

Manuela Ramos Ospina, Camila Acosta,
John Zapata, Dany Palacio
UNIVERSIDAD EAFIT

3/12/2024

Trabajo 1

ST1800 Almacenamiento y
Recuperación de Información

Despliegue de un Lakehouse que integre datos en S3 (*datalake*) y el *data warehouse RedShift* con procesamiento en *Hadoop/Spark* de *AWS EMR, Glue, Athena* y *Redshift (Spectrum y ML)*

Ciclo de vida del proyecto con datos abiertos del área metropolitana

Las actividades humanas derivadas del uso insostenible de la energía, la distribución y uso de la tierra, el acelerado crecimiento urbano, los patrones de consumo y producción han hecho una contribución histórica sinigual a la emisión de gases de efecto invernadero. En particular en el Valle de Aburrá, se cruzan varios factores que causan su catalogación como ‘área fuente de contaminación’, como sus características topográficas y climatológicas. En recientes días ha sido evidente el aumento del nivel de contaminación en la cuenca, lo que nos hace preguntarnos aún más por este fenómeno y su repercusión en la salud humana. **(Referencias:** Informe de síntesis AR6: Cambio Climático 2023, consultado de: <https://www.ipcc.ch/report/ar6/syr/>. Área Metropolitana Valle de Aburrá, Calidad del Aire, consultado de: <https://www.metropol.gov.co/ambiental/calidad-del-aire>. Área Metropolitana Valle de Aburrá, Condiciones especiales del valle de Aburrá, consultado de: <https://www.metropol.gov.co/ambientales/calidad-del-aire/generalidades/condiciones-especiales>)

Fuente de datos

Es por lo anteriormente mencionado que para el desarrollo del presente trabajo se decide estudiar los datos del Portal de Datos Abiertos del Área Metropolitana del Valle de Aburrá (<https://datosabiertos.metropol.gov.co/>), escogiendo inicialmente tres bases de datos:



Cálculo de casos de morbilidad asociados a contaminación del aire

<https://datosabiertos.metropol.gov.co/dataset/7281fa8e-c666-4fed-a8bd-a4af7d0ea749>

“Contiene el número de muertes y enfermedades según causas asociadas a la contaminación del aire, calculadas a partir de los datos de defunciones del sistema nacional SISPRO y cálculos realizados por el Sistema de Vigilancia en Salud Ambiental, con énfasis en calidad del aire (SIVISA)”



Cálculo tasas de mortalidad de enfermedades respiratorias y circulatorias

<https://datosabiertos.metropol.gov.co/dataset/863158fe-8b9e-485c-be99-c55d76d6f15b>

“Tasas de mortalidad según enfermedades respiratorias y circulatorias calculadas en el proyecto "Calidad del aire y sus efectos en la salud de la población de los diez municipios del valle de aburrá, 2008 - 2015””



Resultados del análisis de riesgo por contaminación del aire

<https://datosabiertos.metropol.gov.co/dataset/7382b0c0-cdbd-4d36-80e0-f9bc0eb4163a>

“Resultados del análisis de riesgo calculados en el Proyecto "Calidad del aire y sus efectos en la salud de la población de los diez municipios del valle de aburrá, 2008 - 2015””

Es de resaltar que todas las bases de datos se constituyen de archivos en formato de valores separado por comas (CSV).

Ingesta de datos

Como paso previo para la ingesta de datos se realiza la generación de un *bucket* en AWS S3 con acceso público (lakehouse). En la Figura 1 se observa su creación con la cuenta de Manuela Ramos.

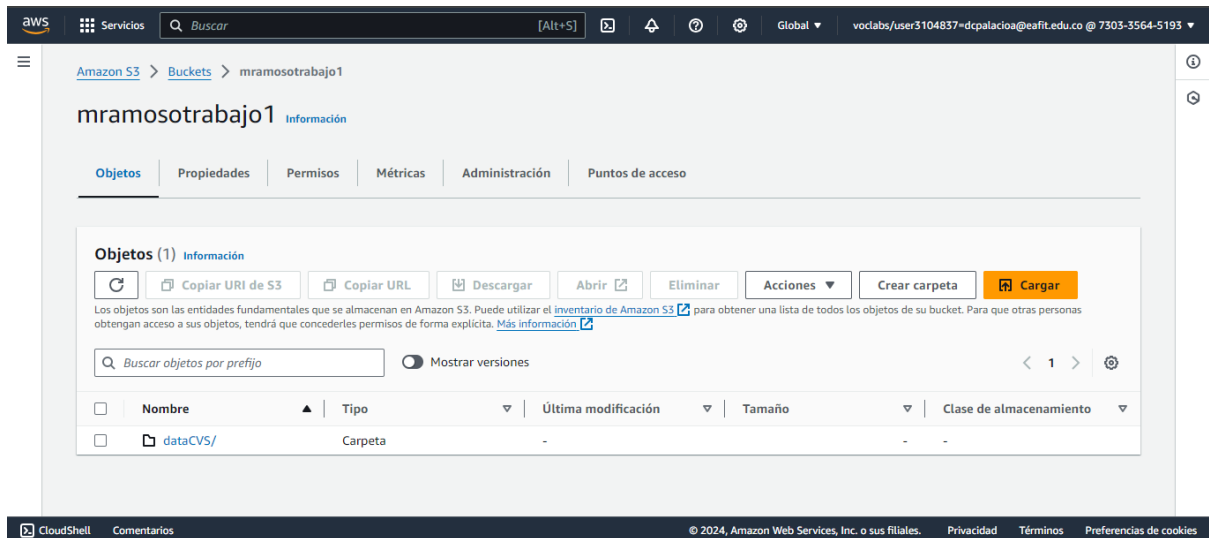


Figura 1. Generación del *bucket* en S3 en el que se depositaron los datasets iniciales

Acto seguido se realiza la creación de un script (ver Figura 2) que permite conectar con la URL de los sitios donde están ubicados los datos con el fin de extraer los datasets y posteriormente migrar los archivos al *bucket* previamente creado, y de esta forma, quedan disponibles para el ecosistema de *Amazon Web Services* (AWS).

Este script fue generado y corrido en *AWS Glue*, bajo la siguiente secuencia:

AWS Glue > ETL jobs > Visual ETL > Script editor > Subir el script de python ETL_Glue.py y ejecutar.

