

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

[Amazon S3](#) > Buckets

▼ Account snapshot

Last updated: Sep 14, 2022 by Storage Lens. Metrics are generated every 24 hours. [Learn more](#)

Total storage328.0 MB

Object count168

Avg. object size2.0 MB

You can enable advanced metrics in the "default-account-dashboard" configuration.

[View Storage Lens dashboard](#)

Buckets (5) [Info](#)

[Refresh](#) [Copy ARN](#) [Empty](#) [Delete](#) [Create bucket](#)

[<](#) 1 [>](#) [Settings](#)

	Name ▲	AWS Region ▼	Access ▼	Creation date ▼
<input type="radio"/>	aaja-trabajo1-datalake	US East (N. Virginia) us-east-1	Public	September 9, 2022, 22:12:13 (UTC-05:00)
<input type="radio"/>	aaja-trabajo1-misc	US East (N. Virginia) us-east-1	Objects can be public	September 9, 2022, 23:01:08 (UTC-05:00)
<input type="radio"/>	aws-glue-assets-801597884486-us-east-1	US East (N. Virginia) us-east-1	Bucket and objects not public	September 10, 2022, 15:27:34 (UTC-05:00)
<input type="radio"/>	ragilu-datalake	US East (N. Virginia) us-east-1	Objects can be public	September 5, 2022, 23:11:00 (UTC-05:00)
<input type="radio"/>	ragilu-misc	US East (N. Virginia) us-east-1	Objects can be public	September 8, 2022, 00:48:23 (UTC-05:00)

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

aaja-trabajo1-datalake [Info](#)

[Publicly accessible](#)

[Objects](#) [Properties](#) [Permissions](#) [Metrics](#) [Management](#) [Access Points](#)

Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

[<](#) 1 [>](#) [Settings](#)

<input type="checkbox"/>	Name ▲	Type ▼	Last modified ▼	Size ▼	Storage class ▼
<input type="checkbox"/>	raw/	Folder	-	-	-
<input type="checkbox"/>	refined/	Folder	-	-	-
<input type="checkbox"/>	trusted/	Folder	-	-	-

2. Creación de las bases de datos en Glue

Databases

A database is a set of associated table definitions, organized into a logical group.

Last updated: September 16, 2022 at 03:14:48 (UTC)

Databases (6)

View and manage all available databases.

Filter databases

<input type="checkbox"/>	Name	Description	Location URI	Created on (UTC)
<input type="checkbox"/>	clima_db	-	-	September 14, 2022 at 04:50:36
<input type="checkbox"/>	climaglobal	-	-	September 10, 2022 at 03:30:16
<input type="checkbox"/>	climaglobaltrabajo1	-	-	September 13, 2022 at 03:31:46

Catalogación a partir del Crawler en Glue

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Last updated: September 16, 2022 at 03:15:26 (UTC)

Crawlers (8)

View and manage all available crawlers.

Filter crawlers

<input type="checkbox"/>	Name	State	Schedule	Last run	Log	Table changes fr...
<input type="checkbox"/>	climacountry	Ready		Succeeded	View log	-
<input type="checkbox"/>	climaglobal	Ready		Succeeded	View log	1 updated
<input type="checkbox"/>	climaglobal-copy	Ready		Succeeded	View log	1 created
<input type="checkbox"/>	countryseries	Ready		Succeeded	View log	1 created
<input type="checkbox"/>	datosETL	Ready		Succeeded	View log	1 updated
<input type="checkbox"/>	datosETL2	Ready		Succeeded	View log	1 created

Lectura de datos en athena sin procesar

Data source
AwsDataCatalog

Database
climaglobal

Tables and views
Create

Filter tables and views

Tables (7)
ccdrcountry
ccdrcountry_series
ccdrdata
ccdrfootnote
ccdrseries
ccdrseries_time
climatrustrusted
Views (0)

Query 1 x Query 2 x Query 3 x

```
1 SELECT * FROM ccdrcountry_series2 se
2 inner join ccdrcountry co on co."country code" = se.countrycode
3
4 SELECT * FROM climacountry
```

