

TRABAJO 1 – IMPLEMENTACIÓN DATALAKE AWS

SANTIAGO ALBERTO MORENO QUEVEDO

DANIEL GARCÍA GARCÍA

PROFESOR: EDWIN MONTOYA

ALMACENAMIENTO Y RECUPERACIÓN DE LA INFORMACIÓN

UNIVERSIDAD EAFIT

MEDELLÍN – ANTIOQUIA

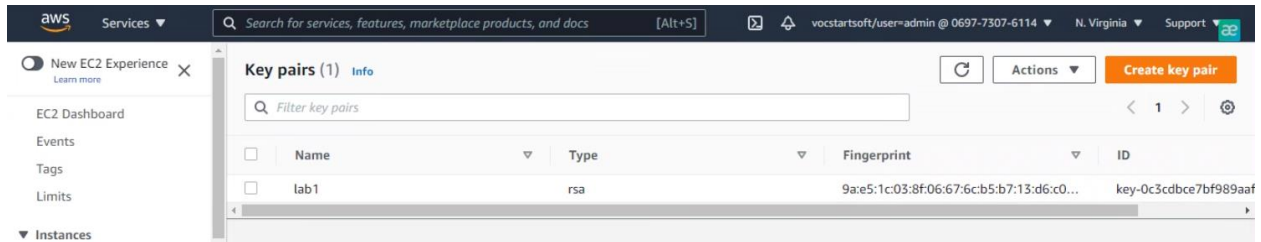
2021 - 2

Tabla de Contenido

Implementación del Datalake	3
EC2 Key Pair	3
Creación del bucket S3.....	3
Particionamiento y zonas del datalake	5
Subida de archivos al datalake.....	7
Cambio de políticas del bucket	11
Creación de primer Crawler	11
Query de prueba en Athena.....	13
Procesos ETL y Crawlers con Glue	14
Consultas en Athena	21
Creación del EMR	23
Consultas con Hive	25

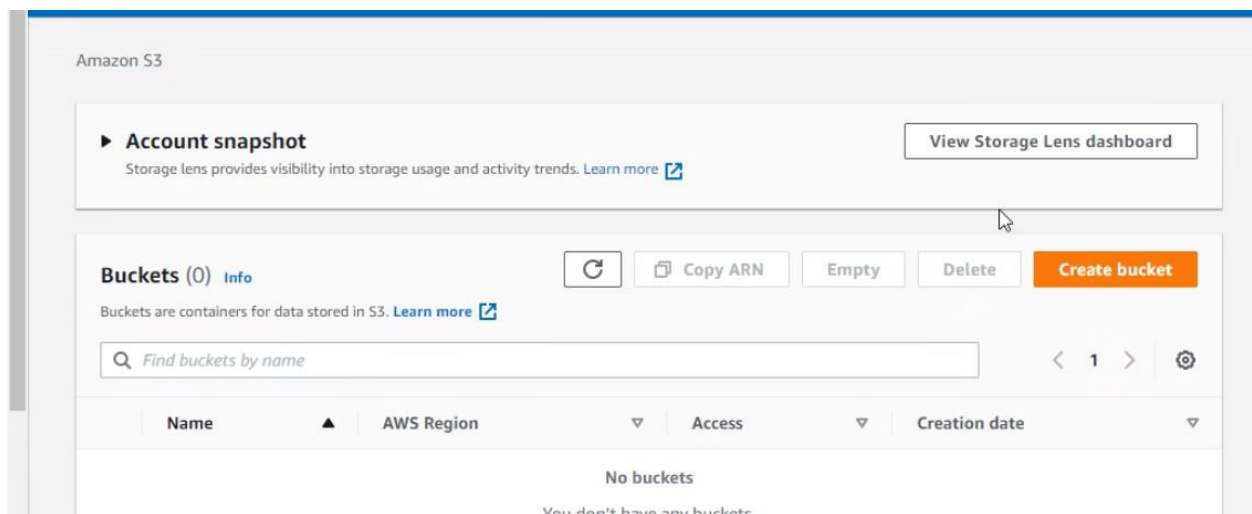
Implementación del Datalake

EC2 Key Pair



Primero debemos crear una Key Pair en AWS EC2 que usaremos posteriormente en la creación de los diferentes elementos que nos proporciona AWS para implementar un datalake.

Creación del bucket S3



Create bucket [Info](#)

Buckets are containers for data stored in S3. [Learn more](#)

General configuration

Bucket name

Bucket name must be unique and must not contain spaces or uppercase letters. [See rules for bucket naming](#)

AWS Region

US East (N. Virginia) us-east-1

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

[Choose bucket](#)

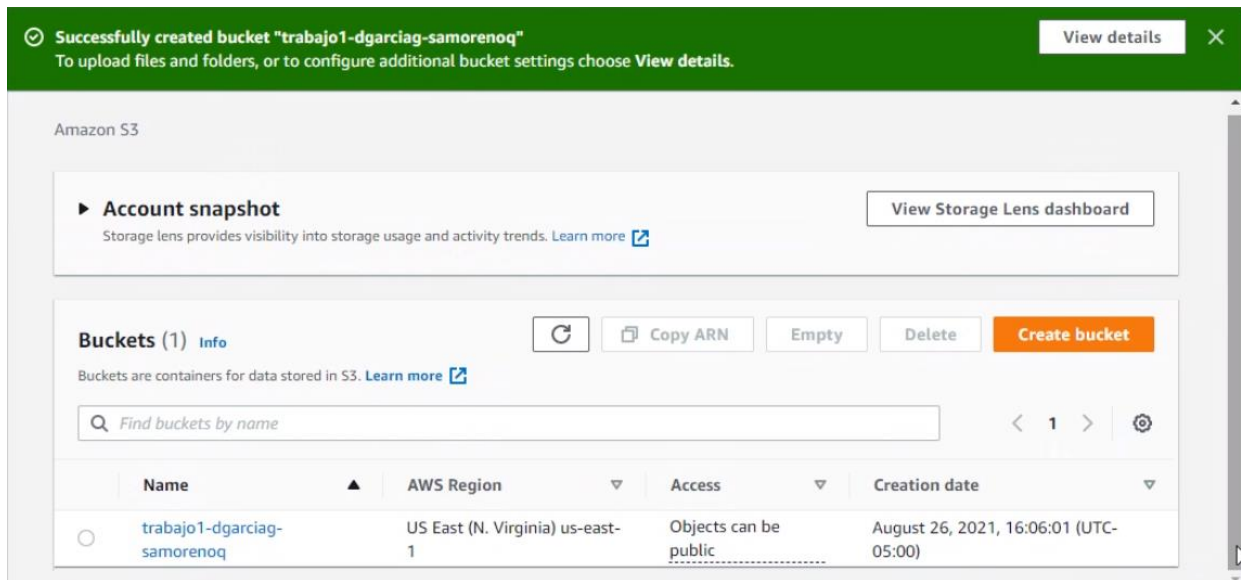
Block Public Access settings for this bucket

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to this bucket and its objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to this bucket or objects within, you can customize the individual settings below to suit your specific storage use cases. [Learn more](#)