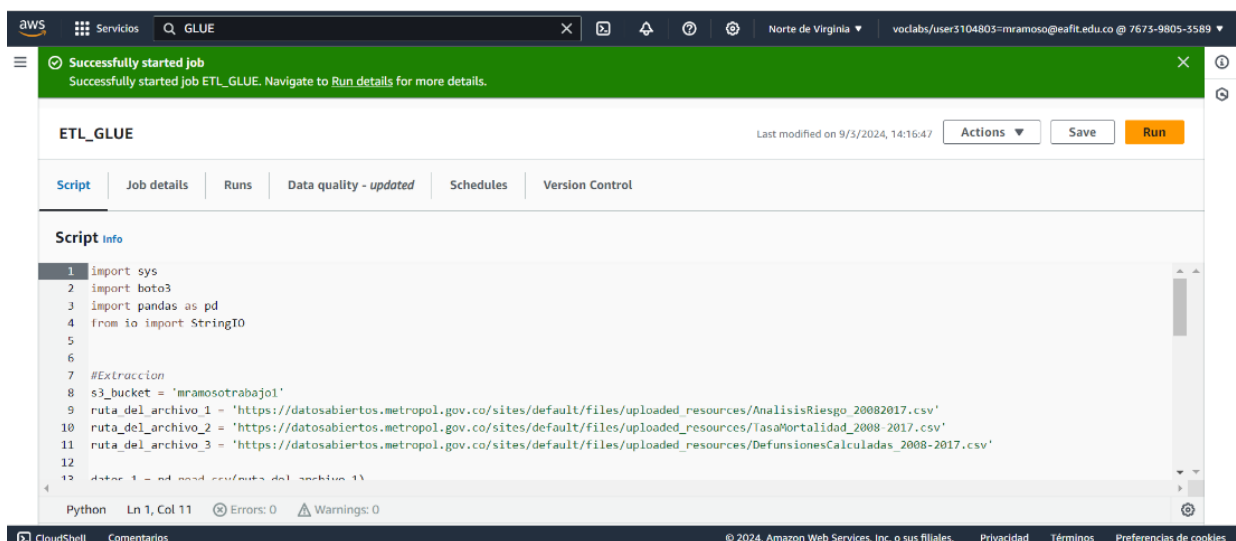


Figura 2. Parte del código del script ETL_Glue.py en AWS GLUE

Almacenamiento de los datos

El script anteriormente mencionado crea una carpeta para cada archivo en su proceso de migración de datos, para posteriormente garantizar una correcta lectura de los archivos en *AWS Glue*. En la Figura 3 se evidencia el almacenamiento de los tres (3) datasets en el *bucket*, cada uno en carpetas separadas, cuyos nombres fueron asignados de manera similar a los datasets con el fin de evitar una posible confusión cuando se procesen. Estas tres (3) carpetas fueron: “Análisis_Riesgo”, “Defunciones” y “Tasa_Mortalidad”.

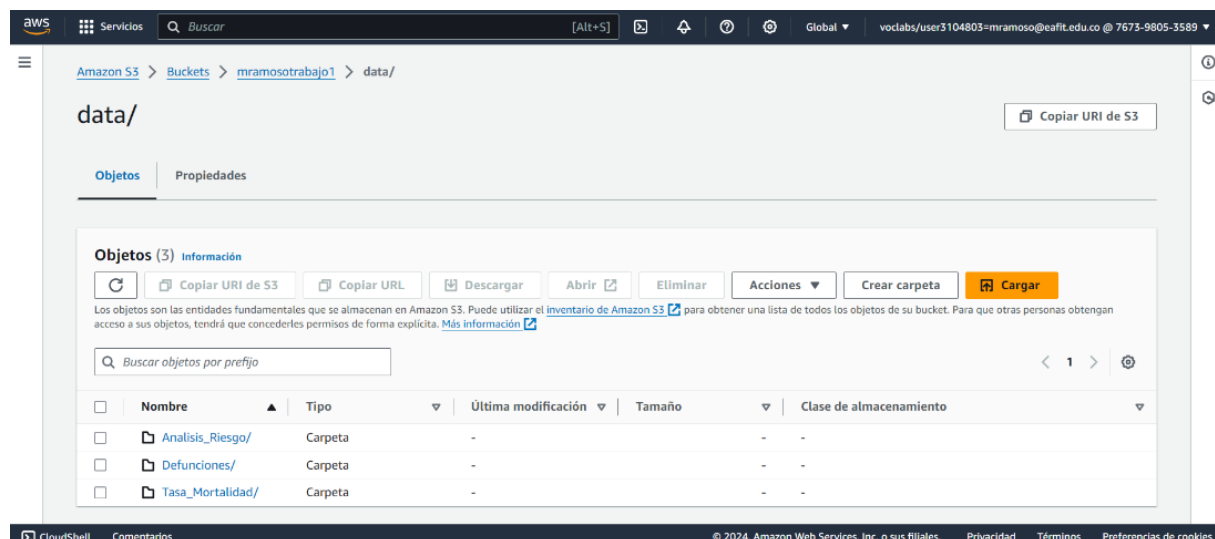


Figura 3. Almacenamiento de datos en AWS S3

Con el fin de realizar un proceso completo ETL se creó un Crawler en AWS Glue (ver Figura 4) para la estructuración de las tablas de los tres (3) archivos depositados en el bucket. De esta manera, quedarán disponibles para su análisis.

