|            41.665|         85|          9|
|        08|   2|    19910| 41.65271966527197|         92|          0|
|        09|   2|   112557| 41.54928017718716|         97|          0|
```