☐ Block all public access

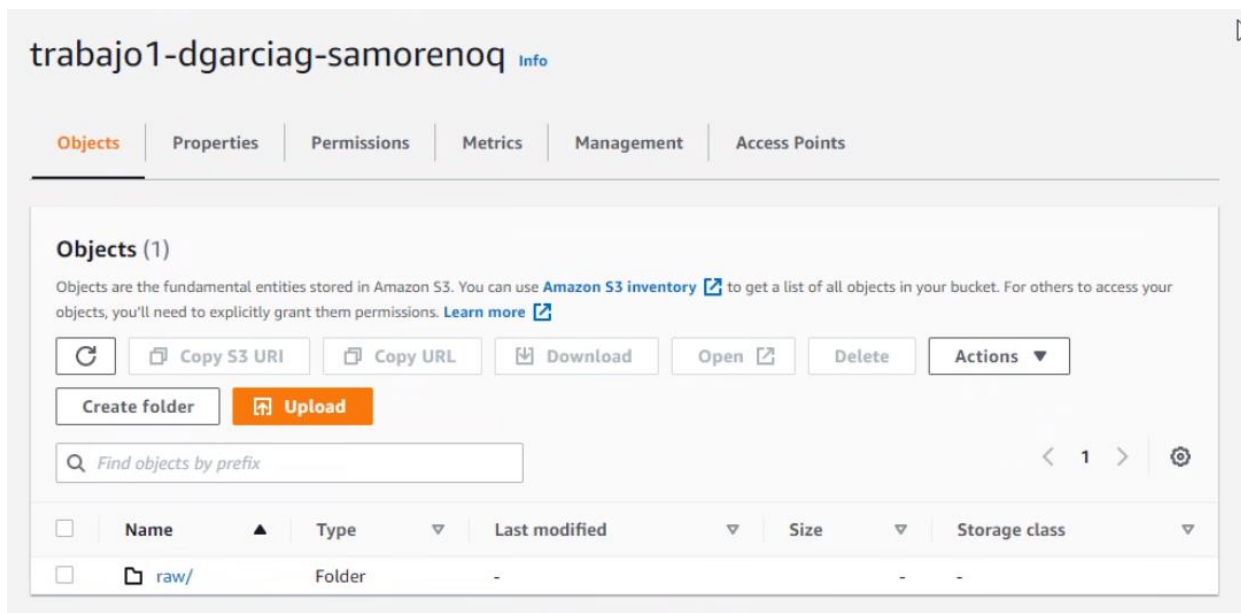
Turning this setting on is the same as turning on all four settings below. Each of the following settings are independent of one another.

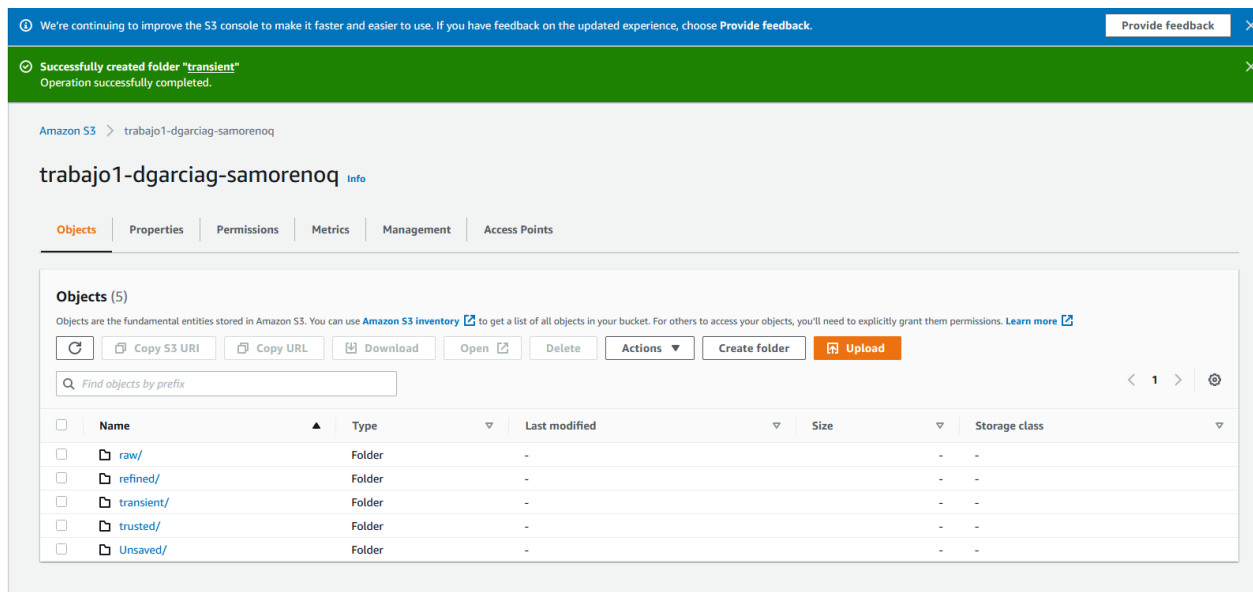
- ☐ **Block public access to buckets and objects granted through *new* access control lists (ACLs)**
S3 will block public access permissions applied to newly added buckets or objects, and prevent the creation of new public access ACLs for existing buckets and objects. This setting doesn't change any existing permissions that allow public access to S3 resources using ACLs.
- ☐ **Block public access to buckets and objects granted through *any* access control lists (ACLs)**
S3 will ignore all ACLs that grant public access to buckets and objects.
- ☐ **Block public access to buckets and objects granted through *new* public bucket or access point policies**
S3 will block new bucket and access point policies that grant public access to buckets and objects. This setting doesn't change any existing policies that allow public access to S3 resources.



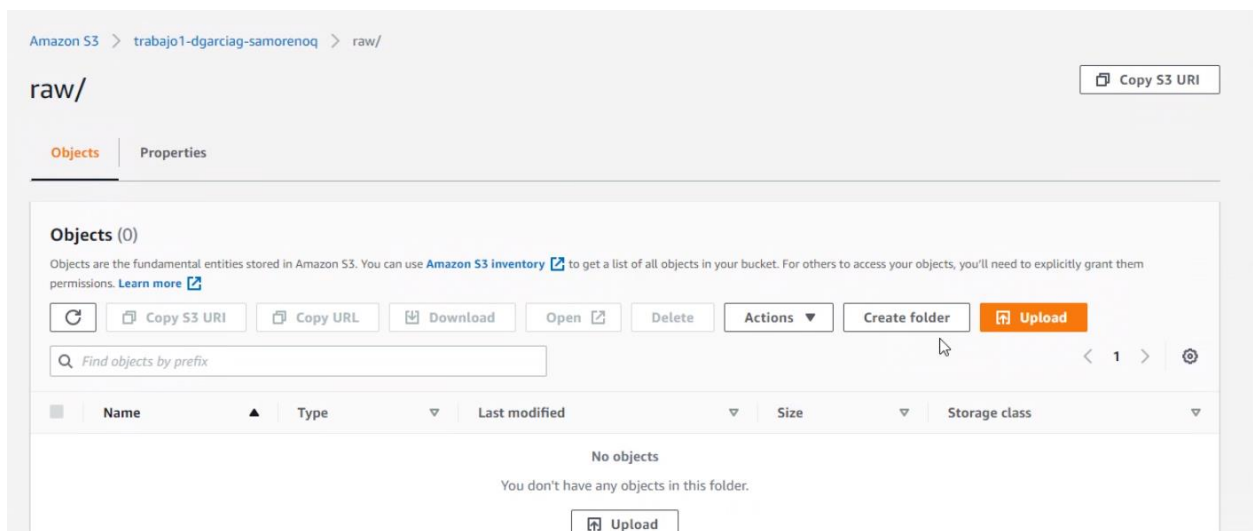
Creamos un bucket en AWS S3 que es el que va funcionar como nuestro datalake para el almacenamiento de los archivos.