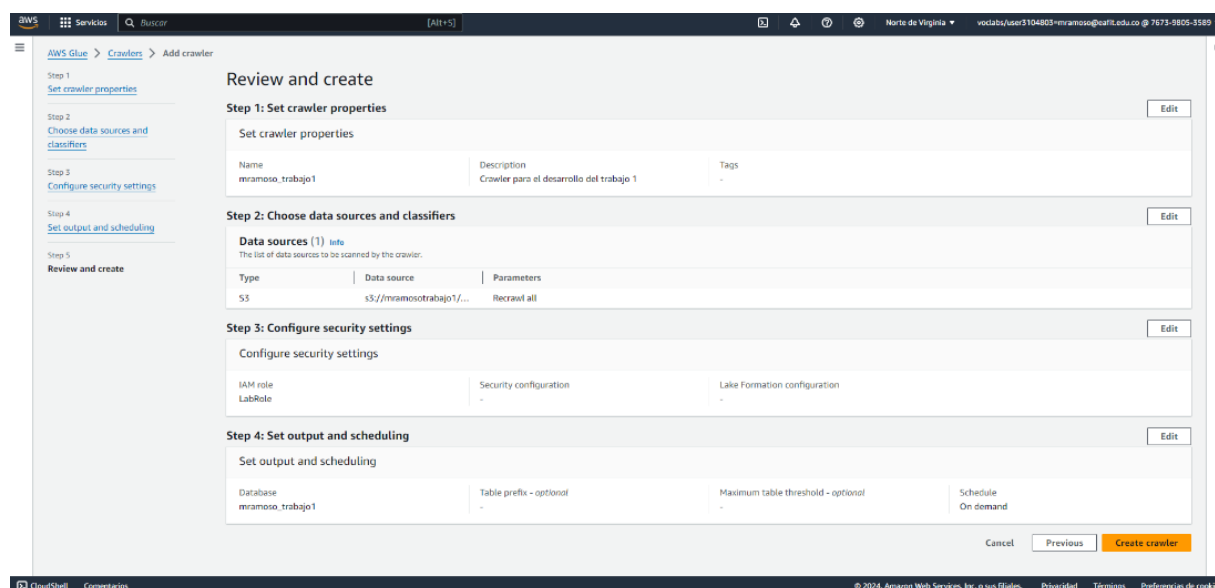


Figura 4. Creación del Crawler

En la Figura 5 se muestran las tres (3) tablas estructuradas en formato csv por AWS Glue, listas para ser analizadas y consultadas usando, por ejemplo, HUE. Estas son: “análisis_riesgo”, “tasa_mortalidad” y “defunciones”.

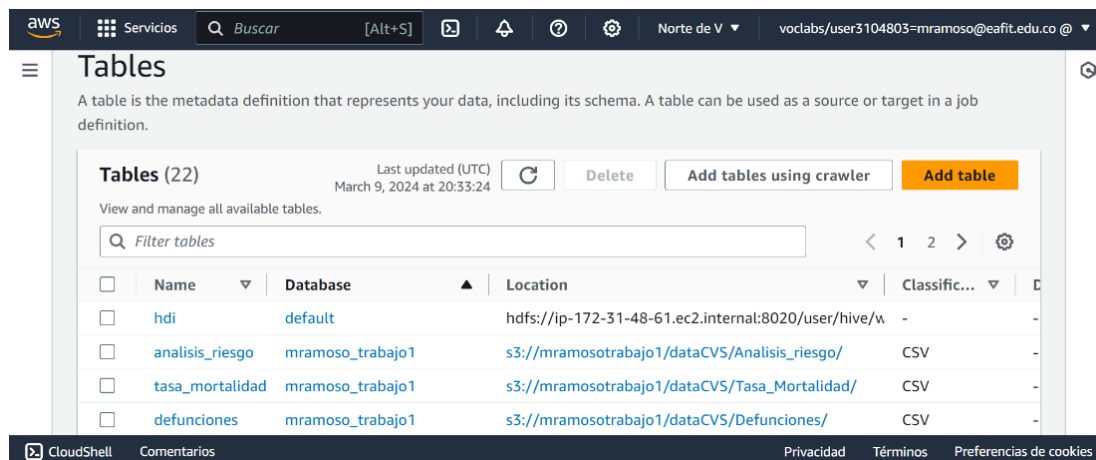


Figura 5. Creación de las tablas en AWS Glue

Para tener mejor noción de la estructura de las tablas crudas, en las Tablas 1-3 se muestra el nombre de las columnas de cada tabla, junto con su tipo de datos.

Tabla 1. Estructura de la tabla análisis_riesgo

#	Nombre de la columna	Tipo de datos
1	tipo de rezago	string
2	número de modelo	bigint
3	municipio	string
4	grupo diagnóstico	string
5	nombre contaminante	string
6	sexo	string
7	edad	string
8	casos	bigint
9	rezago	string
10	beta del contaminante	string
11	exponencial del beta	string
12	limite inferior del exponencial del beta	string
13	limite superior del exponencial del beta	string
14	porcentaje del incremento del riesgo	string

Tabla 2. Estructura de la tabla tasa_mortalidad

#	Nombre de la columna	Tipo de datos
1	año	bigint
2	municipio	string
3	capítulo	string
4	sexo	string
5	tasaajustada	string

Tabla 3. Estructura de la tabla defunciones

#	Nombre de la columna	Tipo de datos
1	municipio	string
2	periodo	bigint
3	sexo	string

4	edad	string
5	capitulo	string
6	grupo	string
7	muerdes	string

Consulta y procesamiento de los datos

HIVE

A continuación, se usó *AWS EMR* para crear un clúster que permitiera enlazar los datos al ecosistema Apache Hadoop. Se utiliza HIVE con la finalidad de hacer consultas sobre las tablas crudas, previamente mencionadas.

- Análisis de riesgo

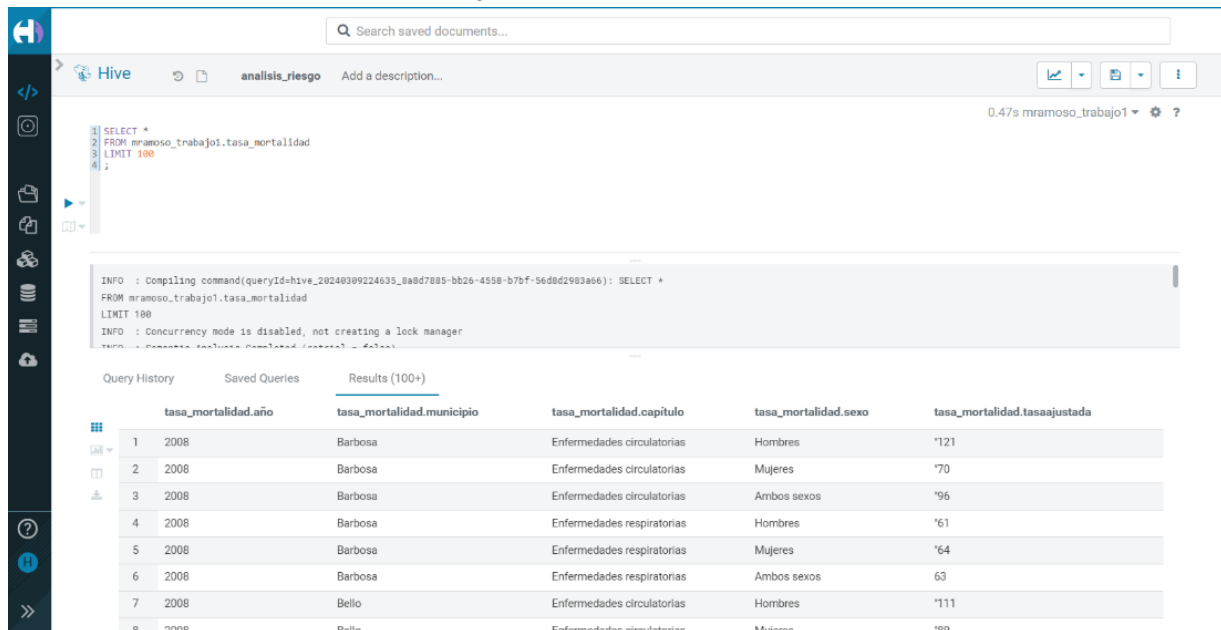
Asimismo, fue consultada mediante lenguaje SQL en HIVE la tabla Análisis de riesgos, la cual contiene 14 columnas (ver Figura 6)