SQL Ln 1, Col 1

Run again Explain Cancel Save Clear Create

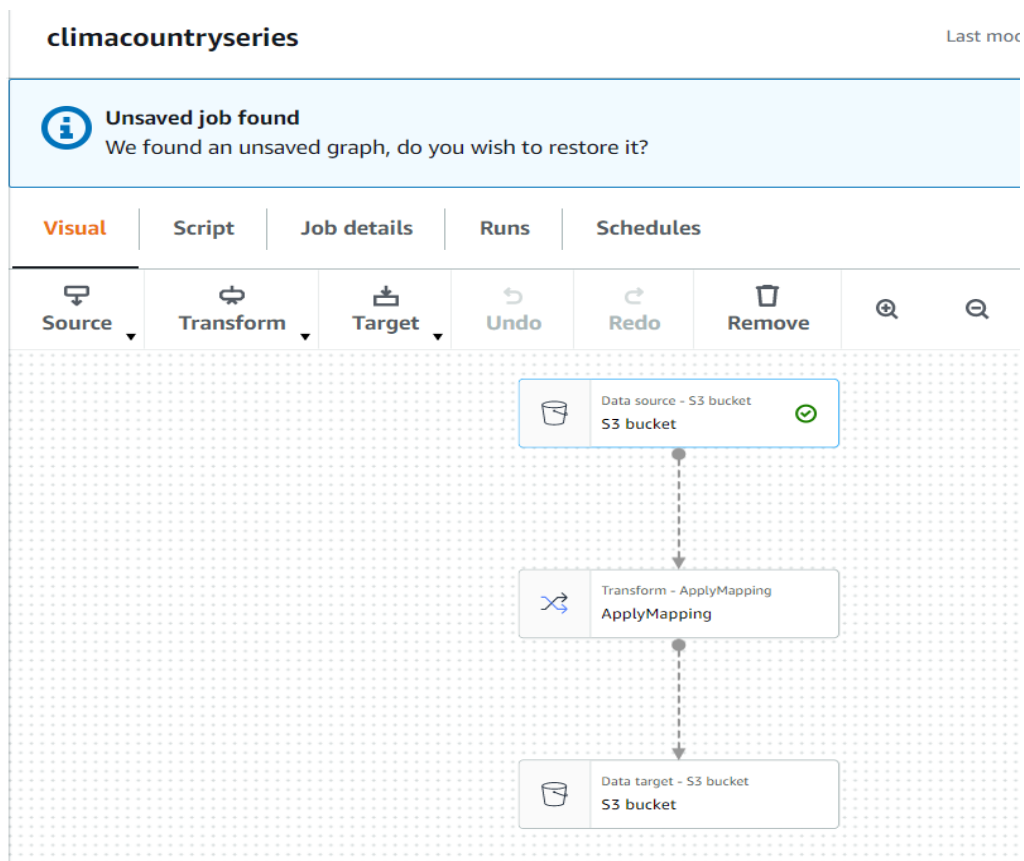
Query results Query stats

Completed Time in queue: 204 ms Run time: 487 ms Data scanned: 5.63 KB

Results (100+)
Search rows

#	col0	col1	col2
1	countrycode	shortname	tablename
2	ABW	Aruba	Aruba
3	AFG	Afghanistan	Afghanistan

3. Creación ETL para renombrar columnas



Lectura de datos procesados en Athena

Data 🔄 <

Query 1 × | Query 2 × | Query 3 ×

1 `SELECT * FROM "climaglobaltrabajo1"."climacountryserie" limit 10;`

Data source:

Database:

Tables and views: Create ⚙️

Tables (5) < **1** >

- ☒ ccdrcountry
- ☒ ccdrcountry_series
- ☒ ccdrcountry_series2
- ☒ climacountry
- ☒ climacountryserie

Views (0) < **1** >

SQL `Ln 1, Col 1`

Run again Explain Cancel Save Clear Create

Query results | Query stats

Completed Time in queue: 107 m

Results (10)

4. Creación cluster redshift para generar data-warehouse

Provisioned clusters dashboard [info](#)

[Try Amazon Redshift Serverless](#) [Purchase reserved nodes](#) [Create cluster](#)

Resources overview
Resource data for US East (N. Virginia) Region.

Total nodes	On-demand nodes	Reserved nodes	Reserved nodes available	Automated snapshots	Manual snapshots
1	1	0	0 (0 of 0 used)	1	2

Cluster overview (1) [Any status](#)

Cluster	Status
redshift-cluster-1	Available

[View all clusters](#)

Datashares
Authorize other AWS accounts to access datashares created in this AWS account. Associate or decline datashares from other AWS accounts.

