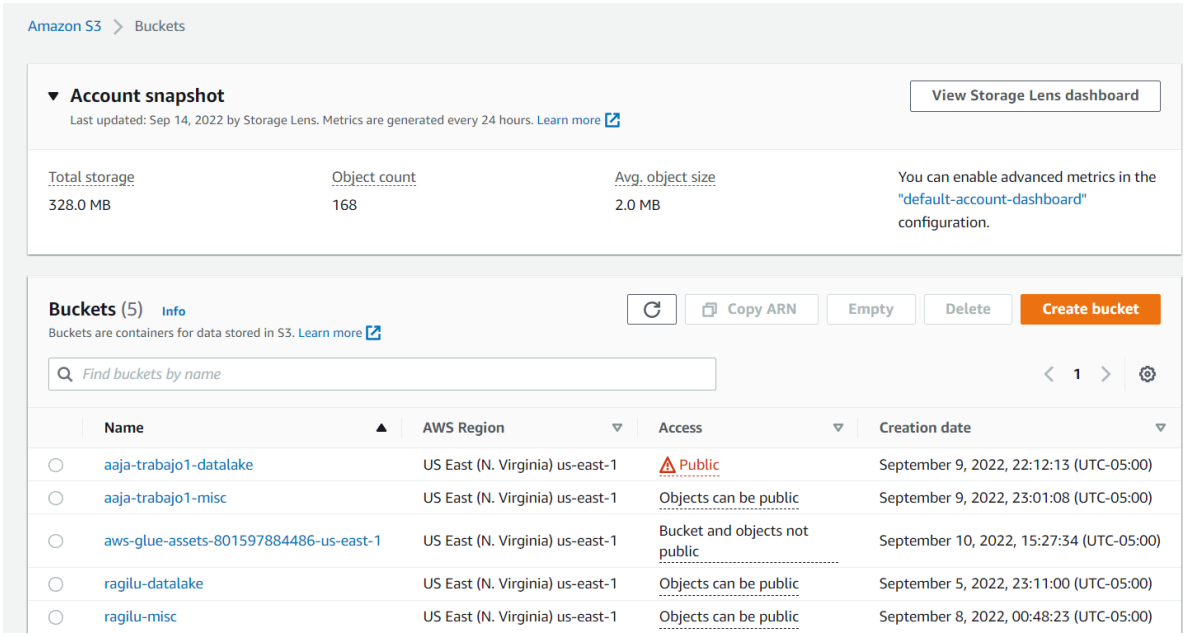
Almacenamiento y recuperación de información

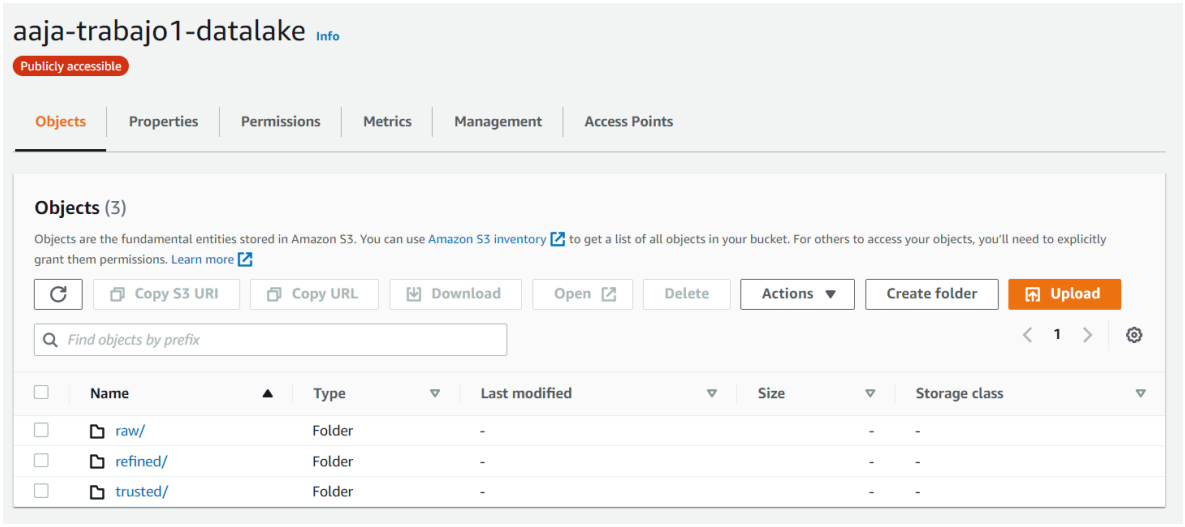
Trabajo 1 AWS-HADOOP-SPARK

Integrantes: Alejandro Arboleda Giron, Jhonatan Montes, Agustin Nieto, Rafael Alejandro Gil.

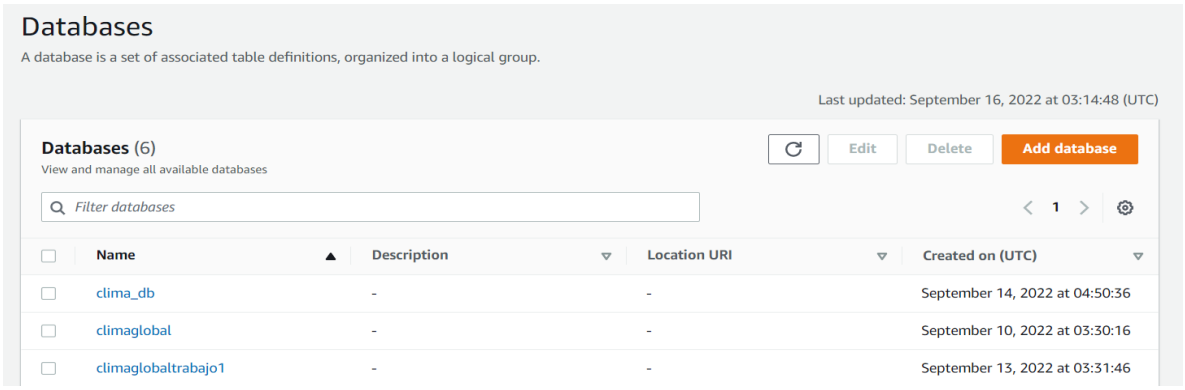
1. Creación del bucket y datalake



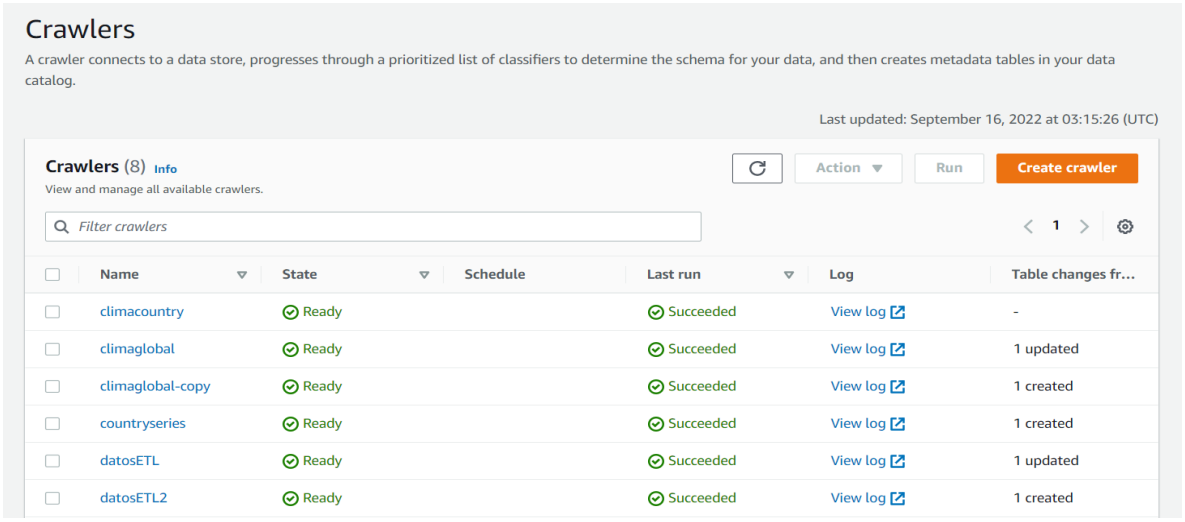
Creación de las zonas y subida de datos a la zona raw