	analisis_riesgo.tipo de rezago	analisis_riesgo.numero de modelo	analisis_riesgo.municipio	analisis_riesgo.grupo diagnóstico	analisis_riesgo.nombre contaminante	analisis_riesgo.sexo	analisis_riesgo.edad	analisis_riesgo.casos	analisis_riesgo.rezago	analisis_riesgo
1	Diarruido	1	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Mujer	Menores de 5 años	14482	0.03	10
2	Diarruido	2	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Mujer	Menores de 5 años	14482	0.07	10
3	Diarruido	3	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Mujer	Menores de 5 años	14482	0.15	10
4	Diarruido	4	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Hombre	Menores de 5 años	14435	0.03	10
5	Diarruido	5	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Hombre	Menores de 5 años	14435	0.07	10
6	Diarruido	6	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Hombre	Menores de 5 años	14435	0.15	10
7	Diarruido	7	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Amboos sexos	Menores de 5 años	30917	0.03	10
8	Diarruido	8	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Amboos sexos	Menores de 5 años	30917	0.07	10
9	Diarruido	9	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Amboos sexos	Menores de 5 años	30917	0.15	10
10	Diarruido	13	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Hombre	5 a 14 años	6476	0.03	10
11	Diarruido	14	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Hombre	5 a 14 años	6476	0.07	10
12	Diarruido	15	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Hombre	5 a 14 años	6476	0.15	10
13	Diarruido	16	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Amboos sexos	5 a 14 años	13011	0.03	10
14	Diarruido	17	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Amboos sexos	5 a 14 años	13011	0.07	10
15	Diarruido	18	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Amboos sexos	5 a 14 años	13011	0.15	10
16	Diarruido	19	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Mujer	Mayores de 65 años	1860	0.03	10
17	Diarruido	20	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Mujer	Mayores de 65 años	1860	0.07	10
18	Diarruido	21	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Mujer	Mayores de 65 años	1860	0.15	10
19	Diarruido	22	Madrid	Enfermedades del tubo medio y de la mastoides (H05.475)	pm10	Hombre	Mayores de 65 años	862	0.03	10

Figura 6. Tabla Análisis de riesgos en “raw”

- Tasa mortalidad

Se realizaron consultas mediante lenguaje SQL en HIVE de la tabla Tasa de mortalidad, la cual contiene cinco (5) columnas (ver Figura 7).



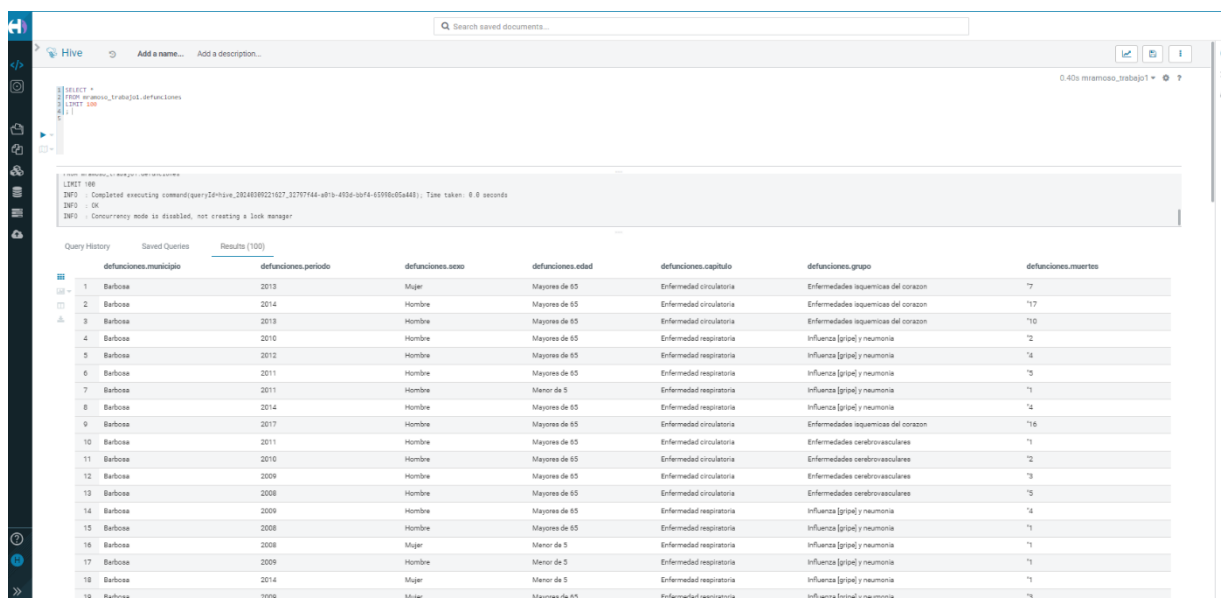
The screenshot shows the Hive web interface with a SQL query executed. The query is: `SELECT * FROM mramoso_trabajo1.tasa_mortalidad LIMIT 100`. The results are displayed in a table with 5 columns: `tasa_mortalidad.año`, `tasa_mortalidad.municipio`, `tasa_mortalidad.capitulo`, `tasa_mortalidad.sexo`, and `tasa_mortalidad.tasaajustada`.

	tasa_mortalidad.año	tasa_mortalidad.municipio	tasa_mortalidad.capitulo	tasa_mortalidad.sexo	tasa_mortalidad.tasaajustada
1	2008	Barbosa	Enfermedades circulatorias	Hombres	'121
2	2008	Barbosa	Enfermedades circulatorias	Mujeres	'70
3	2008	Barbosa	Enfermedades circulatorias	Ambos sexos	'96
4	2008	Barbosa	Enfermedades respiratorias	Hombres	'61
5	2008	Barbosa	Enfermedades respiratorias	Mujeres	'64
6	2008	Barbosa	Enfermedades respiratorias	Ambos sexos	63
7	2008	Bello	Enfermedades circulatorias	Hombres	'111
8	2008	Bello	Enfermedades circulatorias	Mujeres	'89

Figura 7. Tabla de tasa de mortalidad en “raw”

- Defunciones

Finalmente, mediante lenguaje SQL en HIVE con el fin de revisar la tabla defunciones, la cual contiene siete (7) columnas (ver Figura 8).