Require authorization	Require association
0	0

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

The screenshot shows the Redshift query editor v2 interface. On the left, there's a sidebar with a tree view showing the database structure: 'redshift-cluster-1' > 'dev' > 'public' > 'Tables' > 'countryseries'. The main area displays a SQL query:

```
1 //Creamos una tabla externa//
2
3 create external schema clima_schema
4 from data catalog
5 database 'clima_db'
6 iam_role 'arn:aws:iam::88107788486:role/LabRole'
7 create external database if not exists;
8
9
10 create external table clima_schema.countryseries(
11 countrycode character(256),
12 seriescode varchar(256),
13 description character(256))
14 row format delimited
15 fields terminated by ','
16 stored as textfile
17 location 's3://asja-trabajo1-datalake/raw/mundiales2/CCDRCountry-Series/'
18 table properties ('numRows'='172000');
19
20
21 select * from clima_schema.countryseries
22 where seriescode like 'ASPIRE%';
23
```

Below the query, the results are shown in a table with 7 columns: countrycode, shortname, tablename, countrycode, seriescode, and description. The table contains 100 rows of data, including entries for Afghanistan, Angola, and Albania.

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:

The screenshot shows the AWS Management Console 'Create Cluster - Advanced Options' page. The 'Software Configuration' section is expanded, showing the 'Release' dropdown set to 'emr-6.3.1'. Below this, there are checkboxes for various software components:

- ☒ Hadoop 3.2.1
- ☒ JupyterHub 1.2.0
- ☐ Ganglia 3.7.2
- ☒ Hive 3.1.2
- ☐ ZooKeeper 3.4.14
- ☐ Sqoop 1.4.7
- ☐ Oozie 5.2.1
- ☐ TensorFlow 2.4.1
- ☒ Zeppelin 0.9.0
- ☐ Tez 0.9.2
- ☐ HBase 2.2.6
- ☐ Presto 0.245.1
- ☐ JupyterEnterpriseGateway 2.1.0
- ☒ Hue 4.9.0
- ☒ Spark 3.1.1
- ☒ Livy 0.7.0
- ☐ Flink 1.12.1
- ☒ Pig 0.17.0
- ☐ PrestoSQL 350
- ☒ MXNet 1.7.0
- ☐ Phoenix 5.0.0
- ☒ HCatalog 3.1.2

Below the software configuration, there are optional settings:

- ☐ Multiple master nodes (optional)
- ☐ Use multiple master nodes to improve cluster availability. [Learn more](#)
- ☐ AWS Glue Data Catalog settings (optional)
- ☒ Use for Hive table metadata
- ☒ Use for Spark table metadata

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

aws Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia voclabs/user2095735=jmontesc@eafit.edu.co @ 5533-7444-2480

Enter configuration Load JSON from S3

```
{
  "Classification": "jupyter-s3-conf",
  "Properties": {
    "s3.persistence.enabled": "true",
    "s3.persistence.bucket": "aaja-t1-2022"
  }
}
```

Steps (optional)

A step is a unit of work you submit to the cluster. For instance, a step might contain one or more Hadoop or Spark jobs. You can also submit additional steps to a cluster after it is running. [Learn more](#)

Concurrency: ☐ Run multiple steps at the same time to improve cluster utilization

After last step completes: ☒ Clusters enters waiting state ☐ Cluster auto-terminates

Step type

Escribimos el nombre de nuestro clúster

aws Services Search for services, features, blogs, docs, and more [Alt+S] N. Virginia voclabs/user2095735=jmontesc@eafit.edu.co @ 5533-7444-2480

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

General Options

Cluster name

☒ Logging ☐ Log encryption ☒ Debugging ☒ Termination protection

S3 folder

Tags

Key	Value (optional)
<input type="text" value="Add a key to create a tag"/>	<input type="text"/>

Additional Options

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.

Create Cluster - Advanced Options [Go to quick options](#)

- Step 1: Software and Steps
- Step 2: Hardware
- Step 3: General Cluster Settings
- Step 4: Security**

Security Options

EC2 key pair

☒ Cluster visible to all IAM users in account

Permissions

☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR_DefaultRole](#) ☐ Use EMR_DefaultRole_V2

EC2 instance profile [EMR_EC2_DefaultRole](#)

Auto Scaling role [EMR_AutoScaling_DefaultRole](#)

▼ Security Configuration

Security configuration

▼ EC2 security groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, [EMR managed security groups](#) and [additional security groups](#). EMR will [automatically update](#) the rules in the EMR

[Clonar](#)
[Finalizar](#)
[Exportación de la CLI de AWS](#)

Clúster: final_cluster **Esperando** Cluster ready after last step completed.