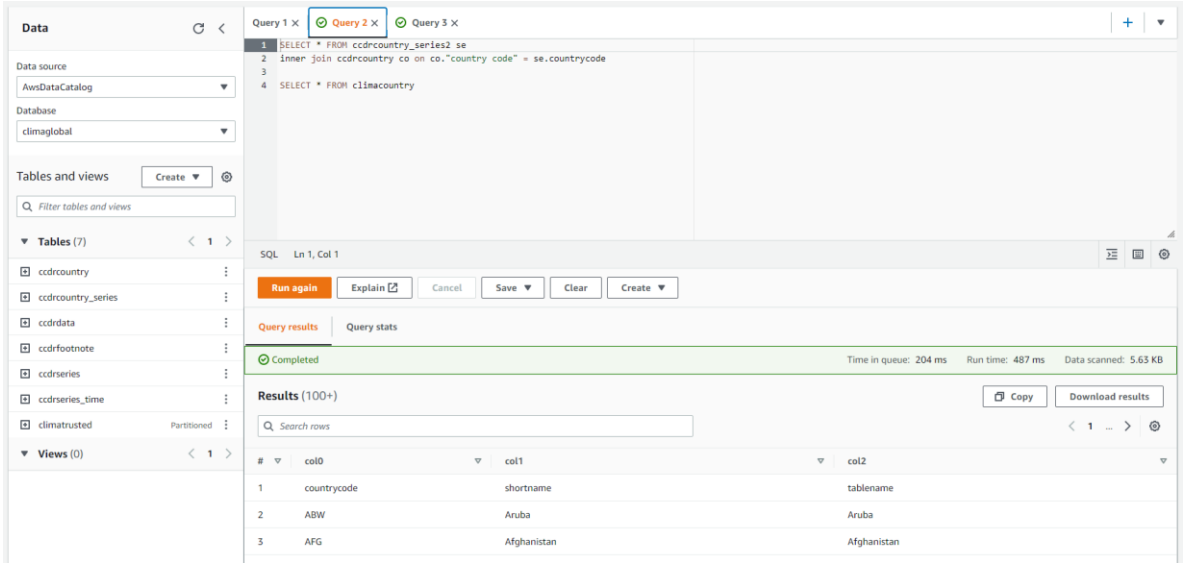
2. Creación de las bases de datos en Glue



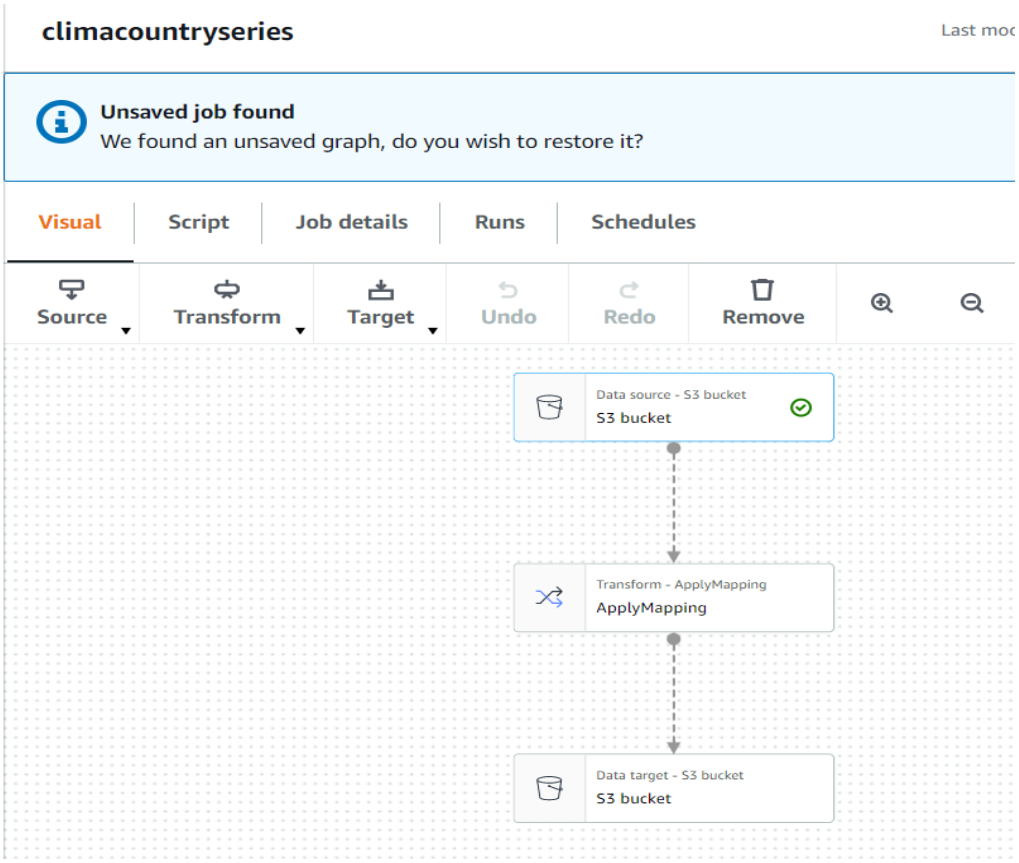
Catalogación a partir del Crawler en Glue