The screenshot shows the Hive web interface with a SQL query executed. The query is: `SELECT * FROM mramoso_trabajo1.defunciones LIMIT 100`. The results are displayed in a table with 7 columns: `defunciones.municipio`, `defunciones.periodo`, `defunciones.sexo`, `defunciones.edad`, `defunciones.capitulo`, `defunciones.grupo`, and `defunciones.muertes`.

	defunciones.municipio	defunciones.periodo	defunciones.sexo	defunciones.edad	defunciones.capitulo	defunciones.grupo	defunciones.muertes
1	Barbosa	2013	Mujer	Mayores de 65	Enfermedad circulatoria	Enfermedades isquémicas del corazón	'7
2	Barbosa	2014	Hombre	Mayores de 65	Enfermedad circulatoria	Enfermedades isquémicas del corazón	'17
3	Barbosa	2013	Hombre	Mayores de 65	Enfermedad circulatoria	Enfermedades isquémicas del corazón	'10
4	Barbosa	2010	Hombre	Mayores de 65	Enfermedad respiratoria	Influenza [gripal] y neumonía	'2
5	Barbosa	2012	Hombre	Mayores de 65	Enfermedad respiratoria	Influenza [gripal] y neumonía	'4
6	Barbosa	2011	Hombre	Mayores de 65	Enfermedad respiratoria	Influenza [gripal] y neumonía	'3
7	Barbosa	2011	Hombre	Menor de 5	Enfermedad respiratoria	Influenza [gripal] y neumonía	'1
8	Barbosa	2014	Hombre	Mayores de 65	Enfermedad respiratoria	Influenza [gripal] y neumonía	'4
9	Barbosa	2017	Hombre	Mayores de 65	Enfermedad circulatoria	Enfermedades isquémicas del corazón	'16
10	Barbosa	2011	Hombre	Mayores de 65	Enfermedad circulatoria	Enfermedades cerebrovasculares	'1
11	Barbosa	2010	Hombre	Mayores de 65	Enfermedad circulatoria	Enfermedades cerebrovasculares	'2
12	Barbosa	2009	Hombre	Mayores de 65	Enfermedad circulatoria	Enfermedades cerebrovasculares	'3
13	Barbosa	2008	Hombre	Mayores de 65	Enfermedad circulatoria	Enfermedades cerebrovasculares	'3
14	Barbosa	2009	Hombre	Mayores de 65	Enfermedad respiratoria	Influenza [gripal] y neumonía	'4
15	Barbosa	2008	Hombre	Mayores de 65	Enfermedad respiratoria	Influenza [gripal] y neumonía	'1
16	Barbosa	2008	Mujer	Menor de 5	Enfermedad respiratoria	Influenza [gripal] y neumonía	'1
17	Barbosa	2009	Hombre	Menor de 5	Enfermedad respiratoria	Influenza [gripal] y neumonía	'1
18	Barbosa	2014	Mujer	Menor de 5	Enfermedad respiratoria	Influenza [gripal] y neumonía	'1
19	Barbosa	2009	Mujer	Mayores de 65	Enfermedad respiratoria	Influenza [gripal] y neumonía	'3

Figura 8. Tabla defunciones en “raw”

Gracias a la consulta de la tabla “Defunciones” por medio de la Tabla 3 y la Figura 8, fue posible analizar que la columna “Muertes” que, es leída como un tipo de dato *String*, se entiende como una agregación del promedio de muertes por grupo, capítulo, edad, sexo, periodo y municipio. Es por ello que posteriormente en PySpark se realiza el proceso de limpieza y transformación de *String* a *Float* para pasar de la zona ‘raw’ a la zona ‘trusted’ de los datos.

SPARK

Para hacer esta transformación, se creó un cluster, desde el que se abrió un *notebook* en jupyterHub, allí se hizo la conversión del campo ‘muertes’ de *string* a *float*, adicionalmente, se eliminaron los registros en null, del mismo campo, y finalmente, se creó una tabla con la cantidad de muertes por municipio, que posteriormente se almacenó nuevamente en S3 en una nueva carpeta pero en el mismo bucket de los datos en crudo.

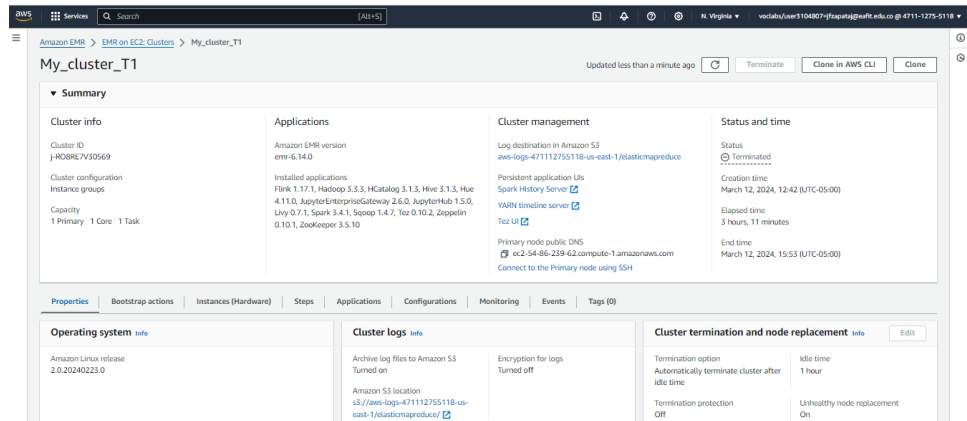


Figura 9. Creación del cluster.