[Resumen](#)
[Historial de aplicaciones](#)
[Monitorización](#)
[Hardware](#)
[Configuraciones](#)
[Eventos](#)
[Pasos](#)
[Acciones de arranque](#)

Resumen

ID: j-21VVK35QUXTZA

Fecha de creación: 2022-09-15 16:48 (UTC-5)

Tiempo transcurrido: 1 hora, 12 minutos

Terminar automáticamente: Cluster waits

Protección contra la Act. [Cambiar](#) terminación:

Etiquetas: -- [Ver todo](#) / [Editar](#)

DNS público principal: ec2-3-236-241-77.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Application user interfaces

Servicio de historial: [Spark history server](#), [YARN timeline server](#), [Tez UI](#)

Conexiones: [Not Enabled](#) [Habilitar conexión web](#)

Seguridad y acceso

Nombre de la clave: vockey

Perfil de instancia EC2: EMR_EC2_DefaultRole

Función de EMR: EMR_DefaultRole

Función de Auto Scaling: EMR_AutoScaling_DefaultRole

Visible para todos los [Todo](#) [Cambiar](#) usuarios:

Grupos de seguridad para [sg-03b36555aceaaf9a0](#) (ElasticMapReduce-principal: master)

Grupos de seguridad para [sg-01b1c647bc1ed33be](#) (ElasticMapReduce-principal y tarea: slave)

Detalles de las configuraciones

Etiqueta de la versión: emr-6.3.1

Distribución Hadoop: Amazon 3.2.1

Aplicaciones: Hive 3.1.2, Pig 0.17.0, Hue 4.9.0, JupyterHub 1.2.0, Sqoop 1.4.7, Zeppelin 0.9.0, Tez 0.9.2, JupyterEnterpriseGateway 2.1.0, Spark 3.1.1, Livy 0.7.0, HCatalog 3.1.2

URI de registro: s3://aws-logs-832038066577-us-east-1/elasticmapreduce/

Vista coherente de EMRFS: Deshabilitados

ID de AMI personalizada: --

Redes y hardware

Zona de disponibilidad: us-east-1f

ID de subred: [subnet-012f95a6ea2e8ed07](#)

Maestro: **En ejecución** 1 m5.xlarge

Principal: **En ejecución** 2 m5.xlarge

Tarea: --

Cluster scaling: Not enabled

Terminación automática: Not enabled

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

Grupos de seguridad (1/2)

Información

Acciones

Exportar los grupos de seguridad a CSV

Crear grupo de seguridad

Filtrar grupos de seguridad

search: sg-03b36555aceaaf9a0

Quitar los filtros

1

	Name	ID del grupo de segu...	Nombre del grupo de se...	ID de la VPC	Descripción	Propietario	Número de
<input type="checkbox"/>	-	sg-01b1c647bc1ed33be	ElasticMapReduce-slave	vpc-0766ff7d91bf9d13b	Slave group for Elastic ...	832038066577	9 Entradas c
<input checked="" type="checkbox"/>	-	sg-03b36555aceaaf9a0	ElasticMapReduce-master	vpc-0766ff7d91bf9d13b	Master group for Elasti...	832038066577	21 Entradas

