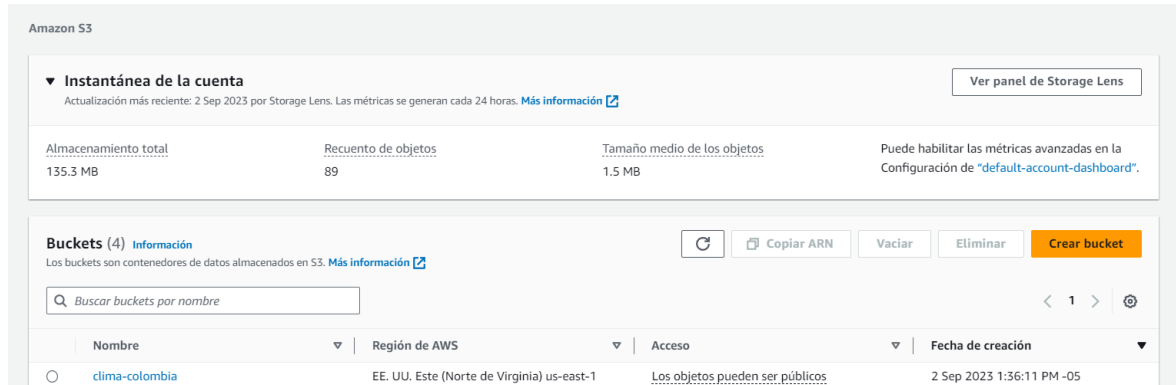


TRABAJO #1 - ALMACENAMIENTO Y RECUPERACIÓN DE INFORMACIÓN

Integrantes: Ricardo Gandica, Samuel Ceballos, Daniela Niño

Parte 1: Ingesta de datos

Creamos un bucket de S3 llamado clima-colombia:



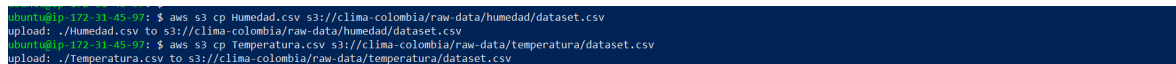
Creamos una instancia de EC2:



Nos conectamos a la instancia de EC2 y descargamos los datasets:



Copiamos los datasets en el bucket clima-colombia en la respectiva carpeta de cada uno:



En el bucket clima-colombia nos deben aparecer los datos:

- Los datos de humedad deben aparecer en las carpetas raw-data/humedad/

Amazon S3 > Buckets > clima-colombia > raw-data/ > humedad/

humedad/ Copiar URI de S3

Objetos Propiedades

Objetos (1)
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Recargar Copiar URI de S3 Copiar URL Descargar Abrir Eliminar Acciones Crear carpeta Cargar

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	dataset.csv	csv	3 Sep 2023 7:18:47 PM -05	73.6 MB	Estándar

- Los datos de temperatura deben aparecer en las carpetas raw-data/temperatura/

Amazon S3 > Buckets > clima-colombia > raw-data/ > temperatura/

temperatura/ Copiar URI de S3

Objetos Propiedades

Objetos (1)
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Recargar Copiar URI de S3 Copiar URL Descargar Abrir Eliminar Acciones Crear carpeta Cargar

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	dataset.csv	csv	3 Sep 2023 7:19:10 PM -05	75.5 MB	Estándar

Parte 2: Modificar y catalogar los datos

Creamos una base de datos en Glue llamada clima-colombia:

AWS Glue > Databases

Databases (3) Last updated (UTC) September 4, 2023 at 24:26:15 Recargar Edit Delete Add database

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

<input type="checkbox"/>	Name	Description	Location URI	Created on (UTC)
<input type="checkbox"/>	clima-colombia	-	-	September 2, 2023 at 21:19:18

Creamos un ETL job para pasar los datos de humedad de la carpeta raw-data a la carpeta trusted-data en el bucket clima-colombia de S3 y para crear un catálogo en Glue a partir de los datos en trusted-data:

Data source properties - S3 | Output schema | Data preview

S3 source type [Info](#)

☒ S3 location
Choose a file or folder in an S3 bucket.

☐ Data Catalog table

S3 URL

☒ Recursive
Read files in all subdirectories.

Data format
CSV

Delimiter
Comma (,)

Escape character - optional
Enter a character to use for escaping

The character which immediately follows is used as-is, except for a small set of well-known escapes (\n, \r, \t, and \0)

Quote character
Double quote (")

☒ First line of source file contains column headers

Transform | Output schema | Data preview

Name
Change Schema

Node parents
Choose which nodes will provide inputs for this one.