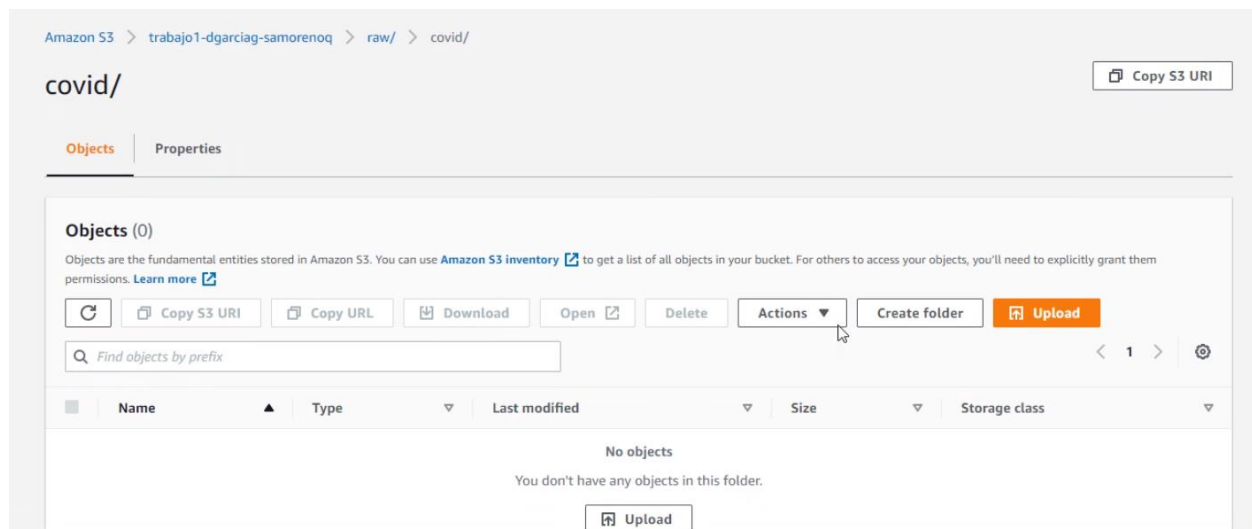
Particionamiento y zonas del datalake





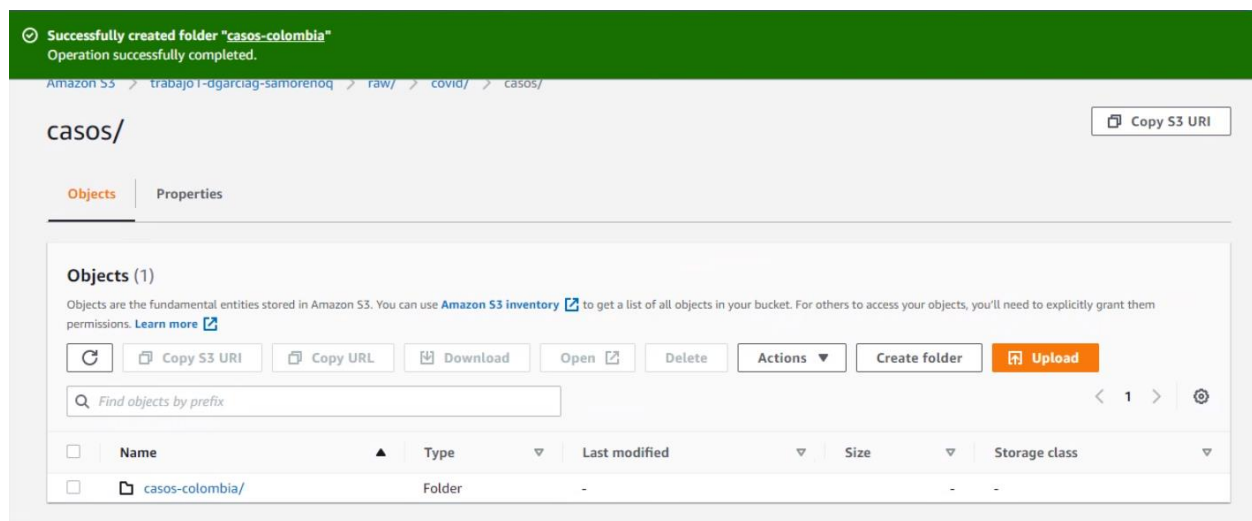
En esta imagen podemos observar las zonas para el datalake donde estarán los diferentes archivos según el proceso ETL que se aplique.





En el directorio de covid creamos las particiones de casos, muertes, vacunaciones y recuperaciones, para su posterior uso en la subida de archivos.

Subida de archivos al datalake



Uploading

Total remaining: 1 file: 806.7 MB(100.00%)

Estimated time remaining: calculating...

Transfer rate: 0 B/s

0%

Cancel

Upload: status

Close

The information below will no longer be available after you navigate away from this page.

Summary

Destination

s3://trabajo1-dgarcia-samorenoq/raw/covid/casos/casos-colombia/

Succeeded

0 files, 0 B (0%)

Failed

0 files, 0 B (0%)

Files and folders

Configuration

OCHA Services

Data Responsibility for COVID-19

FAQ

Log in

Sign up

Switch to HDX Lite

HDX

Search Datasets

DATA | LOCATIONS | ORGANISATIONS | QUICKLINKS

ADD DATA

HOME / DATASETS / CORONAVIRUS (COVID-19) VACCINATIONS

Coronavirus (COVID-19) Vaccinations

This dataset is part of [COVID-19 Pandemic](#)

The map and chart below show the number of COVID-19 vaccination doses administered per 100 people within a given population. Note that this does not measure the total number of people that have been vaccinated (which is usually two doses).

2100+ Downloads

This dataset updates: Live

Contact the contributor

Interactive Data

Data Viz 1

Open fullscreen

Country	Data not available for Dec 2, 2020 Showing closest available data point (Dec 15, 2020)	Cumulative COVID-19 vaccinations per 100 people		
		Aug 25, 2021	Absolute Change	Relative Change
United Kingdom		Aug 24, 2021 132.38	+132.25	+101,731%
United States		109.09	+108.92	+64,071%
Upper middle income	Dec 15, 2020 0.06	101.93	+101.87	+169,783%
Uruguay	Feb 27, 2021 0.01	154.92	+154.91	+1,549,100%
Uzbekistan	Mar 31, 2021 0.00	Aug 13, 2021 36.16	+36.16	
Vanuatu	Jun 1, 2021 0.00	Aug 16, 2021 10.11	+10.11	
Venezuela	Feb 17, 2021 0.00	Aug 17, 2021 16.45	+16.45	
Vietnam	Mar 7, 2021 0.00	Aug 23, 2021 18.13	+18.13	
Wales	Dec 13, 2020 0.26	Aug 24, 2021 142.36	+142.10	+54,654%
Wallis and Futuna	Mar 23, 2021 11.44	Aug 23, 2021 86.04	+74.60	+652%
World	0.00	65.15	+65.15	
Yemen	May 9, 2021 0.06	Jul 27, 2021 1.04	+0.98	+1,633%
Zambia	Apr 14, 2021 0.00	3.04	+3.04	
Zimbabwe	Feb 18, 2021 0.00	26.63	+26.63	

Dec 2, 2020

Aug 25, 2021

Admin0

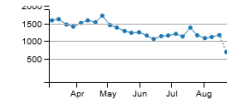
28-Day

Powered by Eari

Weekly

Daily






DOWNLOADS



RELATED SHOWCASES








There are no showcases for this dataset.