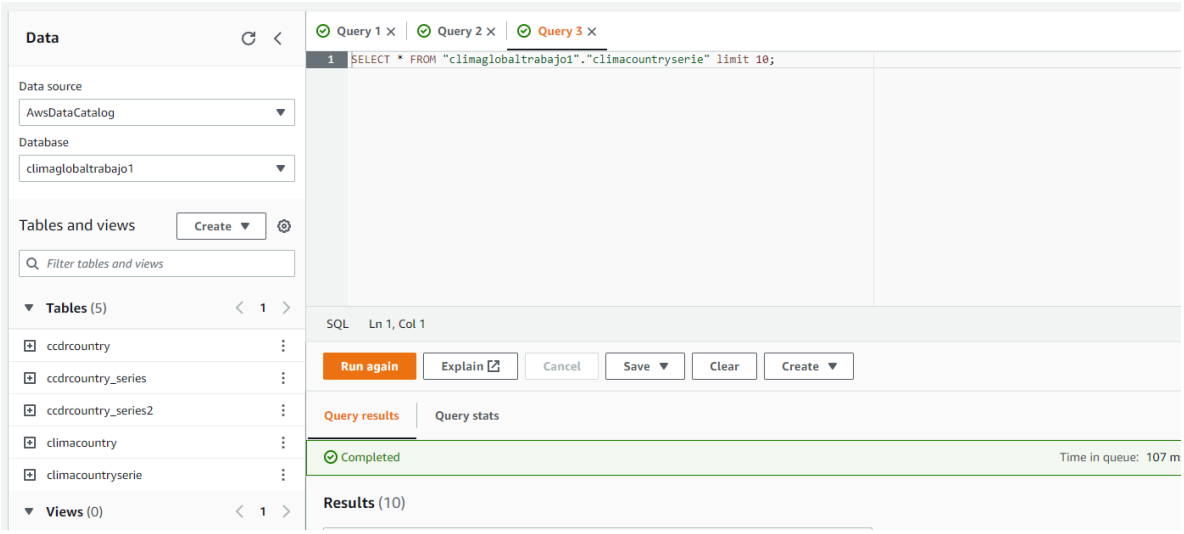
Lectura de datos en athena sin procesar



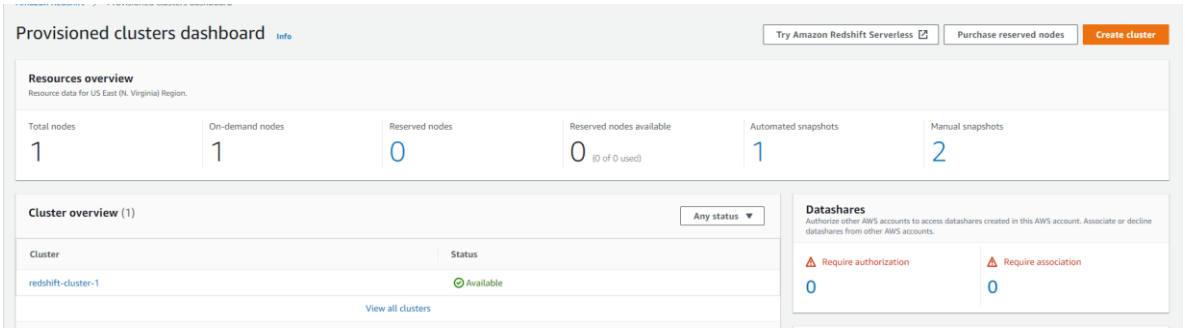
3. Creación ETL para renombrar columnas



Lectura de datos procesados en Athena



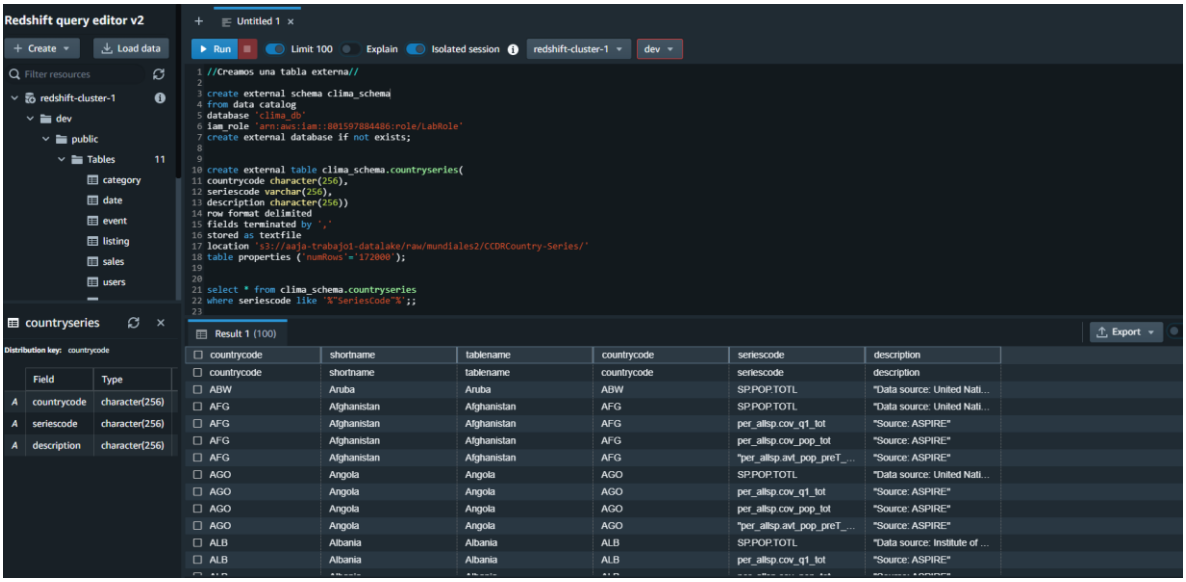
4. Creación cluster redshift para generar data-werehouse



Consumo de datos de datalake para generar el lakehouse

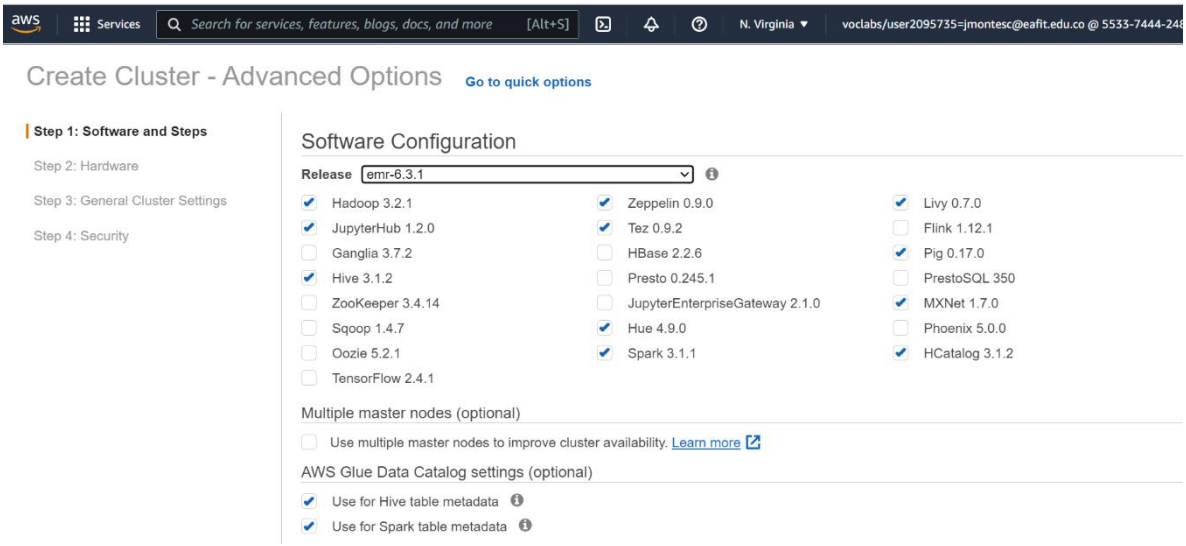
```
//Creamos una tabla externa//

create external schema clima_schema
from data catalog
database 'clima_db'
iam_role 'arn:aws:iam::801597884486:role/LabRole'
create external database if not exists;


create external table clima_schema.countryseries(
countrycode character(256),
seriescode varchar(256),
description character(256))
row format delimited
fields terminated by ','
stored as textfile
location 's3://aaja-trabajo1-datalake/raw/mundiales2/CCDRCountry-Series/'
table properties ('numRows'='172000');


select * from clima_schema.countryseries
where seriescode like '%"SeriesCode"%';;



//Creamos tabla nativa//

Create table ccdrSeriesTime(
SeriesCode varchar(200) distkey,
Year varchar(200) not null,
description varchar(200) not null);



COPY ccdrSeriesTime FROM 's3://aaja-trabajo1-datalake/raw/mundiales2/CCDRSeries-Time/CCDRSeries-Time.csv'
iam_role 'arn:aws:iam::801597884486:role/LabRole'
delimiter ',' timeformat 'YYYY-MM-DD HH:MI:SS' region 'us-east-1';


Create table countryseries(
countrycode VARCHAR(256) distkey,
seriescode VARCHAR(256) not null,
description VARCHAR(256)not null);
```