☒ S3 bucket
S3 - DataSource

Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
codigoestacion	codigo_estacion	bigint	<input type="checkbox"/>
codigosensor			<input checked="" type="checkbox"/>
fechaobservacion	fecha	string	<input type="checkbox"/>
valorobservado	humedad	float	<input type="checkbox"/>
nombrestacion	nombre_estacion	string	<input type="checkbox"/>
departamento	departamento	string	<input type="checkbox"/>
municipio	municipio	string	<input type="checkbox"/>
zonahidrografica	zona	string	<input type="checkbox"/>
latitud			<input checked="" type="checkbox"/>
longitud			<input checked="" type="checkbox"/>
descripcionsensor			<input checked="" type="checkbox"/>
unidadmedida			<input checked="" type="checkbox"/>

Transform | Output schema | Data preview

Name: Change Schema

Node parents: Choose one or more parent node

S3 bucket - DataSource

Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
codigoestacion	codigo_estacion	bigint	<input type="checkbox"/>
codigosensor			<input checked="" type="checkbox"/>
fechaobservacion	fecha	string	<input type="checkbox"/>
valorobservado	temperatura	float	<input type="checkbox"/>
nombrestacion	nombre_estacion	string	<input type="checkbox"/>
departamento	departamento	string	<input type="checkbox"/>
municipio	municipio	string	<input type="checkbox"/>
zonahidrografica	zona	string	<input type="checkbox"/>
latitud			<input checked="" type="checkbox"/>
longitud			<input checked="" type="checkbox"/>
descripcionsensor			<input checked="" type="checkbox"/>
unidadmedida			<input checked="" type="checkbox"/>

Data target properties - S3 | Output schema | Data preview

Name: S3 bucket

Node parents: Choose one or more parent node

Change Schema - ApplyMapping - Transform

Format: Parquet

Compression Type: Choose a compression type

S3 Target Location: Choose an S3 location in the format s3://bucket/prefix/object/ with a trailing slash (/).
s3://clima-colombia/trusted-data/temperatura/

Data Catalog update options: **Info**
Choose how you want to update the Data Catalog table's schema and partitions. These options will only apply if the Data Catalog table is an S3 backed source.
☐ Do not update the Data Catalog
☒ Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions
☐ Create a table in the Data Catalog and on subsequent runs, keep existing schema and add new partitions

Database: Choose the database from the AWS Glue Data Catalog.
clima-colombia

► Use runtime parameters

Table name: Enter a table name for the AWS Glue Data Catalog.
temperatura

Corremos los ETL jobs y en la base de datos de Glue nos deben aparecer las dos tablas:

AWS Glue > Databases > clima-colombia

clima-colombia

Last updated (UTC) September 4, 2023 at 04:03:05

Database properties

Name	Description	Location	Created on (UTC)
clima-colombia	-	-	September 2, 2023 at 21:19:18

Tables (2)
View and manage all available tables.

Last updated (UTC) September 4, 2023 at 04:05:21

Filter tables

Name	Database	Location	Classification	Deprecated	View data	Data quality
humedad	clima-colombia	s3://clima-colombia/truste	Parquet	-	Table data	View data quality
temperatura	clima-colombia	s3://clima-colombia/truste	Parquet	-	Table data	View data quality

- Tabla de humedad

Schema

Partitions

Indexes

Schema (7)

View and manage the table schema.

Edit schema as JSON

Edit schema

Filter schemas

< 1 > ⚙

#	Column name	Data type	Partition key	Comment
1	codigo_estacion	bigint	-	-
2	fecha	string	-	-
3	humedad	float	-	-
4	nombre_estacion	string	-	-
5	departamento	string	-	-
6	municipio	string	-	-
7	zona	string	-	-

- Tabla de temperatura

Schema

Partitions

Indexes

Schema (7)

View and manage the table schema.

Edit schema as JSON

Edit schema

Filter schemas

< 1 > ⚙

#	Column name	Data type	Partition key	Comment
1	codigo_estacion	bigint	-	-
2	fecha	string	-	-
3	temperatura	float	-	-
4	nombre_estacion	string	-	-
5	departamento	string	-	-
6	municipio	string	-	-
7	zona	string	-	-

También nos deben aparecer los datos en el bucket clima-colombia de S3:

- Los datos de humedad deben aparecer en las carpetas trusted-data/humedad/

Amazon S3

Buckets

clima-colombia

trusted-data/

humedad/

Copiar URI de S3

Objetos

Propiedades

Objetos (2)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [Inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

🔄

Copiar URI de S3

Copiar URL

⬇ Descargar

📄 Abrir

🗑 Eliminar

Acciones

Crear carpeta

Cargar

Buscar objetos por prefijo

< 1 > ⚙

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	run-1693800208783-part-block-0-r-00000-snappy.parquet	parquet	3 Sep 2023 11:04:28 PM -05	3.8 MB	Estándar
<input type="checkbox"/>	run-1693800208783-part-block-0-r-00001-snappy.parquet	parquet	3 Sep 2023 11:04:22 PM -05	600.6 KB	Estándar

- Los datos de temperatura deben aparecer en las carpetas trusted-data/temperatura/

Amazon S3 > Buckets > clima-colombia > trusted-data/ > temperatura/

temperatura/ Copiar URI de S3

Objetos | Propiedades

Objetos (2)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [Inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Recargar Copiar URI de S3 Copiar URL Descargar Abrir Eliminar Acciones Crear carpeta Cargar

Buscar objetos por prefijo

	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	run-1693800186725-part-block-0-r-00000-snappy.parquet	parquet	3 Sep 2023 11:04:13 PM -05	4.1 MB	Estándar
<input type="checkbox"/>	run-1693800186725-part-block-0-r-00001-snappy.parquet	parquet	3 Sep 2023 11:04:09 PM -05	772.7 KB	Estándar