ACTIVITY

-  **HDX Metasebya Sahlu** updated the dataset **Novel Coronavirus (COVID-19) Cases Data** 14 hours ago
-  **HDX Metasebya Sahlu** updated the dataset **Novel Coronavirus (COVID-19) Cases Data** 2 days ago
-  **HDX Metasebya Sahlu** updated the dataset **Novel Coronavirus (COVID-19) Cases Data** 3 days ago
-  **HDX Metasebya Sahlu** updated the dataset **Novel Coronavirus (COVID-19) Cases Data** 4 days ago
-  **HDX Metasebya Sahlu** updated the dataset **Novel Coronavirus (COVID-19) Cases Data** 1 week ago

Data and Resources

Metadata

	time_series_covid19_confirmed_global.csv Updated: Live Original wide form (new column for each day)	DOWNLOAD	MORE
	time_series_covid19_deaths_global.csv Updated: Live Original wide form (new column for each day)	DOWNLOAD	MORE
	time_series_covid19_recovered_global.csv Updated: Live Original wide form (new column for each day)	DOWNLOAD	MORE
	time_series_covid19_confirmed_global_iso3_regions.csv Updated: Live Data in wide format with HXL hashtags and ISO3 country codes, region codes and names added	DOWNLOAD	MORE
	time_series_covid19_deaths_global_iso3_regions.csv Updated: Live Data in wide format with HXL hashtags and ISO3 country codes, region codes and names added	DOWNLOAD	MORE
	time_series_covid19_recovered_global_iso3_regions.csv Updated: Live Data in wide format with HXL hashtags and ISO3 country codes, region codes and names added	DOWNLOAD	MORE
	time_series_covid19_confirmed_global_narrow.csv Updated: Live	DOWNLOAD	MORE

Amazon S3 > trabajo1-dgarciag-samorenoq > raw/ > covid/ > casos/ > Create folder

Create folder [Info](#)

Use folders to group objects in buckets. When you create a folder, S3 creates an object using the name that you specify followed by a slash (/). This object then appears as folder on the console. [Learn more](#)



Your bucket policy might block folder creation

If your bucket policy prevents uploading objects without specific tags, metadata, or access control list (ACL) grantees, you will not be able to create a folder using this configuration. Instead, you can use the [upload configuration](#) to upload an empty folder and specify the appropriate settings.

Folder

Folder name

/

Folder names can't contain "/". [See rules for naming](#)

Server-side encryption



The following settings apply only to the new folder object and not to the objects contained within it.

Server-side encryption

☒ Disable

☐ Enable

Cancel

Create folder

Amazon S3 > trabajo1-dgarciag-samorenoq > raw/ > covid/ > vacunaciones/

vacunaciones/

[Copy S3 URI](#)

Objects Properties

Objects (1)

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](#)

Find objects by prefix

< 1 > ⚙

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	COVID-19-Vaccinations.csv	csv	August 26, 2021, 17:16:46 (UTC-05:00)	2.7 MB	Standard

Amazon S3 > trabajo1-dgarciag-samorenoq > raw/ > covid/ > muertes/

muertes/

[Copy S3 URI](#)

Objects Properties

Objects (1)

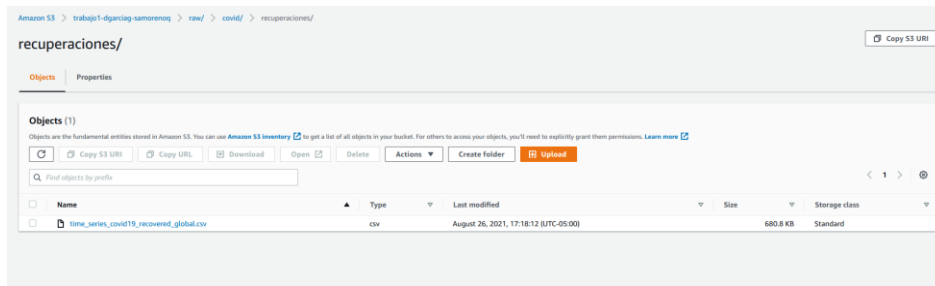
Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

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

Find objects by prefix

< 1 > ⚙

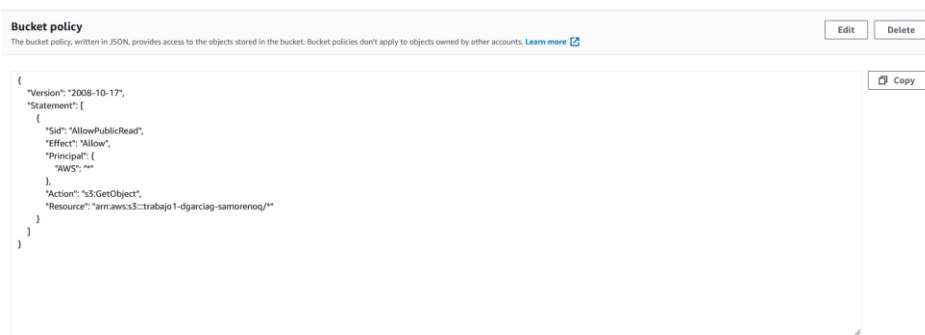
<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	time_series_covid19_deaths_global.csv	csv	August 26, 2021, 17:17:37 (UTC-05:00)	537.4 KB	Standard



En las anteriores imágenes pudimos observar la carga de todos los archivos con los que vamos a trabajar en el datalake.

Cambio de políticas del bucket

Cambio de políticas para el bucket para que sea público:



Creación de primer Crawler

Add information about your crawler

Crawler name

crawler-trabajo1

Tags, description, security configuration, and classifiers (optional)

Tag key

Type tag key...

Tag value

Type tag value...

Description

Crawler para información de casos de Covid-19 en Colombia y a nivel mundial, además de muertes, recuperados y vacunados a nivel mundial.

Security configuration

None

Choose a security configuration to enable at-rest encryption on the logs pushed to CloudWatch.

Classifiers infer the schema of your data. AWS Glue tries to match your data with custom classifiers in the order listed. The first classifier to recognize your data is used. Built-in classifiers are used if you do not supply a classifier that matches.

Custom classifiers

Showing: 0 - 0

Classifier

Classification

No items available

Selected classifiers

Showing: 0 - 0

Classifier

Classification

No classifiers selected.

Add a data store

Choose a data store

S3

Connection

Select a connection

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any future S3 targets will also use the same connection (or none, if left blank).

Add connection

Crawl data in

- ☒ Specified path in my account
☐ Specified path in another account

Include path

s3://trabajo1-dgarciag-samorenoq/raw



All folders and files contained in the include path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Sample size (optional)

Enter an integer between 1 and 249.

This field sets the number of files in each leaf folder to be crawled. If not set, all the files are crawled.

▸ Exclude patterns (optional)

Back

Next

Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. [Learn more](#)

- ☐ Update a policy in an IAM role
☒ Choose an existing IAM role
☐ Create an IAM role

IAM role ⓘ

AWSGlueServiceRole-trabajo1-dgarciag-samorenoq



This role must provide permissions similar to the AWS managed policy, **AWSGlueServiceRole**, plus access to your data stores.