Evidencia de ETL de datos de la columna muertes se pueden observar en la Figura 10.

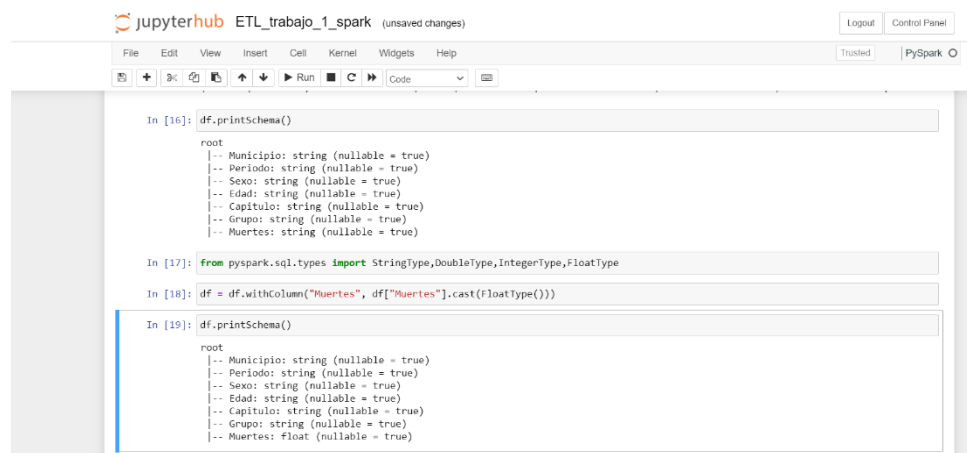


Figura 10. Uso de Spark para la ETL.

A continuación, en la Figura 11 se muestra el almacenamiento del notebook en el que se hizo la transformación de los datos desde Jupyterhub usando el archivo ETL trabajo 1 spark mramoso.ipynb.

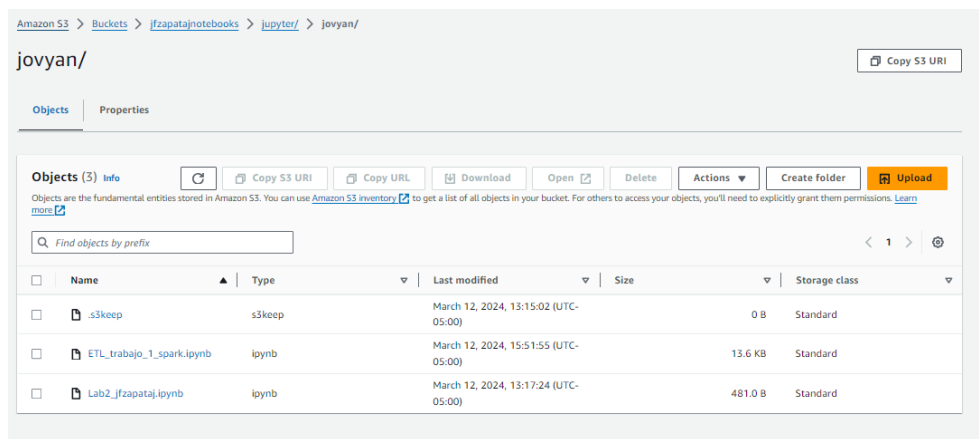


Figura 11. Almacenamiento del notebook

Finalmente, se muestra a continuación, en la Figura 12 la nueva tabla con los datos transformados y resumidos, almacenados en un nuevo folder en S3.

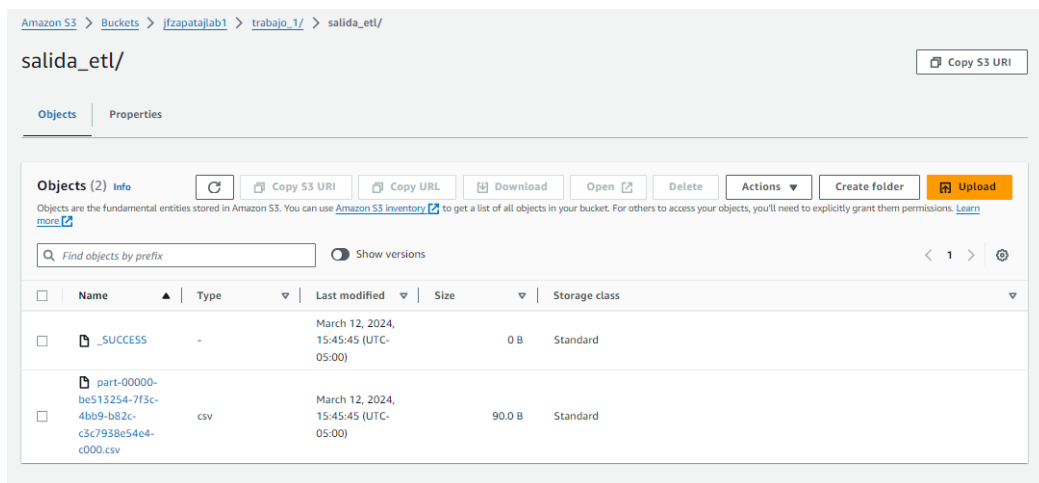


Figura 12. Almacenamiento de los datos trusted en S3.

En la siguiente ruta de Github se podrán ver los archivos mencionados en el documento <https://github.com/jofzapat/Data Science EAFIT.git>

REDSHIFT

De otro lado, también se creó un cluster en AWS Redshift con la finalidad de cargar datos desde S3 a Redshift, tal como se muestra en la Figura 13.

