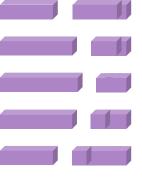


Diplomado en Ciencia de Datos UNAM Modulo 13

Sesión 2 del 22 Sept Uso de SQL Spark

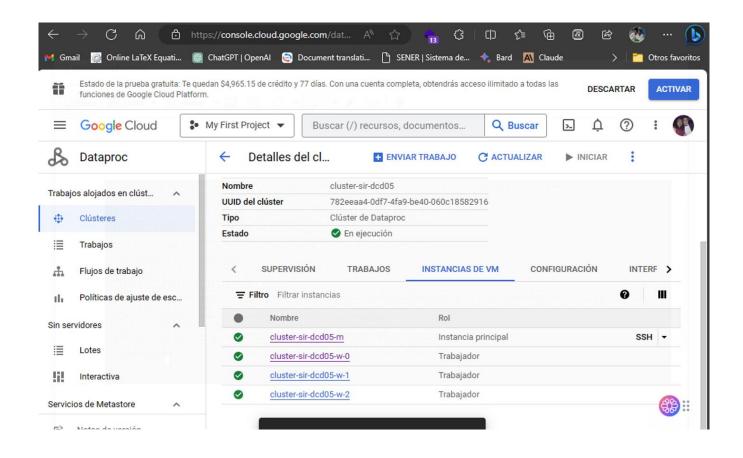
Alumno: Sergio Ibarra Ramírez



### Contenido

- + 1. Cluster con 3 workers
- + 2. Resumen de tareas de escritura (insert into)
- + 3. Visualización de las tablas covid\_avro y covid\_parquet
- + 4. Descripción extendida de la tabla sirilo (DESCRIBE FORMATTED)

#### Cluster con 3 workers



# Resumen de tareas de escritura (insert into)

head: '/tmp/dcd/pyspark/csv/covid/part-/': No	such file or directory
sergio_ibarra1795@cluster-sir-dcd05-m:~\$ hdf	s dfs -ls -R /tmp/dcd/pyspark/csv/covid
-rwxr-xr-x 2 sergio ibarra1795 hadoop	245736 2023-09-23 00:57 /tmp/dcd/pyspark/csv/covid/part-00000-2a263
2d8-250c-4cea-9d20-5bdc436847c0-c000	
-rwxr-xr-x 2 sergio ibarra1795 hadoop	66725 2023-09-23 00:57 /tmp/dcd/pyspark/csv/covid/part-00001-2a263
2d8-250c-4cea-9d20-5bdc436847c0-c000	12/
-rwxr-xr-x 2 sergio ibarra1795 hadoop	66931 2023-09-23 00:57 /tmp/dcd/pyspark/csv/covid/part-00003
2d8-250c-4cea-9d20-5bdc436847c0-c000	₩## II
-rwxr-xr-x 2 sergio ibarra1795 hadoop	65994 2023-09-23 00:57 /tmp/dcd/pyspark/csv/covid/part-00005-2a263
2d8-250c-4cea-9d20-5bdc436847c0-c000	
-rwxr-xr-x 2 sergio ibarra1795 hadoop	46086 2023-09-23 00:57 /tmp/dcd/pyspark/csv/covid/part-00008-2a263
2d8-250c-4cea-9d20-5bdc436847c0-c000	
sergio ibarra1795@cluster-sir-dcd05-m:~\$	

_	_	OL		_			
1	4	32	1	33	2		
1	4	32	1	38	2		
1	4	32	2	29	2		
1	4	32	2	24	2		
1	4	32	2	31	2		
1	4	32	2	24	2		
1	4	32	2	29	2		
2	4	32	1	46	2		
2	4	32	1	42	2		
2	4	10	1	60	2		
1	4	15	2	76	2		
2	9	13	1	72	2		
1	3	15	2	43	2		
1	3	15	1	60	2		
1	3	15	1	54	2		
1	3	15	2	40	2		
1	3	15	2	59	2		
1	3	15	1	51	2		
1	3	15	1	83	2		
1	3	15	2	57	2	į	
2	4	08	1	37	2	1	
2	4	09	1	33	2		
2	9	09	2	51	2		
2	9	09	1	57	2		
Time taken: 7.637 seconds, Fetched 33574 row(s)							
spark-sql>							

### Visualización de las tablas covid\_avro y covid\_parquet

```
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Fri Sep 22 22:36:02 2023 from 35.235.244.32
sergio ibarra1795@cluster-sir-dcd05-m:~$ hdfs dfs -ls -R hdfs:///tmp/dcd/pyspark/covid/
drwxr-xr-x - sergio ibarra1795 hadoop
                                               0 2023-09-22 22:53 hdfs:///tmp/dcd/pyspark/covid/avro
                                               0 2023-09-22 22:53 hdfs:///tmp/dcd/pyspark/covid/avro/ SUCCESS
-rw-r--r-- 2 sergio ibarra1795 hadoop
                                            89066 2023-09-22 22:53 hdfs://tmp/dcd/pyspark/covid/avro/part-000
-rw-r--r-- 2 sergio ibarra1795 hadoop
00-291f4851-b5b9-48f2-b6a2-bc3415e68cee-c000.avro
drwxr-xr-x - sergio ibarra1795 hadoop
                                               0 2023-09-22 22:56 hdfs:///tmp/dcd/pyspark/covid/csv
-rw-r--r-- 2 sergio ibarra1795 hadoop
                                                0 2023-09-22 22:56 hdfs:///tmp/dcd/pyspark/covid/csv/ SUCCESS
-rw-r--r-- 2 sergio ibarra1795 hadoop
                                           245736 2023-09-22 22:56 hdfs:///tmp/dcd/pyspark/covid/csv/part-0000
0-59062acd-bc5d-49e7-adbc-5fdc124c1d23-c000.csv
drwxr-xr-x - sergio ibarra1795 hadoop
                                                0 2023-09-22 22:53 hdfs:///tmp/dcd/pyspark/covid/parquet
                                                0 2023-09-22 22:53 hdfs://tmp/dcd/pyspark/covid/parquet/ SUCC
-rw-r--r-- 2 sergio ibarra1795 hadoop
ESS
-rw-r--r-- 2 sergio ibarra1795 hadoop
                                           38787 2023-09-22 22:53 hdfs:///tmp/dcd/pyspark/covid/parquet/part-
00000-ec6d91bb-ad7a-45cf-b3d2-3160eb89b57b-c000.snappy.parquet
sergio ibarra1795@cluster-sir-dcd05-m:~$
```

```
sergio_ibarra1795@cluster-sir-dcd05-m:~$ hdfs dfs -head /tmp/dcd/pyspark/covid/avro/part-00000-291f4851-b5b9-48
f2-b6a2-bc3415e68cee-c000.avro
Objavro.schema{"type":"record", "name":"topLevelRecord", "fields":[{"name":"ORIGEN", "type":["int", "null"]}, {"name
":"SECTOR", "type":["int", "null"]}, {"name":"ENTIDAD_UM", "type":["string", "null"]}, {"name":"SEXO", "type":["
null"]}, {"name":"EDAD", "type":["int", "null"]}, {"name":"EPOC", "type":["int", "null"]}]} Oorg.apache.spark.vel
":"3.3.2avro.codec
9.2<*rxj030f+K\*27X152Jz8q!DFq+:U22Z!3%09Gd0s,H!d`8=0!r%+9!seI!e,nVI!,G9%e:OArz:s!=~f|:%U;</pre>
```

## Descripción extendida de la tabla sirilo (DESCRIBE FORMATTED)