Lectura y join entre tablas externas con redshift

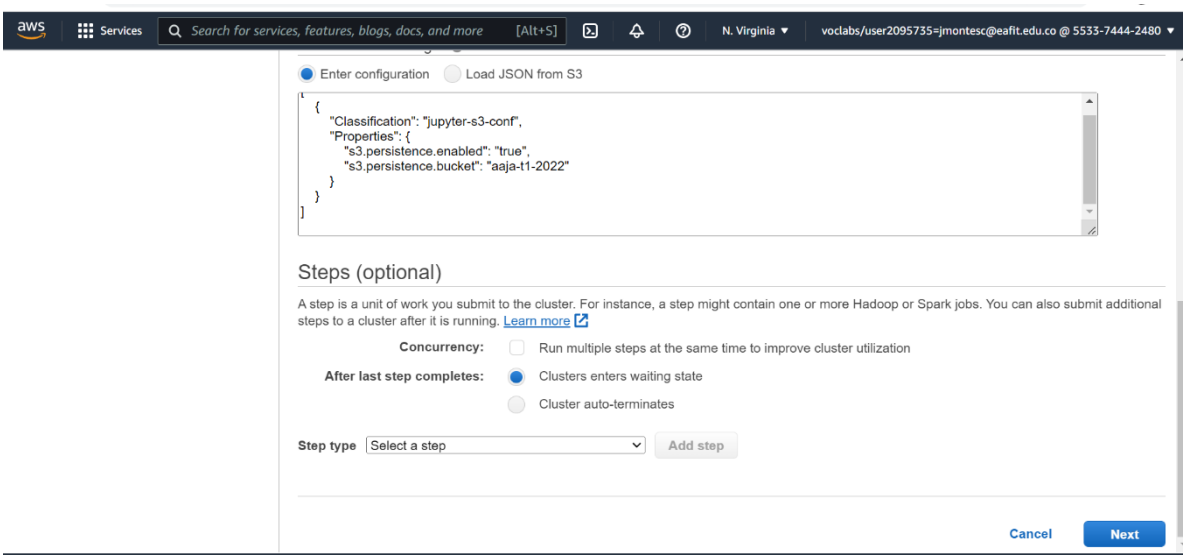


5. Clúster AWS EMR.

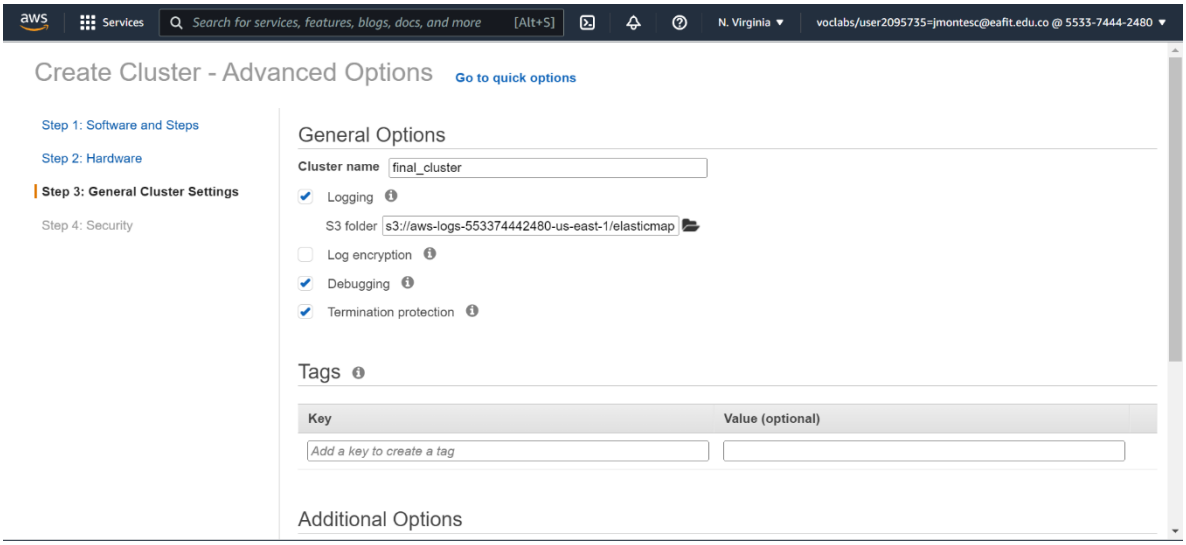
Para crear un clúster en Amazon AWS lo primero es llamar el servicio EMR (Amazon Elastic MapReduce). Se debe ingresar en configuraciones avanzadas y de ahí seleccionar la versión EMR 6.3.1 seleccionar los siguientes servicios:



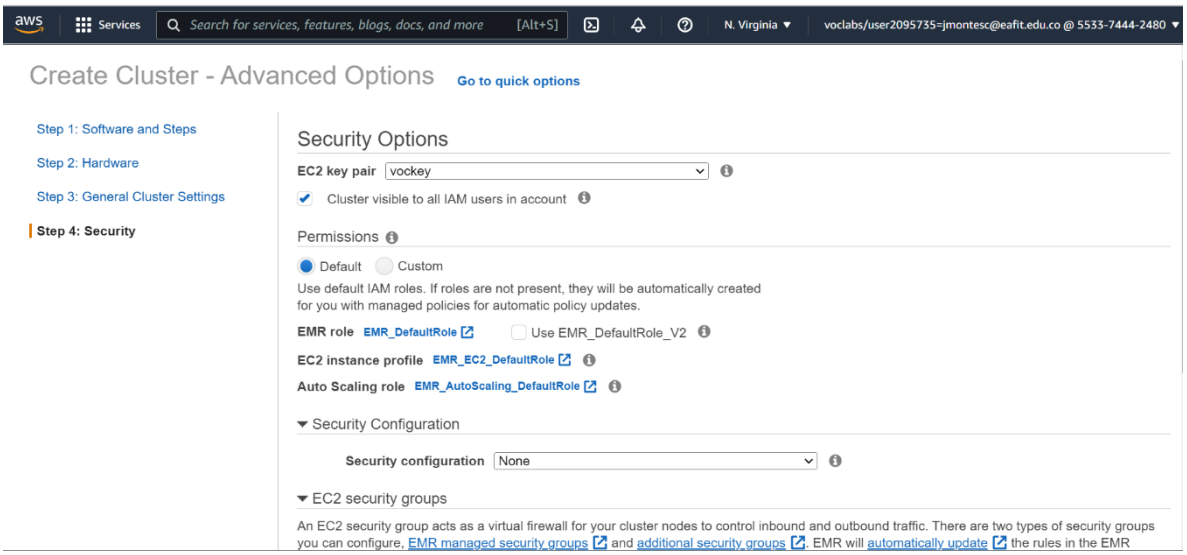
Luego escribimos el código de configuración escribiendo nuestro bucket de amazon de S3 en donde se guardarán nuestros notebooks de jupyter.