- s3://trabajo1-dgarciag-samorenoq/raw

You can also create an IAM role on the [IAM console](#).

Back

Next

Query de prueba en Athena

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.

Crawler "crawler-trabajo1" is now running.

[User preferences](#)

Add crawler	Run crawler	Action ▾	<input type="text" value="Filter by tags and attributes"/>	Showing: 0 - 0 < > 🔄 🛑				
<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
Loading								

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.

Crawler "crawler-trabajo1" completed and made the following changes: 5 tables created, 0 tables updated. See the tables created in database [trabajo1](#).

✕

[User preferences](#)

Add crawler

Run crawler

Action

Filter by tags and attributes

Showing: 1 - 1

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	crawler-trabajo1		Ready	Logs	1 min	1 min	0	5

- Query inicial en Raw con Athena

Este query trae el número total de personas vacunadas y de personas vacunadas por cada 100 habitantes por país/región y lo ordena en orden descendente, según el número de personas vacunadas por cada 100 habitantes. Es importante destacar que estos números de vacunados son acumulativos en las tablas originales. Esto quiere decir que en cada fecha, se reporta el número total de vacunados a la fecha, no el número de vacunados nuevos en esa fecha, lo cual quiere decir que en cada fecha, los vacunados son los de la fecha anterior más los que se vacunaron en esa fecha y, por este motivo, se toma la cifra máxima de vacunados en la tabla y no la suma ni la cuenta.

New query 1

New query 2

New query 3

+

```

1 SELECT location, MAX(people_fully_vaccinated) AS total_vaccinated, ROUND(MAX(people_fully_vaccinated_per_hundred),2) AS total_vaccinated_per_hundred
2 FROM trabajo1.vacunaciones
3 GROUP BY location
4 ORDER BY total_vaccinated_per_hundred
5 DESC;

```

Run query

Save as

Create

(Run time: 0.92 seconds, Data scanned: 2.69 MB)

Format query

Clear

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Athena engine version 2 [Release versions](#)

Results

	location	total_vaccinated	total_vaccinated_per_hundred
1	Gibraltar	39195	116.34
2	Malta	410532	79.78
3	Iceland	262924	77.05
4	Pitcairn	36	76.6
5	United Arab Emirates	7397933	74.8
6	Cayman Islands	49086	74.69
7	Singapore	4348604	74.33
8	Portugal	7427602	72.84

Procesos ETL y Crawlers con Glue

- Glue ETL

AWS Glue Studio > Jobs

Jobs

Create job

Visual with a source and target

Start with a source, ApplyMapping transform, and target.

Visual with a blank canvas

Author using an interactive visual interface.

Spark script editor

Write or upload your own Spark code.

Python Shell script editor

Write or upload your own Python shell script.

Source

Amazon S3

JSON, CSV, or Parquet files stored in S3.

→

Target

Amazon S3

S3 bucket by specifying a bucket path as the data target.

Your jobs (0)

Find jobs

Job name

Type

Last modified

No jobs

You have not created a job yet.

Create job from a blank graph

Quitar espacios de columnas

Unsaved job found
We found an unsaved graph, do you wish to restore it?

Visual Script Job details Runs Schedules

Source Transform Target Undo Redo Remove

Node properties Data source properties - S3 Output schema Data preview

S3 source type info

☐ Data Catalog table

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

S3 URL
s3://trabajo1-dgarciag-samorenou/raaw/covid/casos/casos-colombia/ View Browse S3

☒ 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

Quitar espacios de columnas

Unsaved job found
We found an unsaved graph, do you wish to restore it?

Visual Script Job details Runs Schedules

Source Transform Target Undo Redo Remove

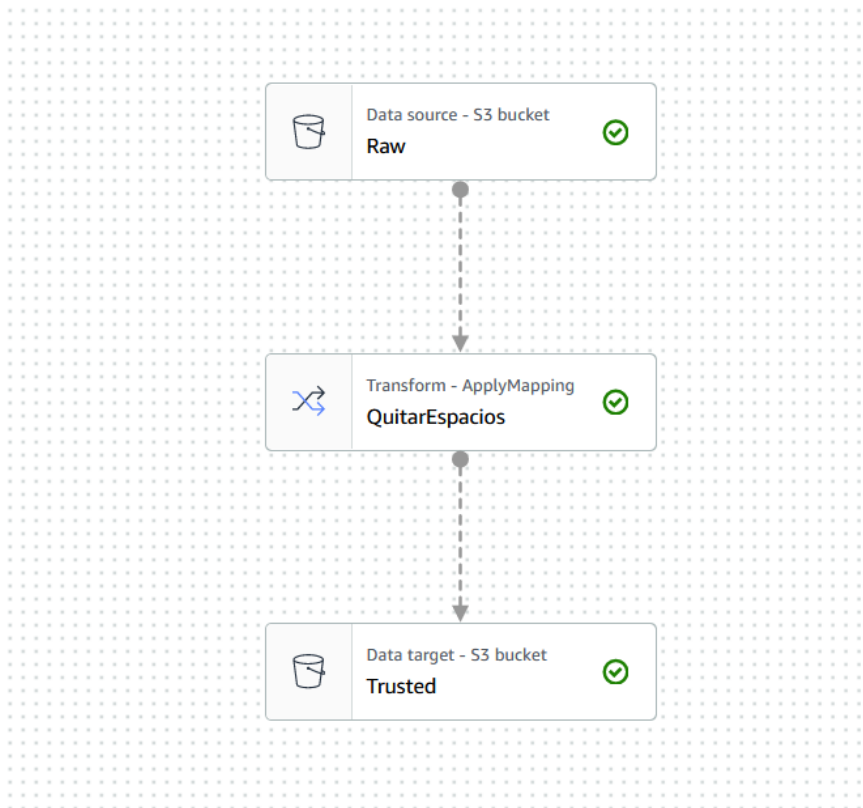
Node properties Transform Output schema Data preview

Name
QuitarEspacios

Node type
Choose which type of node to add to the job.
ApplyMapping
Map fields to new names and types of your choice.

Node parents
Choose which nodes will provide inputs for this one.
Select parents
S3 bucket S3 - DataSource

Se quitaron todos los espacios y las tildes en esta transformación.



Se guardan los datos transformados en la zona “Trusted”.

Quitar espacios de columnas

[Visual](#) | [Script](#) | **[Job details](#)** | [Runs](#) | [Schedules](#)

Basic properties [Info](#)

Name

Quitar espacios de columnas

Description - optional

Descriptions can be up to 2048 characters long.

IAM Role

Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.

AWSGlueServiceRole-trabajo1

Type

The type of ETL job. This is set automatically based on the types of data sources you have selected.

Spark

Glue version [Info](#)

Glue 2.0 - Supports spark 2.4, Scala 2, Python 3

Language

Python 3

Worker type

Set the type of predefined worker that is allowed when a job runs.

G.1X

Propiedades del Job.