Reglas de entrada								
Reglas de salida								
Etiquetas								
<input type="button" value="Ejecutar Reachability Analyzer"/>								
<input type="button" value="Ejecutar Reachability Analyzer"/>								
Reglas de entrada (21)								
<input type="text" value="Filtrar reglas de grupo de seguridad"/>								
<input type="checkbox"/>	Name	ID de la regla del g...	Versión de IP	Tipo	Protocolo	Intervalo de puertos	Origen	
<input type="checkbox"/>	-	sgr-0542cd04873e71f82	IPv4	TCP personalizado	TCP	8443	72.21.198.64,	
<input type="checkbox"/>	-	sgr-057c404cc0a68f7b0	-	Todos los ICMP IPv4	ICMP	Todo	sg-01b1c647l	
<input type="checkbox"/>	-	sgr-0877e6051b0f566...	IPv4	TCP personalizado	TCP	8443	207.171.167.:	
<input type="checkbox"/>	-	sgr-08a62c7718454c212	IPv4	TCP personalizado	TCP	8443	207.171.167.:	
<input type="checkbox"/>	-	sgr-065677cc0e4ad8f73	IPv4	TCP personalizado	TCP	8088	0.0.0.0/0	
<input type="checkbox"/>	-	sgr-037c165d71358ec...	IPv4	TCP personalizado	TCP	8443	72.21.196.64,	
<input type="checkbox"/>	-	sgr-0ca60855f8d354be3	IPv4	TCP personalizado	TCP	8443	207.171.172.i	
<input type="checkbox"/>	-	sgr-0bf34ba329d2c3930	IPv4	TCP personalizado	TCP	8443	54.240.217.6,	
<input type="checkbox"/>	-	sgr-0d70a7ff7a9ee40c9	IPv4	TCP personalizado	TCP	9443	0.0.0.0/0	
<input type="checkbox"/>	-	sgr-0ccaee05df14c3f8a	-	Todos los UDP	UDP	0 - 65535	sg-01b1c647l	
<input type="checkbox"/>	-	sgr-0ad657965d18bd...	IPv4	TCP personalizado	TCP	8443	72.21.217.0/;	
<input type="checkbox"/>	-	sgr-06fe75b525ad0caa7	IPv4	TCP personalizado	TCP	8443	54.239.98.0/;	
<input type="checkbox"/>	-	sgr-0cfd6289fbc229cc9	-	Todos los ICMP IPv4	ICMP	Todo	sg-03b36555.	
<input type="checkbox"/>	-	sgr-0855696e020d2e...	-	Todos los UDP	UDP	0 - 65535	so-03b36555.	

sgr-065677cc0e4ad8f73	TCP personalizado	TCP	8088	Anywhe...	<input type="text" value="0.0.0.0/0"/>	<input type="button" value="Eliminar"/>
sgr-0d70a7ff7a9ee40c9	TCP personalizado	TCP	9443	Anywhe...	<input type="text" value="0.0.0.0/0"/>	<input type="button" value="Eliminar"/>
sgr-00d703b941c8214ca	TCP personalizado	TCP	8888	Persona...	<input type="text" value="0.0.0.0/0"/>	<input type="button" value="Eliminar"/>

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 tunnelling. Set up SSH tunnelling 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 la tabla siguiente se muestran las interfaces web que están disponibles en los nodos esclavos:

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

The screenshot shows the Hue web interface running on a browser. The URL is `ec2-3-236-241-77.compute-1.amazonaws.com:8888/hue/editor?editor=56`. The interface is in Spanish. On the left, there is a sidebar with a file explorer showing a directory named `clima_procesado` containing a table named `ccrdata`. The main area shows a Hive query editor with the query `select * from ccrdata;` and a button to execute it. Below the query editor, there is a log showing the execution status: `INFO : Completed executing command(queryId=hive_20220916034225_71847d69-4fa4-4d95-9c19-b1b33c841ed5); Time taken: 0.0 seconds`. At the bottom, there is a table of results with 9 rows and 3 columns: `ccrdata.countryname`, `ccrdata.countrycode`, and `ccrdata.indicatorname`. The results are as follows:

	ccrdata.countryname	ccrdata.countrycode	ccrdata.indicatorname
1	countryname	countrycode	indicatorname
2	"Country Name"	"Country Code"	"Indicator Name"
3	"Cook Islands"	COK	"Access to electricity (% of population)"
4	"Cook Islands"	COK	"Account (% age 15+)"
5	"Cook Islands"	COK	"Account"
6	"Cook Islands"	COK	"Additional people below \$1.90 as % of total population"
7	"Cook Islands"	COK	"Additional people below \$1.90 as % of total population"
8	"Cook Islands"	COK	"Additional people below \$1.90 as % of total population"
9	"Cook Islands"	COK	"Additional people below \$1.90 as % of total population"

Query en Hive conectado a la base de datos.

The screenshot shows the Hive web interface. On the left, a sidebar lists tables under the 'clima_procesado' database, including 'ccrdata' and 'ccrdata_v1'. The main area displays a query: `select * from ccrdata where countryname like 'Colombia';`. The query has been executed, and the results are shown in a table with 10 rows. The table has three columns: 'ccrdata.countryname', 'ccrdata.countrycode', and 'ccrdata.indicatorname'. The results show data for Colombia with country code 'COL' and various indicators.