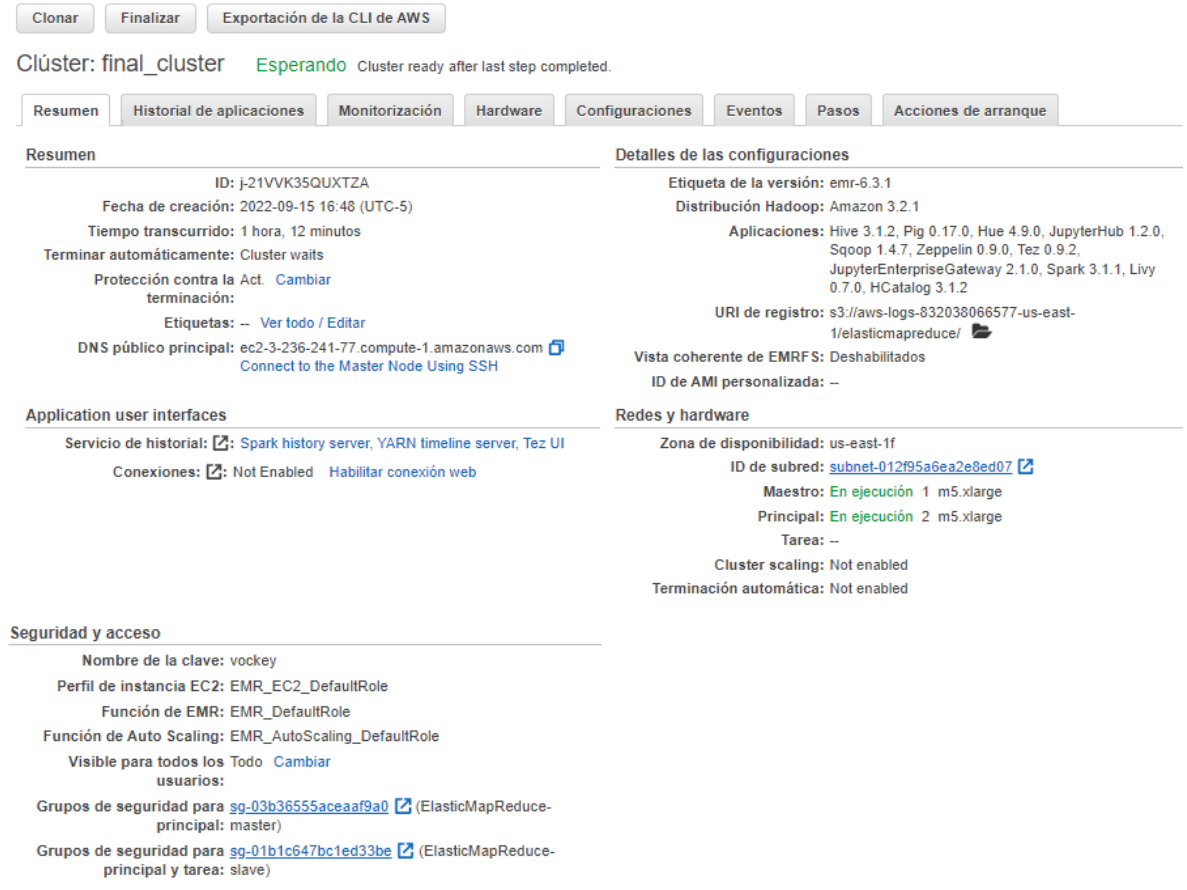


Escribimos el nombre de nuestro clúster

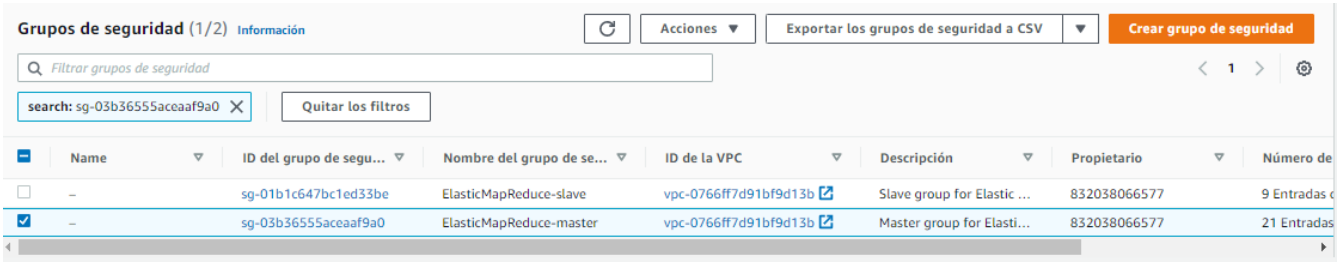


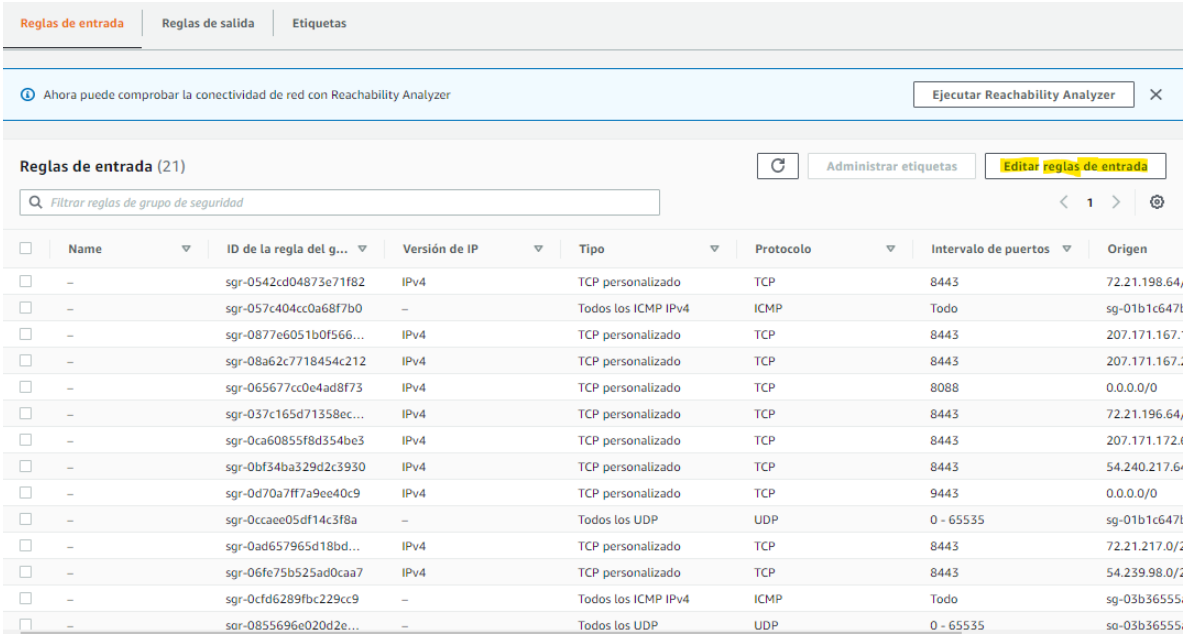
Y por último, antes de crear el clúster, seleccionamos la llave vockey que usamos para conectarnos a las máquinas. Finalmente damos click en crear el clúster.

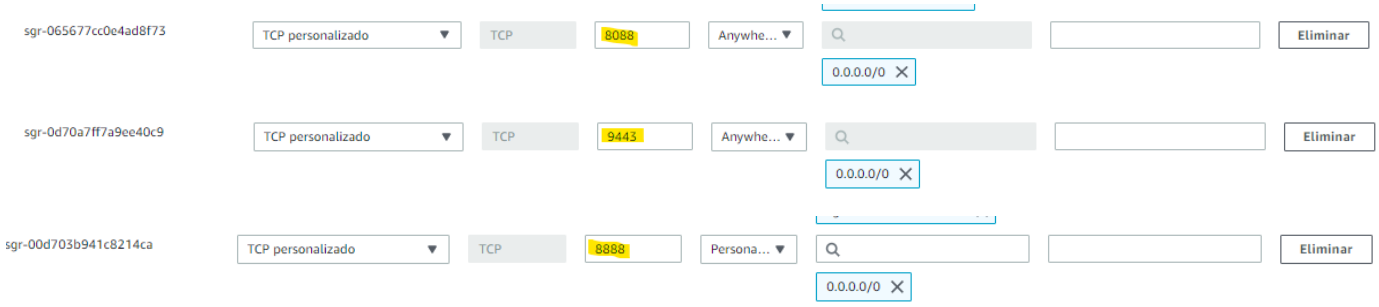




Configuración de reglas de entrada: ingreso a grupos de seguridad en el master y adición de puertos 8888, 9443 y 8088.







6. Conexión Apache

Para conectarse con servicios como Hue, Zepelling, spark, etc, se debe dirigir a la sección de historial de aplicaciones del cluster, en esta sección se deben haber generado links que envían a los servicios antes mencionados.