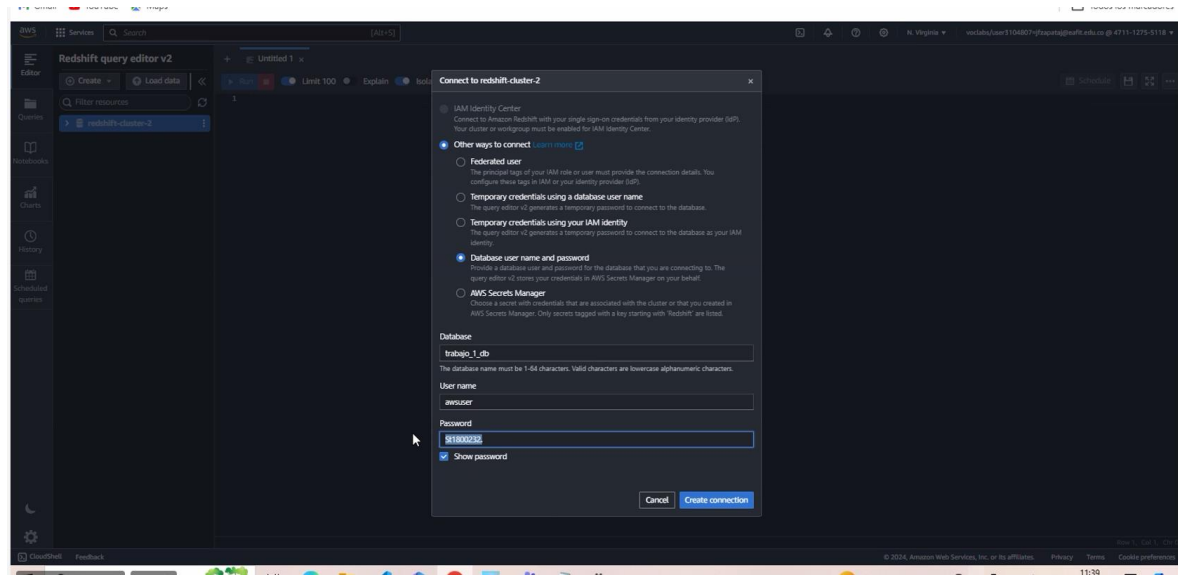


Figura 13. Creación del cluster en Redshift Spectrum.

Inicialmente se crea una tabla externa con la misma estructura de las Tabla 1, Tabla 2 y Tabla 3. Los registros se extraen del *bucket* en S3. En este caso se analiza la tabla de análisis de riesgo (ver Figura 14).

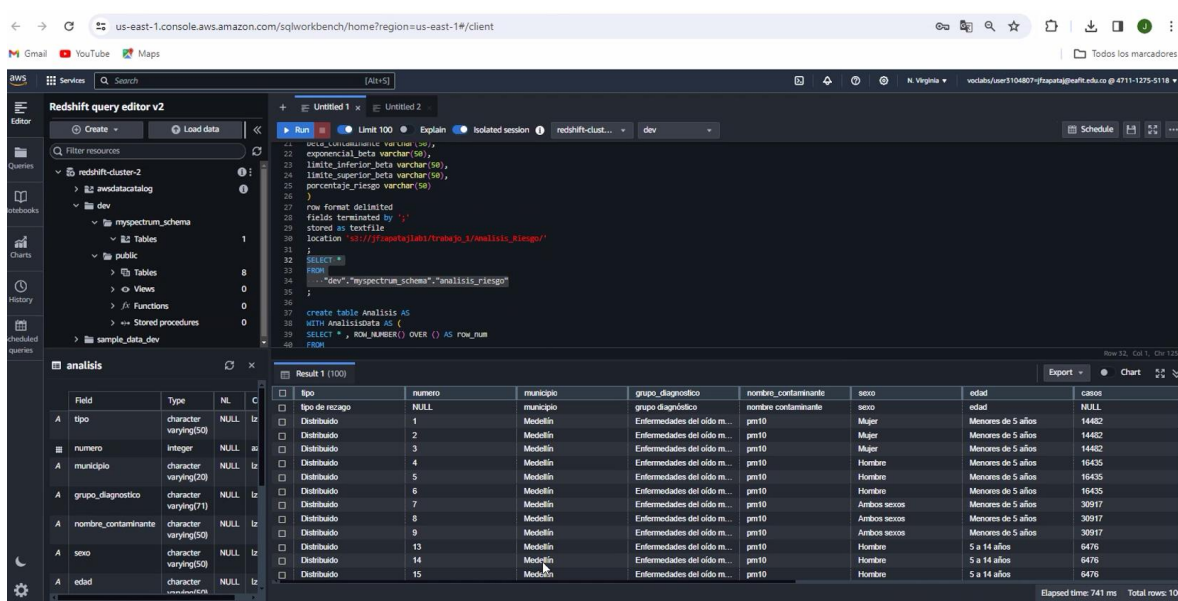


Figura 14. Creación de la tabla externa a Redshift Spectrum.

Para realizar un proceso ETL es necesario crear una tabla nativa en *Redshift Spectrum* con la misma información de la tabla externa, en particular se realizó una copia de la tabla previamente creada, tal como se muestra en la Figura 15.

The screenshot shows the Redshift query editor v2 interface. The SQL query in the editor is as follows:

```

37 create table Analisis AS
38 WITH AnalisisData AS (
39 SELECT *, ROW_NUMBER() OVER () AS row_num
40 FROM
41 "dev"."myspectrum_schema"."analisi_riesgo"
42 )
43 SELECT *
44 FROM AnalisisData
45 WHERE row_num > 1
46 ;
47
48 create table Analisis AS
49 WITH etl AS (
50 SELECT *
51 FROM Analisis
52 )

```

The result set shows the following data:

Field	Type	NL	C
A tipo	character varying(50)	NULL	lz
A numero	integer	NULL	ai
A municipio	character varying(20)	NULL	lz
A grupo_diagnostico	character varying(7)	NULL	lz
A nombre_contaminante	character varying(50)	NULL	lz
A sexo	character varying(50)	NULL	lz
A edad	character varying(50)	NULL	lz

The data table below shows the results of the query:

edad	casos	rezago	beta_contaminante	exponencial_beta	limite_inferior_beta	limite_superior_beta	porcentaje_riesgo
Menores de 5 años	16435	0.07	0.0919421838	1.055736253	1.06373917	1.107054673	9.58
Menores de 5 años	16435	0.15	0.126650884	1.135028867	1.117676595	1.152656818	13.51
Menores de 5 años	30917	0.03	0.046325275	1.047415953	1.038556245	1.056349427	4.75
Menores de 5 años	30917	0.07	0.061466682	1.067810206	1.075452985	1.100270747	8.79
Menores de 5 años	30917	0.15	0.122333265	1.139153363	1.117864154	1.148623544	13.02
5 a 14 años	6476	0.03	0.029337863	1.029772436	1.022529331	1.037667885	2.98
5 a 14 años	6476	0.07	0.036894451	1.037376894	1.029681554	1.045229462	3.74
5 a 14 años	6476	0.15	0.054750385	1.056276899	1.042577783	1.063737689	5.63
5 a 14 años	13011	0.03	0.011546434	1.011613352	1.004065333	1.019279822	1.17
5 a 14 años	13011	0.07	0.016108941	1.016326892	1.00883253	1.02381296	1.64
5 a 14 años	13011	0.15	0.018201473	1.018207887	1.007478184	1.031436598	1.94
Mayores de 65 años	1860	0.03	0.023163386	1.023387687	1.017273981	1.034163355	2.34
Mayores de 65 años	1860	0.07	0.024374884	1.024627483	1.014486425	1.034964854	2.47

Figura 15. Creación de la tabla nativa en Redshift Spectrum.