Query: `select * from ccrdata where countryname like 'Colombia';`

Results (100+):

	ccrdata.countryname	ccrdata.countrycode	ccrdata.indicatorname
1	Colombia	COL	"Access to electricity (% of population)"
2	Colombia	COL	"Account (% age 15+)"
3	Colombia	COL	"Account"
4	Colombia	COL	"Additional people below \$1.90 as % of total population"
5	Colombia	COL	"Additional people below \$1.90 as % of total population"
6	Colombia	COL	"Additional people below \$1.90 as % of total population"
7	Colombia	COL	"Additional people below \$1.90 as % of total population"
8	Colombia	COL	"Additional people below \$1.90 as % of total population"
9	Colombia	COL	"Additional people below \$1.90 as % of total population"
10	Colombia	COL	"Additional people below \$1.90 as % of total population"

The screenshot shows the Hive web interface. On the left, a sidebar lists tables under the 'clima_procesado' database, including 'ccrdata' and 'ccrdata_v1'. The main area displays a query: `select countryname, count(*) as c from ccrdata group by countryname order by c desc;`. The query has been executed, and the results are shown in a table with 9 rows. The table has two columns: 'countryname' and 'c'. The results show the count of records for each country, ordered by count in descending order.

Query: `select countryname, count(*) as c from ccrdata group by countryname order by c desc;`

Results (200+):

	countryname	c
1	"Congo"	864
2	"Korea"	864
3	"American Samoa"	432
4	"Bosnia and Herzegovina"	432
5	"Brunei Darussalam"	432
6	"Central African Republic"	432
7	"Costa Rica"	432
8	"Egypt"	432
9	"El Salvador"	432

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

Zeppelin Notebook Job

test

spark

res3: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@6642319a

Took 0 sec. Last updated by anonymous at September 16 2022, 10:06:34 PM.

sc

res4: org.apache.spark.SparkContext = org.apache.spark.SparkContext@7a5fe95f

Took 0 sec. Last updated by anonymous at September 16 2022, 10:06:36 PM.

%spark.pyspark

```
df = spark.read.csv('s3://aoja-t1-2022/trusted/clima-v1/ccrdata/run-S3bucket_node3-1-part-r-00000', inferSchema=True, header=True)
df = df[df['countryname'] != 'Country Name']
z.show(df.limit(30))
```

SPARK JOB FINISHED

countryname	countrycode	indicatorname	indicatorcode
Cook Islands	COK	Access to electricity (% of population)	EG.ELC.ACCS.ZS
Cook Islands	COK	Account (% age 15+)	account.t.d
Cook Islands	COK	Account, income, poorest 40% (% ages 15+)	account.t.d.7
Cook Islands	COK	Additional people below \$1.90 as % of total population by impact - Agriculture Revenues	CC.AVPB.PTPIAR
Cook Islands	COK	Additional people below \$1.90 as % of total population by impact - All impacts	CC.AVPB.PTPIAI
Cook Islands	COK	Additional people below \$1.90 as % of total population by impact - Disasters	CC.AVPB.PTPIDI

Eliminamos la primera fila que se agregó como header

%spark.pyspark

```
df = spark.read.csv('s3://aoja-t1-2022/trusted/clima-v1/ccrdata/run-S3bucket_node3-1-part-r-00000', inferSchema=True, header=True)
df = df[df['countryname'] != 'Country Name']
z.show(df.limit(30))
```

SPARK JOB FINISHED

countryname	countrycode	indicatorname	indicatorcode
Cook Islands	COK	Annual investment needs for coastal protection, by risk strategy (% of GDP) - optimal protection	CC.INCP.SPMC
Cook Islands	COK	Annual investments needed to make transport infrastructure more resilient by 2030 (% of baseline investment costs)	CC.TNET.INV.ZS
Cook Islands	COK	Annual methane emissions at operating coal mines (Mt CO2e 100 years)	CC.COAL.EMIS.CH
Cook Islands	COK	Annual methane emissions at proposed coal projects (Mt CO2e 100 years)	CC.COAL.EMPR.CH
Cook Islands	COK	Arable land (% of land area)	AG.LND.ARBL.ZS

Took 1 sec. Last updated by anonymous at September 16 2022, 10:42:53 PM.

%spark.pyspark

```
df.count()
```

95040

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

Contamos la cantidad de registros y filtramos por país = colombia

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

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

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.PTPIAR
Colombia	COL	Additional people below \$1.90 as % of total population by impact - All impacts	CC.AVPB.PTPLAI
Colombia	COL	Additional people below \$1.90 as % of total population by impact - Disasters	CC.AVPB.PTPLDI

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

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()
```

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
```

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

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
Cook Islands	COK

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