Parte 3: Consultas desde Athena

Hacemos consultas desde Athena usando los catálogos:

- Join entre las dos tablas

Query 8

```

1 SELECT temperatura.fecha, temperatura.temperatura, humedad.humedad, temperatura.departamento, temperatura.municipio
2 FROM "clima-colombia".temperatura
3 JOIN "clima-colombia".humedad
4 ON temperatura.fecha = humedad.fecha
5 AND temperatura.codigo_estacion = humedad.codigo_estacion
6 LIMIT 10;

```

SQL Ln 1, Col 91

Run again Explain Cancel Clear Create

Query results | Query stats

Completed Time in queue: 174 ms Run time: 1.054 sec Data scanned: 3.16 MB

Results (10)

#	fecha	temperatura	humedad	departamento	municipio
1	01/27/2023 11:25:00 AM	25.29858	62.45414	TOLIMA	LIBANO
2	01/27/2023 11:25:00 AM	29.76772	39.60738	META	LEJANIAS
3	01/27/2023 11:25:00 AM	26.97998	51.78831	QUINDIO	MONTENEGRO
4	01/27/2023 11:25:00 AM	27.94687	49.9254	QUINDIO	BUENAVISTA
5	01/27/2023 11:25:00 AM	26.34907	50.36818	CALDAS	CHINCHINA
6	01/27/2023 11:25:00 AM	26.92156	56.75153	VALLE DEL CAUCA	CAicedonia
7	01/27/2023 11:25:00 AM	23.35229	71.03864	CAUCA	BALBOA
8	01/27/2023 11:25:00 AM	25.93711	65.558	BOYACA	PAUNA
9	01/27/2023 11:25:00 AM	26.62858	51.96909	RISARALDA	GUATICA
10	01/27/2023 11:25:00 AM	18.78267	85.16475	TOLIMA	DOLORES

- Join entre las dos tablas y contar el número de registros donde la temperatura es mayor a 23° y la humedad es mayor a 60%

Query 8

```

1 SELECT count(*)
2 FROM "clima-colombia".temperatura
3 JOIN "clima-colombia".humedad
4 ON temperatura.fecha = humedad.fecha
5 AND temperatura.codigo_estacion = humedad.codigo_estacion
6 WHERE temperatura.temperatura > 23
7 AND humedad.humedad > 60;

```

SQL Ln 7, Col 26

Run again Explain Cancel Clear Create

Query results | Query stats

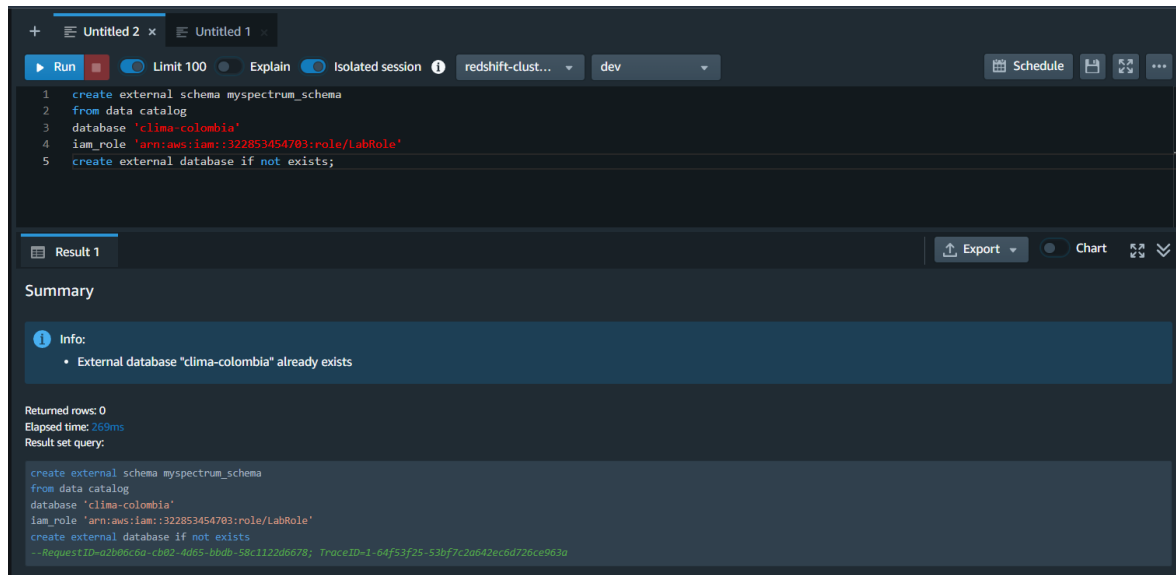
Completed Time in queue: 161 ms Run time: 1.318 sec Data scanned: 5.43 MB

Results (1)

#	_col0
1	66471

Parte 4: Consultas desde Redshift

Creamos la base de datos externa:

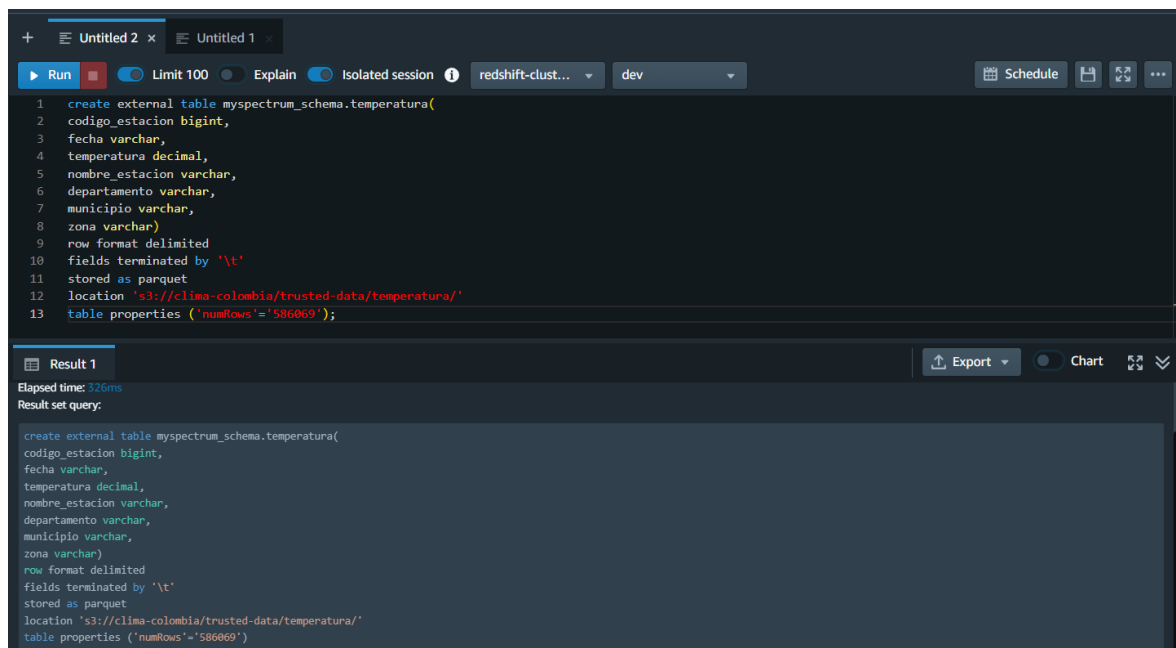


The screenshot shows the AWS Redshift console interface. At the top, there are tabs for 'Untitled 2' and 'Untitled 1'. Below the tabs, there are buttons for 'Run', 'Limit 100', 'Explain', 'Isolated session', and a dropdown menu showing 'redshift-clust...' and 'dev'. To the right of these buttons are icons for 'Schedule', 'Save', 'Share', and a menu icon. The main area contains a SQL query:

```
1 create external schema myspectrum_schema
2 from data catalog
3 database 'clima-colombia'
4 iam_role 'arn:aws:iam::322853454703:role/LabRole'
5 create external database if not exists;
```

Below the query, there is a 'Result 1' tab. The 'Summary' section shows an 'Info' message: 'External database "clima-colombia" already exists'. Below this, it says 'Returned rows: 0' and 'Elapsed time: 269ms'. The 'Result set query:' section shows the same SQL query as above, followed by a green log entry: '--RequestID=a2b06c6a-cb02-4d65-bb0b-58c1122d6678; TraceID=1-64f53f25-53bf7c2a642ec6d726ce963a'.

Creamos una tabla con los datos de temperatura los cuales se encuentran en el bucket clima-colombia de S3 en las carpetas trusted-data/temperatura/:



The screenshot shows the AWS Redshift console interface. At the top, there are tabs for 'Untitled 2' and 'Untitled 1'. Below the tabs, there are buttons for 'Run', 'Limit 100', 'Explain', 'Isolated session', and a dropdown menu showing 'redshift-clust...' and 'dev'. To the right of these buttons are icons for 'Schedule', 'Save', 'Share', and a menu icon. The main area contains a SQL query:

```
1 create external table myspectrum_schema.temperatura(
2 codigo_estacion bigint,
3 fecha varchar,
4 temperatura decimal,
5 nombre_estacion varchar,
6 departamento varchar,
7 municipio varchar,
8 zona varchar)
9 row format delimited
10 fields terminated by '\t'
11 stored as parquet
12 location 's3://clima-colombia/trusted-data/temperatura/'
13 table properties ('numRows'='586069');
```

Below the query, there is a 'Result 1' tab. The 'Elapsed time: 326ms' is shown. The 'Result set query:' section shows the same SQL query as above, followed by a green log entry: '--RequestID=a2b06c6a-cb02-4d65-bb0b-58c1122d6678; TraceID=1-64f53f25-53bf7c2a642ec6d726ce963a'.