Roles > AWSGlueServiceRole-trabajo1

Summary Delete role

Role ARN	arn:aws:iam::069773076114:role/service-role/AWSGlueServiceRole-trabajo1
Role description	Edit
Instance Profile ARNs	
Path	/service-role/
Creation time	2021-08-26 17:20 EST
Last activity	Not accessed in the tracking period
Maximum session duration	1 hour Edit

Permissions Trust relationships Tags Access Advisor Revoke sessions

▼ Permissions policies (3 policies applied)

[Attach policies](#) [Add inline policy](#)

Policy name	Policy type	
AmazonS3FullAccess	AWS managed policy	✕
AWSGlueServiceRole	AWS managed policy	✕
AWSGlueServiceRole-trabajo1	Managed policy	✕

Se agregó AmazonS3FullAccess al rol IAM de Glue.

Quitar espacios de columnas Save Delete Run

🟢 Successfully started job
Successfully started job Quitar espacios de columnas. Navigate to [Run Details](#) for more details.

Visual Script Job details **Runs** Schedules

Recent job runs (9) [Info](#) 🔄 Stop job run

August 29, 2021 1:10 AM

Id
[j_ccd5d44e39706c62b9cedc2c73e8f3277e7f04fcec35d71aa6fa71a3604aa9](#)

Recent attempt
-

Run status
🟢 Succeeded

Glue version
2.0

Logs
[Cloudwatch logs](#)

Output Logs
[Cloudwatch output logs](#)

Error logs
[Cloudwatch error logs](#)

Execution time
2 minutes

Start time
August 29, 2021 1:10 AM

End time
August 29, 2021 1:12 AM

Corrida del Job exitosa.

Crawlers > crawler-trabajo1-trusted

[Run crawler](#) [Edit](#)

Name	crawler-trabajo1-trusted
Description	Create a single schema for each S3 path
Table level	false
Security configuration	
Tags	-
State	Ready
Schedule	
Last updated	Sun Aug 29 01:17:41 GMT-500 2021
Date created	Sun Aug 29 01:17:41 GMT-500 2021
Database	trabajo1-trusted
Service role	service-role/AWSGlueServiceRole-trabajo1
Selected classifiers	
Data store	S3
Include path	s3://trabajo1-dgarcia-samorenog/trusted
Connection	
Exclude patterns	

Configuration options

Schema updates in the data store	Update the table definition in the data catalog.
Object deletion in the data store	Mark the table as deprecated in the data catalog.

Nuevo crawler para la zona Trusted.

aws Services Search for services, features, marketplace products, and docs [Alt+S] vocstartsoft/user=admin @ 0697-7307-6114 N. Virginia Support

Untitled job Job has not been saved Save Delete Run

Visual Script Job details Runs Schedules

Source Transform Target Undo Redo Remove

Node properties Transform Output schema Data preview

DropFields

Field	Data type
<input type="checkbox"/> edad	long
<input checked="" type="checkbox"/> unidad_de_medida_de_edad	long
<input type="checkbox"/> sexo	string
<input checked="" type="checkbox"/> tipo_de_contagio	string
<input type="checkbox"/> ubicacion_del_caso	string
<input type="checkbox"/> estado	string

En un nuevo job eliminamos algunas columnas que para efectos de este trabajo son innecesarias y no nos brindan mucha información. Vamos a recibir la nueva tabla que tenemos en la zona trusted, hacemos la transformación y almacenamos en la zona refined.

aws Servicios Buscar servicios, características, productos del Marketplace y documentos [Alt+S] vocstartsoft/user=admin @ 0697-7307-6114 Norte de Virginia

AWS Glue Studio

Jobs

- Monitoring
- Connectors
- What's new
- Glue console
 - Glue catalog
 - Crawlers
 - Security configurations
- Marketplace
- Documentation

Visual with a source and target Start with a source, ApplyMapping transform, and target.

Visual with a blank canvas Author using an interactive visual interface.

Spark script editor Write or upload your own Spark code.

Python Shell script editor Write or upload your own Python shell script.

Source Amazon S3 JSON, CSV, or Parquet files stored in S3.

Target Amazon S3 S3 bucket by specifying a bucket path as the data t...

Your jobs (2) Info

Find jobs

Job name	Type	Last modified
Quitar columnas	Glue ETL	30/8/2021 1:50:04
Quitar espacios de columnas	Glue ETL	29/8/2021 1:06:22

Ahora tenemos 2 procesos ETL.

Rastreadores > crawler_trabajo1refined

Ejecutar rastreador

Editar

Nombre	crawler_trabajo1refined
Descripción	
Cree un único esquema para cada ruta de S3	false
Table level	
Configuración de seguridad	
Etiquetas	-
Estado	Ready
Programación	
Last updated	Mon Aug 30 01:54:56 GMT-500 2021
Date created	Mon Aug 30 01:54:56 GMT-500 2021
Base de datos	trabajo1refined
Rol de servicio	service-role/AWSGlueServiceRole-trabajo1
Clasificadores seleccionados	
Almacén de datos	S3
Ruta de inclusión	s3://trabajo1-dgarcia-g-samorenog/refined
Connection	
Patrones de exclusión	

Crawler para agregar la nueva tabla a la zona refined.

Rastreadores

Un rastreador se conecta a un almacén de datos, avanza por una lista de clasificadores ordenados por prioridad para determinar el esquema de sus datos y, a continuación, crea tablas de metadatos en el catálogo de datos.

[Preferencias del usuario](#)

Añadir un rastreador

Ejecutar rastreador

Acción

🔍 Filtrar por etiquetas y atributos

Mostrando: 1 - 3 < > ↺ ⓘ

<input type="checkbox"/>	Nombre	Programación	Estado	Registros	Última ejecución	Tiempo de ejecución medio	Tablas actualizadas	Tablas añadidas
<input type="checkbox"/>	crawler-trabajo1		Ready	Registros	1 minuto	1 minuto	0	5
<input type="checkbox"/>	crawler-trabajo1-trusted		Ready	Registros	1 minuto	1 minuto	0	1
<input type="checkbox"/>	crawler_trabajo1refined		Ready	Registros	40 segundos	40 segundos	0	1

Tenemos 3 crawlers en total.

Data source

Connect data source

AwsDataCatalog

Database

trabajo1-trusted

Filter tables and views...

▼ Tables (1)

Create table

▼ trusted

:

fecha_reporte_web (string)

id_de_caso (bigint)

fecha_de_notificacion (string)

codigo_divipola_departamento (bigint)

nombre_departamento (string)

codigo_divipola_municipio (bigint)

nombre_municipio (string)

edad (bigint)

unidad_de_medida_de_edad (bigint)

sexo (string)

tipo_de_contagio (string)

ubicacion_del_caso (string)

estado (string)

codigo_iso_del_pais (bigint)

nombre_del_pais (string)

recuperado (string)

fecha_de_inicio_de_sintomas (string)

fecha_de_muerte (string)

fecha_de_diagnostico (string)

fecha_de_recuperacion (string)

tipo_de_recuperacion (string)

pertenencia_etnica (bigint)

nombre _del_grupo_etnico (string)

Resultado después del proceso de ETL con Glue y el crawler en trusted: ya las columnas están bien nombradas.



Resultado después del proceso de ETL con Glue y el crawler en refined: ya las columnas están bien nombradas y no tenemos columnas innecesarias como unidad de medida de edad, código iso del país, nombre del país (En cualquiera de los casos estamos hablando de Colombia), tipo de recuperación, etc.