Clúster: final_cluster Comenzando Configuring cluster software

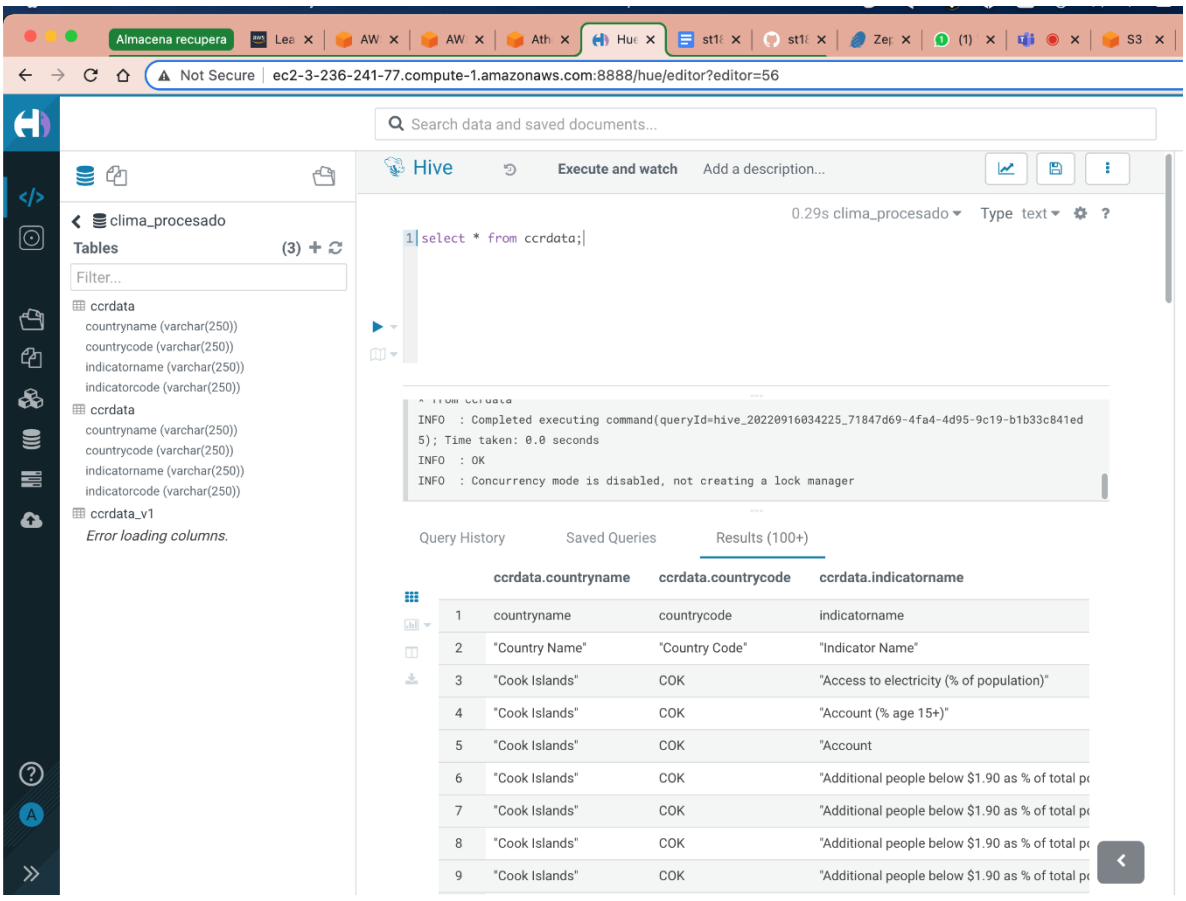
Resumen | Historial de aplicaciones | Monitorización | Hardware | Configuraciones | Eventos | Pasos | Acciones de arranque

On-cluster application user interfaces

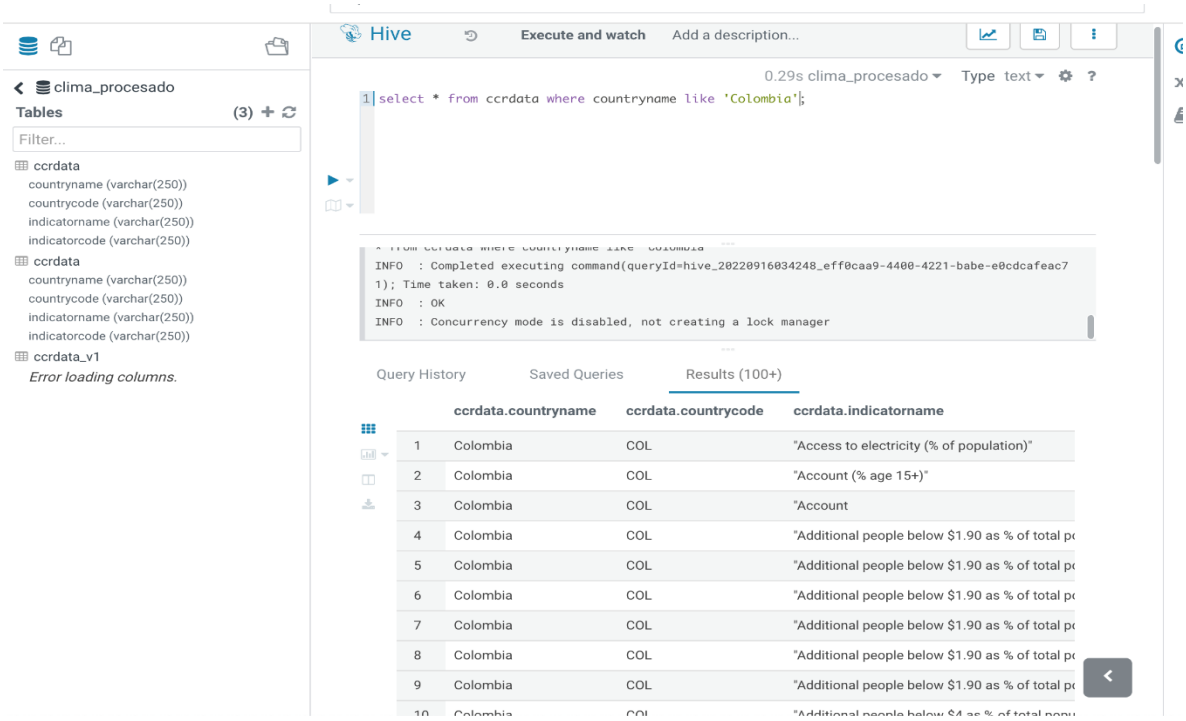
On-cluster UI are available only while clusters are running. Because they are hosted on the master node, on-cluster UI require a connection via SSH tunneling. Set up SSH tunneling before accessing these application UI. Learn more

Application	User interface URL	Status
Nodo del nombre de HDFS	http://ec2-44-200-104-33.compute-1.amazonaws.com:9870/	SSH tunnel not enabled
Tonalidad	http://ec2-44-200-104-33.compute-1.amazonaws.com:8888/	SSH tunnel not enabled
JupyterHub	https://ec2-44-200-104-33.compute-1.amazonaws.com:9443/	SSH tunnel not enabled
Zeppelin	http://ec2-44-200-104-33.compute-1.amazonaws.com:8890/	SSH tunnel not enabled
Tez UI	http://ec2-44-200-104-33.compute-1.amazonaws.com:8080/tez-ui	SSH tunnel not enabled
Spark History Server	http://ec2-44-200-104-33.compute-1.amazonaws.com:18080/	SSH tunnel not enabled
Livy	http://ec2-44-200-104-33.compute-1.amazonaws.com:8998/	SSH tunnel not enabled
Administrador de recursos	http://ec2-44-200-104-33.compute-1.amazonaws.com:8088/	SSH tunnel not enabled