Hacemos una consulta a la tabla temperatura:

The screenshot shows a Redshift SQL client interface. At the top, there are tabs for 'Untitled 2' and 'Untitled 1'. Below the tabs, there are buttons for 'Run', 'Limit 100', 'Explain', 'Isolated session', and a dropdown menu showing 'redshift-clust...' and 'dev'. The main area contains a single SQL query: `1 SELECT COUNT(*) FROM myspectrum_schema.temperatura;`. Below the query, there is a section labeled 'Result 1 (1)' which displays a table with two rows: 'count' and '586069'. To the right of the result, there are buttons for 'Export', 'Chart', and a refresh icon.

count
586069

Creamos una tabla nativa para los datos de humedad:

The screenshot shows a Redshift SQL client interface. At the top, there are tabs for 'Untitled 2' and 'Untitled 1'. Below the tabs, there are buttons for 'Run', 'Limit 100', 'Explain', 'Isolated session', and a dropdown menu showing 'redshift-clust...' and 'dev'. The main area contains a SQL query to create a table: `1 create table humedad(
2 codigo_estacion bigint,
3 fecha varchar,
4 humedad DOUBLE PRECISION,
5 nombre_estacion varchar,
6 departamento varchar,
7 municipio varchar,
8 zona varchar);`. Below the query, there is a section labeled 'Result 1' which displays a 'Summary' of the query execution. The summary includes 'Returned rows: 0', 'Elapsed time: 48ms', and 'Result set query:'. The query text is repeated in the result area. At the bottom, there is a green log message: `--RequestID=030620cc-c843-42e1-b714-55a8d4ef170e; TraceID=1-64f54921-411457061bbb71a363da589d`.

Summary
Returned rows: 0
Elapsed time: 48ms
Result set query:

create table humedad(
codigo_estacion bigint,
fecha varchar,
humedad DOUBLE PRECISION,
nombre_estacion varchar,
departamento varchar,
municipio varchar,
zona varchar)
--RequestID=030620cc-c843-42e1-b714-55a8d4ef170e; TraceID=1-64f54921-411457061bbb71a363da589d

Cargamos los datos a la tabla humedad usando los datos que se encuentran en el bucket clima-colombia de S3 en las carpetas trusted-data/humedad/:

The screenshot shows a Redshift SQL client interface. At the top, there are tabs for 'Untitled 2' and 'Untitled 1'. Below the tabs, there are buttons for 'Run', 'Limit 100', 'Explain', 'Isolated session', and a dropdown menu showing 'redshift-clust...' and 'dev'. The main area contains a SQL query to load data from S3: `1 COPY humedad
2 FROM 's3://clima-colombia/trusted-data/humedad/'
3 IAM_ROLE 'arn:aws:iam::322853454703:role/LabRole'
4 FORMAT AS PARQUET;`. Below the query, there is a section labeled 'Result 1' which displays a 'Summary' of the query execution. The summary includes 'Returned rows: 0', 'Elapsed time: 20.3s', and 'Result set query:'. The query text is repeated in the result area. Below the query text, there is a blue information box with the text: 'Info: Load into table 'humedad' completed, 549562 record(s) loaded successfully.' At the bottom, there is a green log message: `--RequestID=1fd88ac-6598-4c09-92c6-d5dc1c0d82b8; TraceID=1-64f54796-6a7b0c0c25ee6b314592044`.

Summary
Info:
• Load into table 'humedad' completed, 549562 record(s) loaded successfully.

Returned rows: 0
Elapsed time: 20.3s
Result set query:

COPY humedad
FROM 's3://clima-colombia/trusted-data/humedad/'
IAM_ROLE 'arn:aws:iam::322853454703:role/LabRole'
FORMAT AS PARQUET
--RequestID=1fd88ac-6598-4c09-92c6-d5dc1c0d82b8; TraceID=1-64f54796-6a7b0c0c25ee6b314592044

Hacemos una consulta a la tabla humedad:

+

Untitled 2 x

Untitled 1

▶ Run

Limit 100

Explain

Isolated session

redshift-clust...

dev

Schedule

...

1 SELECT COUNT(*) FROM humedad;

Result 1 (1)

Export Chart

count	
549562	

Hacemos una consulta usando la tabla externa temperatura y la tabla nativa humedad:

+

Untitled 2 x

Untitled 1

▶ Run

Limit 100

Explain

Isolated session

redshift-clust...

dev

Schedule

...

1 SELECT humedad.fecha, myspectrum_schema.temperatura.temperatura, humedad.humedad, humedad.departamento, humedad.municipio
2 FROM myspectrum_schema.temperatura
3 JOIN humedad
4 ON myspectrum_schema.temperatura.fecha = humedad.fecha
5 AND myspectrum_schema.temperatura.codigo_estacion = humedad.codigo_estacion
6 LIMIT 10;

Result 1 (10)

Export Chart

fecha	humedad	departamento	municipio	
01/01/2023 12:00:00 AM	99.97608947753906	QUINDÍO	PIJAO	
01/01/2023 12:00:00 AM	94.90614318847656	CALDAS	CHINCHINA	
01/01/2023 12:00:00 AM	0	CAQUETA	SOLANO	
01/01/2023 12:00:00 AM	81.2428970336914	TOLIMA	DOLORES	
01/01/2023 12:00:00 AM	99.40284729003906	QUINDÍO	CALARCÁ	
01/01/2023 12:00:00 AM	88	ATLANTICO	JUAN DE ACOSTA	
01/01/2023 12:00:00 AM	93.56640625	CALDAS	SUPIA	
01/01/2023 12:00:00 AM	93	ANTIOQUIA	MEDELLÍN	
01/01/2023 12:00:00 AM	84.355712890625	RISARALDA	PEREIRA	
01/01/2023 12:00:00 AM	77.6642074584961	CALDAS	MANZANARES	