Consultas en Athena

New query 1 New query 2

```

1 SELECT nombre_municipio, count(*) as num_casos
2 FROM "trabajol-trusted".trusted
3 GROUP BY nombre_municipio
4 ORDER BY num_casos
5 DESC
6

```

Run query Save as Create (Run time: 2.11 seconds, Data scanned: 859.31 MB) Format query Clear

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete Athena engine version 2 Release versions

Results

	nombre_municipio	num_casos
1	BOGOTA	1435900
2	MEDELLIN	397742
3	CALI	273082
4	BARRANQUILLA	197265
5	CARTAGENA	121076
6	BUCARAMANGA	106387
7	IBAGUE	74478
8	MANIZALES	68771
9	"SANTA MARTA"	64133

Query en Athena usando los nuevos nombres de columnas.

New query 1 New query 4

```

1 SELECT edad, count(*) as num_casos
2 FROM refined
3 GROUP BY edad
4 ORDER BY num_casos
5 DESC

```

Run query Save as Create (Run time: 2.31 seconds, Data scanned: 751.78 MB) Format query Clear

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete Athena engine version 2 Release versions

Results

	edad	num_casos
1	30	120032
2	28	118989
3	29	118908
4	27	117312
5	26	117082
6	31	116988

Query en Athena a nuestro archivo en la zona refined, donde podemos observar las edades donde se ha presentado un número mayor de contagios.

Creación del EMR

Elegir las herramientas adecuadas para el cluster. (Nota: en este caso no se usarán todas, pero igualmente es útil porque quedará como plantilla para trabajos futuros en los que sí se usen Spark, notebooks, etc.)

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

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

Software Configuration

Release **emr-6.3.0**

<input checked="" type="checkbox"/> Hadoop 3.2.1	<input checked="" type="checkbox"/> Zeppelin 0.9.0	<input checked="" type="checkbox"/> Livy 0.7.0
<input checked="" type="checkbox"/> JupyterHub 1.2.0	<input checked="" type="checkbox"/> Tez 0.9.2	<input type="checkbox"/> Flink 1.12.1
<input type="checkbox"/> Ganglia 3.7.2	<input type="checkbox"/> HBase 2.2.6	<input type="checkbox"/> Pig 0.17.0
<input checked="" type="checkbox"/> Hive 3.1.2	<input type="checkbox"/> Presto 0.245.1	<input type="checkbox"/> PrestoSQL 350
<input type="checkbox"/> ZooKeeper 3.4.14	<input checked="" type="checkbox"/> JupyterEnterpriseGateway 2.1.0	<input type="checkbox"/> MXNet 1.7.0
<input type="checkbox"/> Sqoop 1.4.7	<input checked="" type="checkbox"/> Hue 4.9.0	<input type="checkbox"/> Phoenix 5.0.0
<input type="checkbox"/> Oozie 5.2.1	<input checked="" type="checkbox"/> Spark 3.1.1	<input checked="" type="checkbox"/> HCatalog 3.1.2
<input type="checkbox"/> TensorFlow 2.4.1		

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

Edit software settings

☒ Enter configuration ☐ Load JSON from S3

```
classification=config-file-name,properties={myKey1=myValue1,myKey2=myValue2}
```

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](#)

Se escogen los tipos de nodos, que en este caso serán m4.xlarge y se pondrán en Spot para que salga más barato.

Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. [Learn more about instance purchasing options](#)

Console options for automatic scaling have changed. [Learn more](#)

Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m4.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	1 Instances	<input type="radio"/> On-demand <input checked="" type="radio"/> Spot Use on-demand as max price
Core Core - 2	m4.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	2 Instances	<input type="radio"/> On-demand <input checked="" type="radio"/> Spot Use on-demand as max price

- Nombre del cluster

General Options

Cluster name **emr-trabajo1**

☒ Logging
S3 folder **s3://aws-logs-069773076114-us-east-1/elasticmapred**

☐ Log encryption

☒ Debugging

☒ Termination protection

- Key-pair del cluster

Security Options

EC2 key pair

lab1

☒ 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

EC2 security groups

Cancel

Previous

Create cluster

- Se debe abrir el puerto 8888 para tener acceso a Hue para hacer las consultas de Hive.

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
HDFS Name Node	http://ec2-18-207-252-214.compute-1.amazonaws.com:9870/	SSH tunnel not enabled
Hue	http://ec2-18-207-252-214.compute-1.amazonaws.com:8888/	SSH tunnel not enabled
JupyterHub	https://ec2-18-207-252-214.compute-1.amazonaws.com:9443/	SSH tunnel not enabled
Zeppelin	http://ec2-18-207-252-214.compute-1.amazonaws.com:8890/	SSH tunnel not enabled
Tez UI	http://ec2-18-207-252-214.compute-1.amazonaws.com:8080/tez-ui	SSH tunnel not enabled
Spark History Server	http://ec2-18-207-252-214.compute-1.amazonaws.com:18080/	SSH tunnel not enabled
Livy	http://ec2-18-207-252-214.compute-1.amazonaws.com:8998/	SSH tunnel not enabled
Resource Manager	http://ec2-18-207-252-214.compute-1.amazonaws.com:8088/	SSH tunnel not enabled

The following table lists web interfaces you can view on the task nodes:

- En security groups, se le cambiarán las inbound rules al nodo maestro del EMR para abrir el puerto 8888 y permitir acceso desde cualquier dirección IP.

Security Groups (3)

Filter security groups

Actions

Create security group

Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rules count	Outbound rules count
-	sg-04730ca3696c4a733	ElasticMapReduce-slave	vpc-500a742d	Slave group for Elastic ...	069773076114	6 Permission entries	1 Permission entry
-	sg-0ee33ac1d3d0e2804	ElasticMapReduce-mas...	vpc-500a742d	Master group for Elasti...	069773076114	19 Permission entries	1 Permission entry
-	sg-a844afb9	default	vpc-500a742d	default VPC security gr...	069773076114	1 Permission entry	1 Permission entry