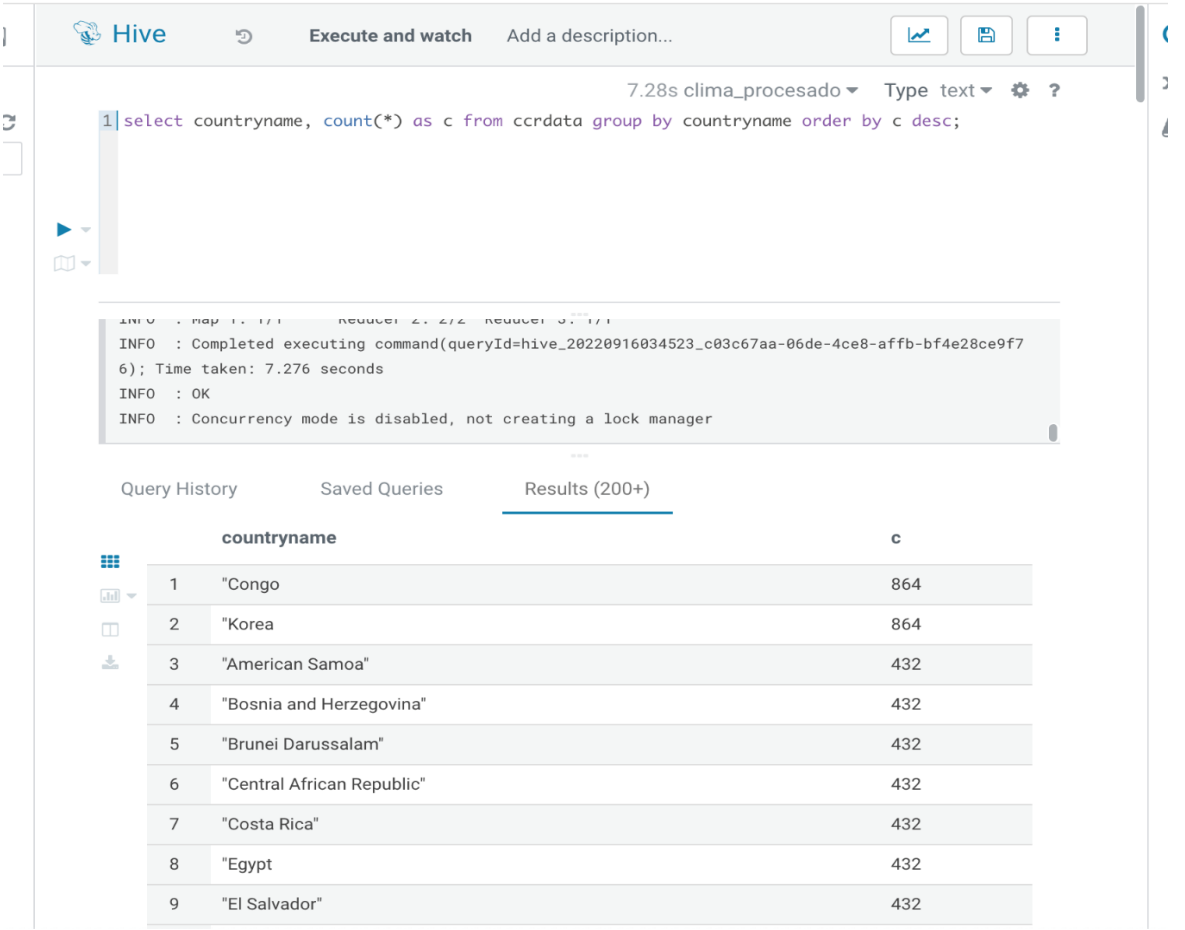
En Hive realizan consultas a la base de datos con los pasos de conexión del cluster a este servicio.



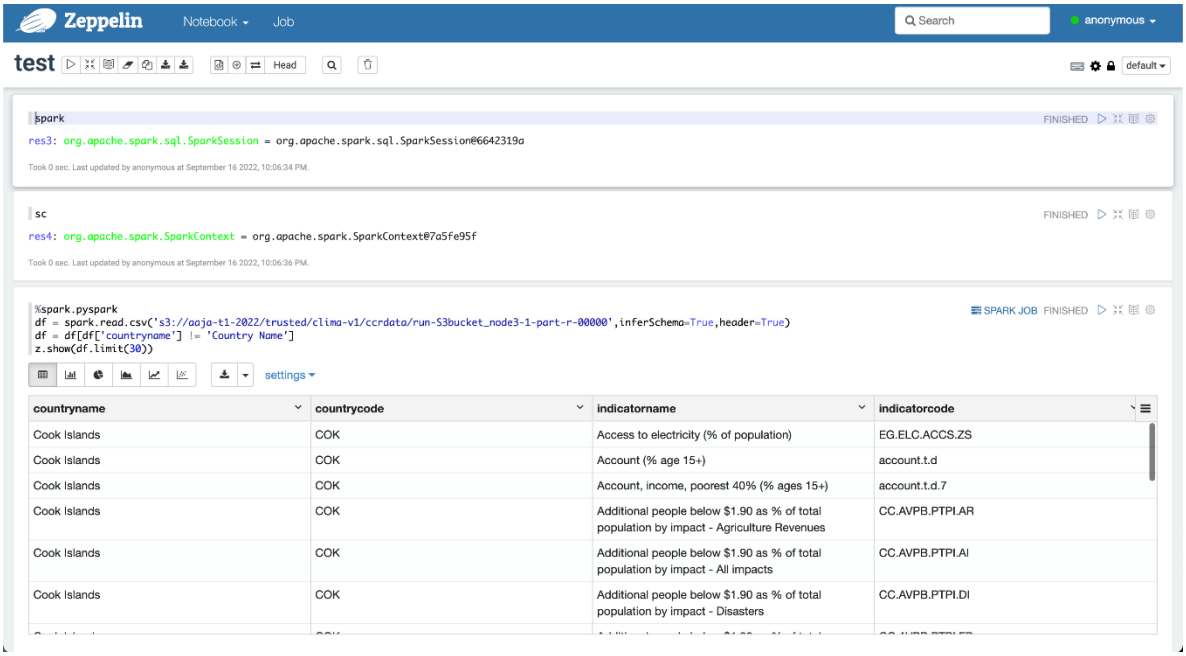
Query en Hive conectado a la base de datos.



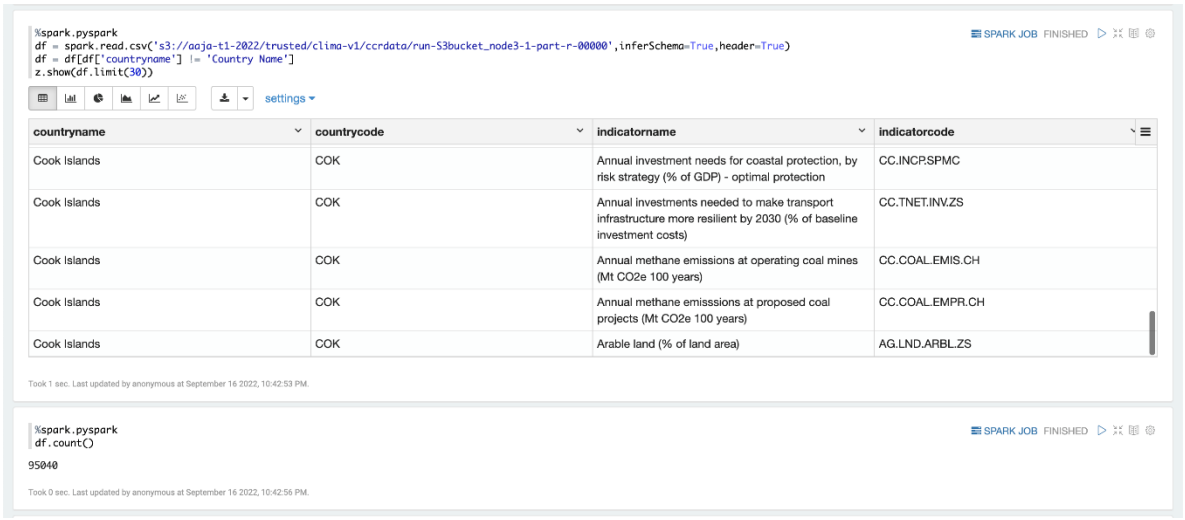
Search data and saved documents...