Parte 5: Implementación de cluster EMR

Entramos al servicio de EMR y creamos un cluster con la siguiente configuración:

Create cluster [Info](#)

Name and applications [Info](#)

Name

T1

Amazon EMR release [Info](#)

A release contains a set of applications which can be installed on your cluster.

emr-6.12.0

Application bundle

Spark

Core Hadoop

Flink

HBase

Presto

Trino

Custom

▼ Customize your application bundle

Applications included in bundle

☐ Flink 1.17.0

☒ HCatalog 3.1.3

☒ Hue 4.11.0

☒ Livy 0.7.1

☐ Phoenix 5.1.3

☒ Spark 3.4.0

☐ Tez 0.10.2

☐ ZooKeeper 3.5.10

☐ Ganglia 3.7.2

☒ Hadoop 3.3.3

☒ JupyterEnterpriseGateway 2.6.0

☐ MXNet 1.9.1

☐ Pig 0.17.0

☐ Sqoop 1.4.7

☐ Trino 414

☐ HBase 2.4.17

☒ Hive 3.1.3

☒ JupyterHub 1.4.1

☐ Oozie 5.2.1

☐ Presto 0.281

☐ TensorFlow 2.11.0

☒ Zeppelin 0.10.1

Summary [Info](#)

Name and applications

Name

T1

Amazon EMR release

emr-6.12.0

Application bundle

Custom (HCatalog 3.1.3, Hadoop 3.3.3, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2...)

Amazon Linux release

2.0.20230822.0

Cluster configuration

Instance groups

Primary (m4.xlarge), Core (m4.xlarge), Task (m4.xlarge)

Cluster scaling and provisioning option

Cancel

Create cluster

Cluster configuration [Info](#)

Choose a configuration method for the primary, core, and task node groups for your cluster.

☒ Instance groups

Choose one instance type per node group

☐ Instance fleets

Choose any combination of instance types within each node group

Instance groups

Primary

Choose EC2 instance type

m4.xlarge

4 vCore 16 GiB memory EBS only storage

On-Demand price: - Lowest Spot price: -

Actions ▼

☐ Use multiple primary nodes

To improve cluster availability, use 3 primary nodes with the same configuration and bootstrap actions. You can not use multiple primary nodes with instance fleets.

► Node configuration - optional

Core

Choose EC2 instance type

m4.xlarge

4 vCore 16 GiB memory EBS only storage

On-Demand price: - Lowest Spot price: -

Actions ▼

Identity and Access Management (IAM) roles [Info](#)

Choose or create a service role and instance profile for the EC2 instances in your cluster.

Amazon EMR service role [Info](#)

The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.

☒ Choose an existing service role

Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.

☐ Create a service role

Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.

Service role

EMR_DefaultRole



EC2 instance profile for Amazon EMR

The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

☒ Choose an existing instance profile

Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.

☐ Create an instance profile

Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

Instance profile

EMR_EC2_DefaultRole



Parte 5.1: Consultas desde Hive

Consultamos los datos usando los catálogos de Glue:

- Join entre las dos tablas

Search saved documents...

Hive Add a name... Add a description...

42.60s clima-colombia

SELECT temperatura.fecha, temperatura.temperatura, humedad.humedad, temperatura.departamento, temperatura.municipio
FROM 'clima-colombia'.temperatura
JOIN 'clima-colombia'.humedad
ON temperatura.fecha = humedad.fecha
AND temperatura.codigo_estacion = humedad.codigo_estacion
LIMIT 10;

INFO : Map 1: 1/1 Map 2: 1/1
INFO : Completed executing command(queryId=hive_28238984238486_cdc71443-c98c-4fe7-898b-3f72be4cbeaf); Time taken: 42.028 seconds
INFO : OK

Query History Saved Queries Results (10)

	temperatura.fecha	temperatura.temperatura	humedad.humedad	temperatura.departamento	temperatura.municipio
1	01/01/2023 12:00:00 AM	15.24925	99.97609	QUINDÍO	PIJAO
2	01/01/2023 12:00:00 AM	17.76134	94.90614	CALDAS	CHINCHINA
3	01/01/2023 12:00:00 AM	21.8	0	CAQUETA	SOLANO
4	01/01/2023 12:00:00 AM	16.19377	81.2429	TOLIMA	DOLORES
5	01/01/2023 12:00:00 AM	16.56269	99.40285	QUINDÍO	CALARCÁ
6	01/01/2023 12:00:00 AM	25.6	88	ATLANTICO	JUAN DE ACOSTA
7	01/01/2023 12:00:00 AM	19.52354	93.56641	CALDAS	SUPIA
8	01/01/2023 12:00:00 AM	19.3	93	ANTIOQUIA	MEDELLÍN
9	01/01/2023 12:00:00 AM	19.91131	84.35571	RISARALDA	PEREIRA
10	01/01/2023 12:00:00 AM	17.44153	77.66421	CALDAS	MANZANARES