sgr-09ba7885528a34476

All ICMP - IPv4

ICMP

All

Custom

0.0.0.0/0

069773076114/sg-0ee33ac1d3d0e2804

Delete

sgr-09d4ac4c57749632a

Custom TCP

TCP

8443

Custom

54.239.98.0/24

069773076114/sg-04730ca3696c4a733

Delete

sgr-09796d4bb87ac7a39

All UDP

UDP

0 - 65535

Custom

207.171.167.101/32

0.0.0.0/0

Delete

sgr-0e14a7c4a6ca242f0

Custom TCP

TCP

8443

Custom

Delete

-

Custom TCP

TCP

8888

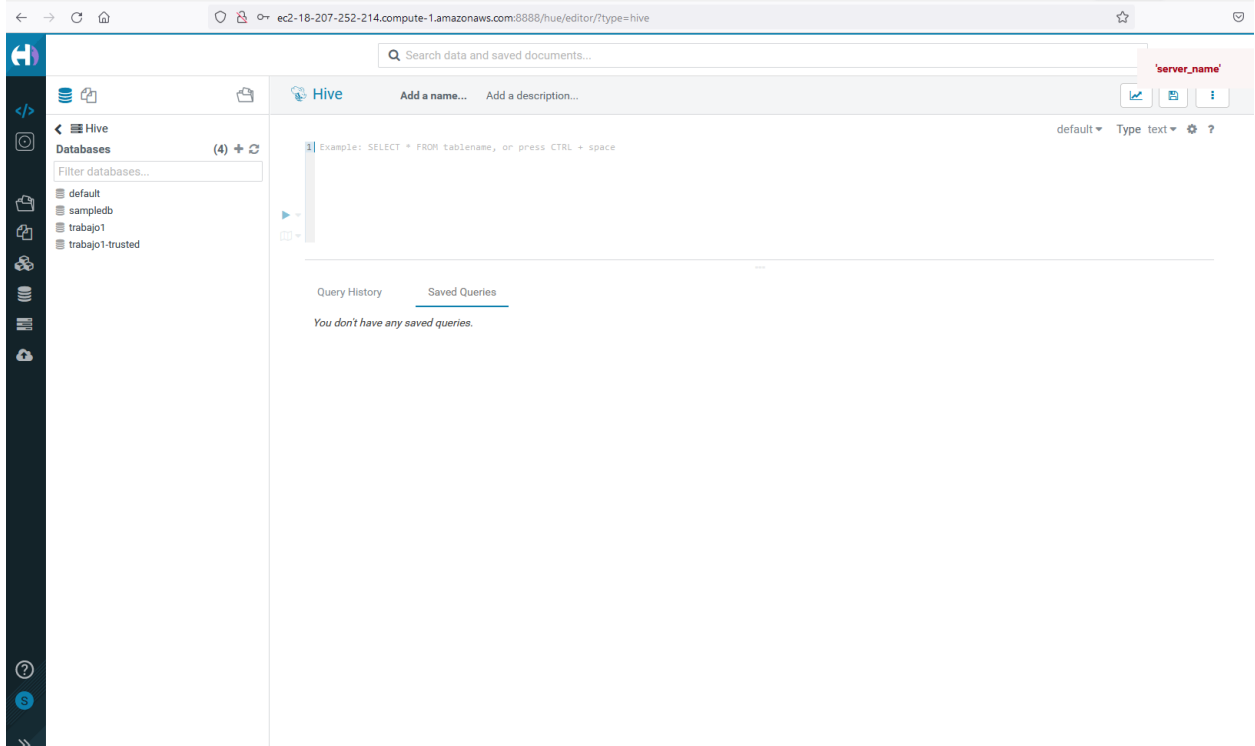
Anywhere...

Delete

Add rule

- Ya se tiene acceso a la interfaz gráfica de Hue y se ven las bases de datos creadas por los crawlers.

Consultas con Hive



- Ejemplos de consultas en Hive, tanto en la base de datos original como en la modificada en trusted.

8.72s trabajo1 Type text ?

```

1 SELECT 'country/region',
2 ('1/22/20' + '1/23/20' + '1/24/20' + '1/25/20' + '1/26/20' + '1/27/20' + '1/28/20' + '1/29/20' + '1/30/20' + '1/31/20' + '2/1/20' + '2/2/20' + '2/3/20' + '2/4/20' + '2/5/20' + '2/6/20' + '2/7/20' +
3 AS total_muertes
4 FROM muertes
5 ORDER BY total_muertes
6 DESC;

```

INFO : Session is already open
 INFO : Dag name: SELECT 'country/region'...total_muertes
 DESC (Stage-1)
 Status: Running (Executing on YARN cluster with App id application_1630256648670_0002)

country/region	total_muertes
1 US	177128675
2 Brazil	119217084
3 India	79879911
4 Mexico	63885533
5 Peru	52358276
6 United Kingdom	40227081
7 Italy	37237099
8 France	32282679
9 Russia	31003685
10 Spain	26217547
11 Colombia	23717194
12 Iran	23498114
13 Germany	21370857

Esta consulta en Hive muestra el número total de muertes por país en orden descendente.

52.25s trabajo1-trusted Type text ?

```

1 SELECT tipo_de_contagio, COUNT(*) as total_casos
2 FROM trusted
3 GROUP BY tipo_de_contagio
4 ORDER BY total_casos
5 DESC;

```

INFO : Compiling command(queryId=hive_20218829194234_64d88b9-48e2-4352-a226-b125f7b34cf7): SELECT tipo_de_contagio, COUNT(*) as total_ca
 FROM trusted
 GROUP BY tipo_de_contagio
 ORDER BY total_casos

tipo_de_contagio	total_casos
1 "En estudio"	3182451
2 Comunitaria	981596
3 Relacionado	716794
4 Importado	3101

Esta consulta en Hive muestra el número total de casos por cada tipo de contagio. La consulta fue hecha sobre los datos en la zona Trusted, que se obtuvieron tras quitar los espacios en los nombres de las columnas.

Search data and saved documents...

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

trabajo1refined (1) +

Tables Filter... refined

```

1 SELECT recuperado, count(*) as num_casos
2 FROM refined
3 GROUP BY recuperado
4 ORDER BY num_casos
5 DESC;
6

```

INFO : Session is already open
INFO : Dag name: SELECT recuperado, count(*)...num_casos
DESC (Stage-1)

Query History Saved Queries Results (5)

	recuperado	num_casos
1	Recuperado	4710952
2	Fallecido	123728
3	Activo	34530
4	N/A	14427

Tables Filter...
trabajo1refined.refined
fecha_reporte_web string
id_de_caso bigint
fecha_de_notificacion string
codigo_divipola_departam... bigint
nombre_departamento string
codigo_divipola_municipio bigint
nombre_municipio string
edad bigint
sexo string
ubicacion_del_caso string
estado string
recuperado string
fecha_de_inicio_de_sinto... string
fecha_de_muerte string
fecha_de_diagnostico string
fecha_de_recuperacion string
pertenencia_etnica bigint
nombre_del_grupo_etnico string

Esta consulta se realizó sobre la zona refined, acá podemos observar la cantidad de casos de covid19 que están recuperados, que llevaron a la muerte o que aún están activos, de igual manera, podemos observar que la mayoría de los casos se han recuperado y que hay un número considerable de fallecidos.

Search data and saved documents...

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

trabajo1refined (1) +

Tables Filter... refined

```

1 SELECT fecha_de_muerte, count(*) as num_casos
2 FROM refined
3 GROUP BY fecha_de_muerte
4 ORDER BY num_casos
5 DESC;
6

```

INFO : Compiling command(queryId=hive_20210630131036_186c930d-250f-4f56-8141-fab83ea2cc85): SELECT fecha...
FROM refined
GROUP BY fecha_de_muerte
ORDER BY num_casos
DESC

Query History Saved Queries Results (100+)

	fecha_de_muerte	num_casos
1		4742868
2	"21/6/2021 0:00:00"	711
3	"24/6/2021 0:00:00"	683
4	"23/6/2021 0:00:00"	674
5	"20/6/2021 0:00:00"	673
6	"14/6/2021 0:00:00"	672
7	"29/6/2021 0:00:00"	672
8	"13/6/2021 0:00:00"	670
9	"17/6/2021 0:00:00"	667

Tables Filter...
trabajo1refined.refined
fecha_reporte_web string
id_de_caso bigint
fecha_de_notificacion string
codigo_divipola_departam... bigint
nombre_departamento string
codigo_divipola_municipio bigint
nombre_municipio string
edad bigint
sexo string
ubicacion_del_caso string
estado string
recuperado string
fecha_de_inicio_de_sinto... string
fecha_de_muerte string
fecha_de_diagnostico string
fecha_de_recuperacion string
pertenencia_etnica bigint
nombre_del_grupo_etnico string

Para esta consulta el objetivo era observar las fechas con el mayor número de muertes, podemos ver que el 21 de junio fue la fecha de más muertes, encontrando también que las demás fechas también se encuentran en el mes de junio, por lo que fue el mes más complicado en cuanto a muertes. Esta consulta se realizó también en la zona refined.