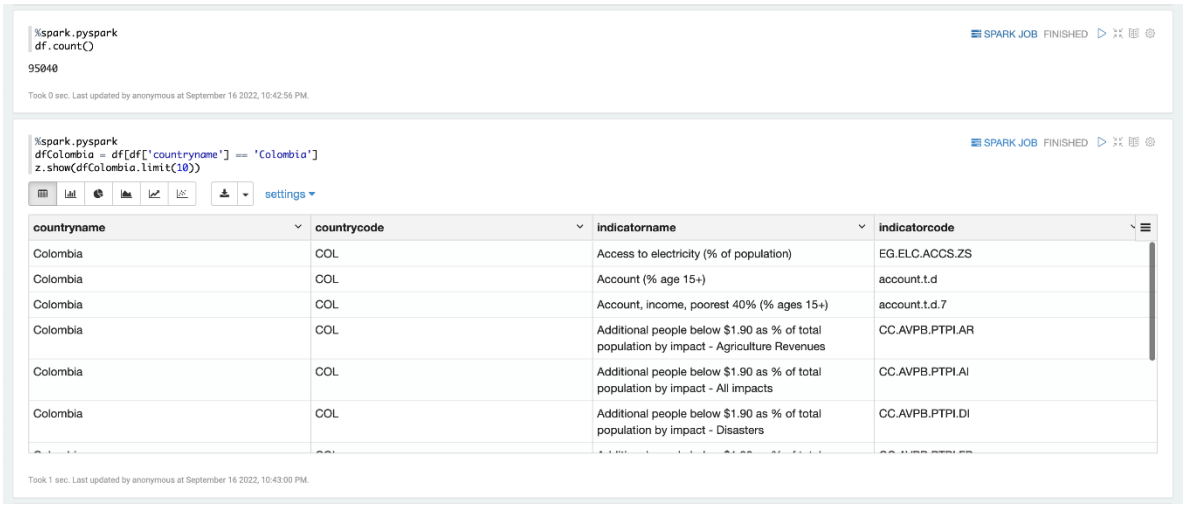
7. Scripts en Zeppelin, comprobación de Spark y cague de DF



Eliminamos la primera fila que se agregó como header



Contamos la cantidad de registros y filtramos por país = colombia



```
%spark.pyspark
df.count()
```

SPARK JOB FINISHED

```
95040
```

Took 0 sec. Last updated by anonymous at September 16 2022, 10:42:56 PM.

```
%spark.pyspark
dfColombia = df[df['countryname'] == 'Colombia']
z.show(dfColombia.limit(10))
```

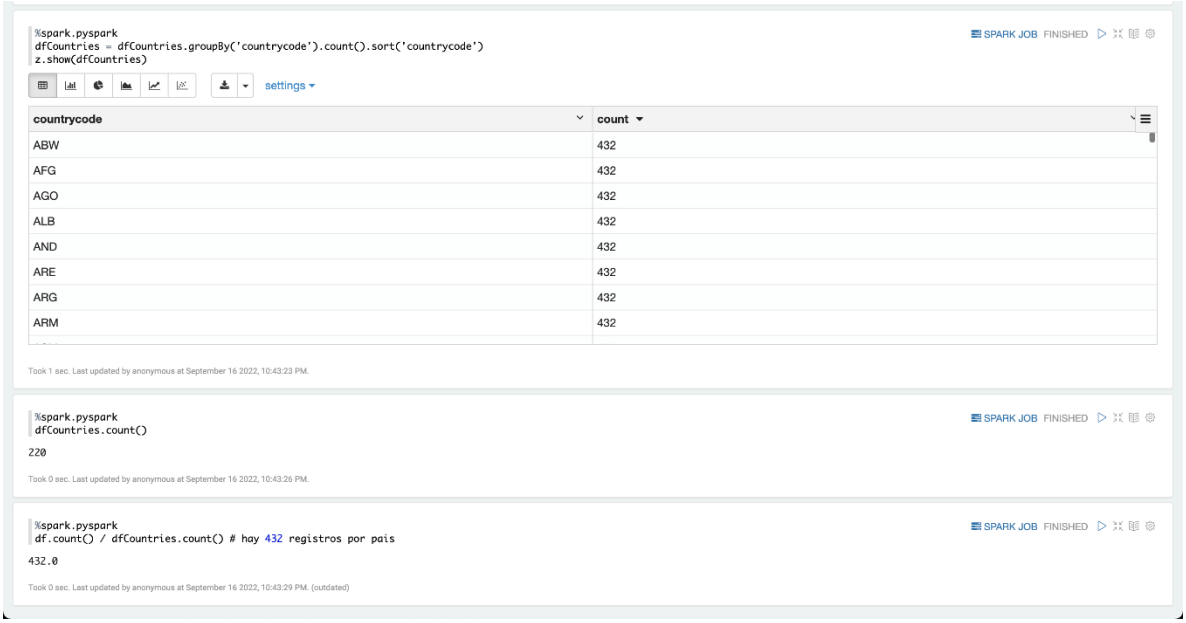
SPARK JOB FINISHED

settings

countryname	countrycode	indicatorname	indicatorcode
Colombia	COL	Access to electricity (% of population)	EG.ELC.ACCS.ZS
Colombia	COL	Account (% age 15+)	account.t.d
Colombia	COL	Account, income, poorest 40% (% ages 15+)	account.t.d.7
Colombia	COL	Additional people below $1.90 as % of total population by impact - Agriculture Revenues	CC.AVPB.PTPI.AR
Colombia	COL	Additional people below $1.90 as % of total population by impact - All impacts	CC.AVPB.PTPI.AI
Colombia	COL	Additional people below $1.90 as % of total population by impact - Disasters	CC.AVPB.PTPI.DI

Took 1 sec. Last updated by anonymous at September 16 2022, 10:43:00 PM.

Contamos la cantidad de registros por país



```
%spark.pyspark
dfCountries = dfCountries.groupBy('countrycode').count().sort('countrycode')
z.show(dfCountries)
```

SPARK JOB FINISHED

settings

countrycode	count
ABW	432
AFG	432
AGO	432
ALB	432
AND	432
ARE	432
ARG	432
ARM	432

Took 1 sec. Last updated by anonymous at September 16 2022, 10:43:23 PM.

```
%spark.pyspark
dfCountries.count()
```

SPARK JOB FINISHED

```
220
```

Took 0 sec. Last updated by anonymous at September 16 2022, 10:43:26 PM.

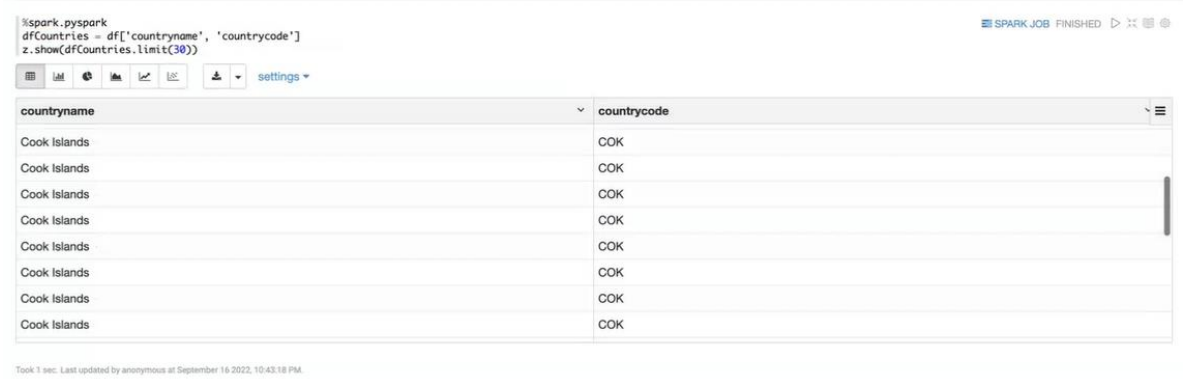
```
%spark.pyspark
df.count() / dfCountries.count() # hay 432 registros por pais
```

SPARK JOB FINISHED

```
432.0
```

Took 0 sec. Last updated by anonymous at September 16 2022, 10:43:29 PM. (outdated)

Segregamos el DF con las columnas countrycode



```
%spark.pyspark
dfCountries = df['countryname', 'countrycode']
z.show(dfCountries.limit(30))
```

SPARK JOB FINISHED

settings

countryname	countrycode
Cook Islands	COK
Cook Islands	COK
Cook Islands	COK
Cook Islands	COK
Cook Islands	COK
Cook Islands	COK
Cook Islands	COK
Cook Islands	COK

Took 1 sec. Last updated by anonymous at September 16 2022, 10:43:18 PM.