- Join entre las dos tablas y contar el número de registros donde la temperatura es mayor a 23° y la humedad es mayor a 60%

The screenshot shows the Hive console interface. On the left, the 'Databases' sidebar lists 'clima-colombia', 'default', 'onu', and 'ticket'. The main area displays a SQL query:

```
1 SELECT count(*)
2 FROM `clima-colombia`.temperatura
3 JOIN `clima-colombia`.humedad
4 ON temperatura.fecha = humedad.fecha
5 AND temperatura.codigo_estacion = humedad.codigo_estacion
6 WHERE temperatura.temperatura > 23
7 AND humedad.humedad > 60;
```

Below the query, the execution status is shown: 'INFO : Completed executing command(queryId=hive_20238904233125_169f71ae-41d6-4ec8-adc3-b0aas2a7e7a); Time taken: 25.706 seconds', 'INFO : OK', and 'INFO : Concurrency mode is disabled, not creating a lock manager'. The 'Results (1)' tab shows a single row with the value '66471'.

Desde Hive creamos una base de datos llamada clima_colombia_hive (en esta se pondrán los datos leídos desde S3):

The screenshot shows the 'Create a new database' form in the Hive console. The 'Name' field is filled with 'clima_colombia_hive'. The 'Description' field is empty. The 'Default location' checkbox is checked. The 'DESTINATION' section is empty, and the 'PROPERTIES' section is also empty. A green checkmark icon and the text 'No source data' are visible at the top right of the form.

Leemos los datos de temperatura desde S3 y los añadimos a la base de datos clima_colombia_hive:

The screenshot shows the Hive console interface. On the left, the 'Tables' sidebar lists 'humedad' and 'temperatura'. The main area displays a SQL command to create an external table:

```
1 CREATE EXTERNAL TABLE temperatura (
2   codigo_estacion bigint,
3   fecha string,
4   temperatura float,
5   nombre_estacion string,
6   departamento string,
7   municipio string,
8   zona string)
9 ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
10 STORED AS PARQUET
11 LOCATION 's3://clima-colombia/trusted-data/temperatura';
```

Below the command, the execution status is shown: 'INFO : Completed executing command(queryId=hive_20238904235638_d7fcd081-2959-47a4-ad81-317d07127a0b); Time taken: 0.482 seconds', 'INFO : OK', and 'INFO : Concurrency mode is disabled, not creating a lock manager'. A green checkmark icon and the text 'Success.' are visible at the bottom.

Hacemos consultas desde la base de datos clima_colombia_hive:

- Join entre las dos tablas

```
SELECT temperatura.fecha, temperatura.temperatura, humedad.humedad, temperatura.departamento, temperatura.municipio
FROM clima_colombia_hive.temperatura
JOIN clima_colombia_hive.humedad
ON temperatura.fecha = humedad.fecha
AND temperatura.codigo_estacion = humedad.codigo_estacion
LIMIT 10;
```

	temperatura.fecha	temperatura.temperatura	humedad.humedad	temperatura.departamento	temperatura.municipio
1	01/01/2023 12:00:00 AM	15.24925	99.97509	QUINDIO	PIJAO
2	01/01/2023 12:00:00 AM	17.76134	94.90614	CALDAS	CHINCHINA
3	01/01/2023 12:00:00 AM	21.8	0	CAQUETA	SOLANO
4	01/01/2023 12:00:00 AM	16.19377	81.2429	TOlima	DOLORES
5	01/01/2023 12:00:00 AM	16.56269	99.40285	QUINDIO	CALARCA
6	01/01/2023 12:00:00 AM	25.6	88	ATLANTICO	JUAN DE ACOSTA
7	01/01/2023 12:00:00 AM	19.52354	93.56641	CALDAS	SUPIA
8	01/01/2023 12:00:00 AM	19.3	93	ANTIOQUIA	MEDELLIN
9	01/01/2023 12:00:00 AM	19.91131	84.35571	RISARALDA	PEREIRA
10	01/01/2023 12:00:00 AM	17.44153	77.66421	CALDAS	MANZANARES

- Join entre las dos tablas y contar el número de registros donde la temperatura es mayor a 23° y la humedad es mayor a 60%

```
SELECT count(*)
FROM clima_colombia_hive.temperatura
JOIN clima_colombia_hive.humedad
ON temperatura.fecha = humedad.fecha
AND temperatura.codigo_estacion = humedad.codigo_estacion
WHERE temperatura.temperatura > 23
AND humedad.humedad > 60;
```

	_c0
1	66471

Si entrenamos a Glue y vemos las bases de datos, podremos ver los catálogos que creamos desde Hive:

Database properties

Name	Description	Location	Created on (UTC)
clima_colombia_hive	-	hdfs://ip-172-31-83-86.ec2.internal:8020/user/hive/warehouse/clima_colombia_hive.db	September 4, 2023 at 23:54:24

Tables (2)

Name	Database	Location	Classification	Deprecated	View data	Data quality
humedad	clima_colombia_hive	s3://clima-colombia/trusted-dat	-	-	Table data	View data quality
temperatura	clima_colombia_hive	s3://clima-colombia/trusted-dat	-	-	Table data	View data quality

Parte 5.2: Ambiente de procesamiento

Creamos un ambiente de procesamiento de pyspark en el cluster de EMR y desde allí accedimos a los datos en S3 mediante SparkSQL:

No seguro | https://ec2-44-212-32-127.compute-1.amazonaws.com:9443/user/jovyan/notebooks/clima_notebook.ipynb

jupyterhub clima_notebook Last Checkpoint: a few seconds ago (autosaved)

LogoutControl Panel

FileEditViewInsertCellKernelWidgetsHelpKernel starting, please wait...Not TrustedPySpark

In [1]: spark

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
3	application_1693866466115_0007	pyspark	idle	Link	Link	None	✓

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...

SparkSession available as 'spark'.

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...

<pyspark.sql.session.SparkSession object at 0x7ff223adce10>

In [2]: spark.catalog.setCurrentDatabase("clima_colombia_hive")

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...

In [3]: df_temperatura = spark.sql("select * from temperatura")

df_temperatura.show()

FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),...

[codigo_estacion]	[fecha]	[temperatura]	[nombre_estacion]	[departamento]	[municipio]	[zona]
23025502	01/01/2023 12:00:...	1.888466	ALMACAFE LETRAS ...	CALDAS	MANIZALES	MEDIO MAGDALENA
26125509	01/01/2023 12:00:...	15.24925	ESPERANZA LA - AUT	QUINDIO	PIJAO	CAUCA
26165501	01/01/2023 12:00:...	15.46122	EL CIPRES - AUT	CALDAS	SALAMINA	CAUCA
52055501	01/01/2023 12:00:...	15.26959	OSPINA PEREZ - AUT	NARIÑO	CONSACÁ	PATÍA
26135503	01/01/2023 12:00:...	17.76134	NARANJAL - AUT	CALDAS	CHINCHINA	CAUCA
44055010	01/01/2023 12:00:...	21.8	TRES ESQUINAS	CAQUETA	SOLANO	CAQUETA
16015501	01/01/2023 12:00:...	23.9	APTO CAMILO DAZA	NORTE DE SANTANDER	CÚCUTA	CATATUMBO
21165501	01/01/2023 12:00:...	16.19377	DOLORES - AUT	TOLIMA	DOLORES	ALTO MAGDALENA
16025501	01/01/2023 12:00:...	17.98149	CUCUTILLA - AUT	NORTE DE SANTANDER	CUCUTILLA	CATATUMBO
21205791	01/01/2023 12:00:...	12.4	APTO EL DORADO - ...	BOGOTA D.C.	BOGOTA, D.C.	ALTO MAGDALENA
26125505	01/01/2023 12:00:...	16.56269	LA BELLA - AUT	QUINDIO	CALARCÁ	CAUCA
29045110	01/01/2023 12:00:...	25.6	JUAN DE ACOSANTA	ATLANTICO	JUAN DE ACOSTA	CARIBE - LITORAL
26175501	01/01/2023 12:00:...	19.52354	RAFAEL ESCOBAR - ...	CALDAS	SUPIA	CAUCA
27015330	01/01/2023 12:00:...	19.3	APTO OLAYA HERRER...	ANTIOQUIA	MEDELLIN	NECHÍ
26135501	01/01/2023 12:00:...	19.91131	EL PILANO - AUT	RISARALDA	PEREIRA	CAUCA
23025501	01/01/2023 12:00:...	17.44153	MANZANARES - AUT	CALDAS	MANZANARES	MEDIO MAGDALENA

Parte 5.3: HDFS

Copiamos los archivos del bucket clima-colombia a HDFS:

File Browser

Search for file nameActionsCopy PathOpen in Importer

Home

/user/hadoop/clima_colombia

Name





↑

.





humedad

temperatura

 Home /user/hadoop/clima_colombia/**humedad**

<input type="checkbox"/>	Name
<input type="checkbox"/>	 ↑
<input type="checkbox"/>	 .
<input type="checkbox"/>	 run-1693800208783-part-block-0-r-00000-snappy.parquet
<input type="checkbox"/>	 run-1693800208783-part-block-0-r-00001-snappy.parquet

 Home /user/hadoop/clima_colombia/**temperatura**

<input type="checkbox"/>	Name
<input type="checkbox"/>	 ↑
<input type="checkbox"/>	 .
<input type="checkbox"/>	 run-1693800186725-part-block-0-r-00000-snappy.parquet
<input type="checkbox"/>	 run-1693800186725-part-block-0-r-00001-snappy.parquet

Información adicional

Bucket de clima-colombia: s3://clima-colombia/

GitHub: <https://github.com/sceballosp/trabajos-ari-st1800/tree/master/trabajo-1/>

- Jupyter notebook del ambiente de procesamiento:
<https://github.com/sceballosp/trabajos-ari-st1800/tree/master/trabajo-1/code/emr/pyspark>