A esta última tabla se le realizó un proceso ETL, el cual consistió básicamente en reducir el número de decimales a dos (2) cifras, de las columnas tipo *float*, así como reducir la longitud de las columnas “sexo” y “edad”, en este último caso renombrarla a “edad_años”, para que la unidad de medida quedara en el *header* y no en los registros. Finalmente, se eliminaron columnas cuya información no era relevante, como la columna “rezago”, la cual está asociada con la hora en formato a.m. y p.m. Los resultados del proceso ETL se muestran en términos generales en la Figura 16 y el proceso completo se describe en el anexo “ETL_Redshift.sql”

The screenshot shows the Redshift query editor v2 interface. The SQL query in the editor is as follows:

```

49 FROM "dev"."public"."analisi"
50 ;
51
52 create table Analisis_etl AS
53 select municipio
54 , grupo_diagnostico
55 , CASE WHEN sexo = 'Mujer' THEN 'F' WHEN sexo = 'Hombre' THEN 'M' ELSE 'FPM' END AS sexo
56 , CASE WHEN SUBSTR(edad, 1, 7) = 'Menores' THEN 'M' WHEN SUBSTR(edad, 1, 7) = 'Menores' THEN '5 - 14' END AS edad_años
57 , CAST(casos AS integer) AS casos
58 , CAST(beta_contaminante AS decimal(5,2)) AS beta_contaminante
59 , CAST(exponencial_beta AS decimal(5,2)) AS exponencial_beta
60 , CAST(limite_inferior_beta AS decimal(5,2)) AS limite_inferior_beta
61 , CAST(limite_superior_beta AS decimal(5,2)) AS limite_superior_beta
62 , CAST(percentage_risk AS decimal(5,2)) AS percentage_risk
63 FROM
64 "dev"."public"."analisi";
65
66
67

```

The result set shows the following data:

Field	Type	NL	C
A tipo	character varying(50)	NULL	lz
A numero	integer	NULL	ai
A municipio	character varying(20)	NULL	lz
A grupo_diagnostico	character varying(7)	NULL	lz
A nombre_contaminante	character varying(50)	NULL	lz
A sexo	character varying(50)	NULL	lz
A edad	character varying(50)	NULL	lz

Figura 16. Proceso ETL en Redshift Spectrum

En síntesis, en la Figura 17 se muestra a la izquierda una tabla con los datos en la zona *raw*, y en la derecha una tabla en la zona *trusted*, ya lista para usar sus datos en un modelo o como parte de una app.

Tabla en zona *raw*

	grupo_diagnostico	nombre_contaminante	sexo	edad	casos	rezago	beta_contaminante	exponencial_beta	limite_inferior_beta	limite_superior_beta	porcentaje_riesgo	row_num
A	grupo_diagnostico	character varying(71)	NULL	lzo								
A	nombre_contaminante	character varying(50)	NULL	lzo								
A	sexo	character varying(50)	NULL	lzo								
A	edad	character varying(50)	NULL	lzo								
A	casos	integer	NULL	az64								
A	rezago	character varying(50)	NULL	lzo								
A	beta_contaminante	character varying(50)	NULL	lzo								
A	exponencial_beta	character varying(50)	NULL	lzo								
A	limite_inferior_beta	character varying(50)	NULL	lzo								
A	limite_superior_beta	character varying(50)	NULL	lzo								
A	porcentaje_riesgo	character varying(50)	NULL	lzo								
A	row_num	bigint	NULL	az64								

Tabla en zona *trusted*

	Field	Type	NL	OMP
A	municipio	character varying(20)	NULL	lzo
A	grupo_diagnostico	character varying(71)	NULL	lzo
A	sexo	character varying(3)	NULL	lzo
A	edad_años	character varying(6)	NULL	lzo
A	casos	integer	NULL	az64
A	beta_contaminante	numeric(5,2)	NULL	az64
A	exponencial_beta	numeric(5,2)	NULL	az64
A	limite_inferior_beta	numeric(5,2)	NULL	az64
A	limite_superior_beta	numeric(5,2)	NULL	az64
A	porcentaje_riesgo	numeric(5,2)	NULL	az64

Figura 17. Comparación de datos en *raw* vs en *trusted* haciendo ETL en Redshift Spectrum

La tabla final, resultante del proceso ETL y los registros asociados se muestran parcialmente luego de un *query* en la Figura 18.

Result 1 (100)

municipio	grupo_diagnostico	sexo	edad_años	casos	beta_contaminante	exponencial_beta	limite_inferior_beta	limite_superior_beta	porcentaje_riesgo
Modelín	Enfermedades del oído m...	M	< 5	16435	0.05	1.05	1.04	1.06	5.78
Modelín	Enfermedades del oído m...	F/M	< 5	30917	0.08	1.08	1.07	1.1	8.79
Modelín	Enfermedades del oído m...	M	5 - 14	6476	0.05	1.05	1.04	1.06	5.63
Modelín	Enfermedades del oído m...	F	> 65	1860	0.02	1.02	1.01	1.03	2.34
Modelín	Enfermedades del oído m...	M	> 65	802	0.03	1.03	1.02	1.03	3.22
Modelín	Enfermedades del oído m...	F/M	> 65	2722	0.03	1.03	1.02	1.05	3.9
Modelín	Enfermedades del oído m...	F/M	< 5	30894	0.1	1.11	1.09	1.13	11.55
Modelín	Enfermedades del oído m...	F/M	5 - 14	13003	0.02	1.03	1.02	1.03	3.01
Modelín	Enfermedades del oído m...	F	> 65	1858	0.04	1.05	1.03	1.06	5.12
Modelín	Enfermedades del oído m...	F/M	> 65	2717	0.08	1.08	1.07	1.09	8.58
Modelín	Enfermedades del oído m...	F	< 5	13979	0.09	1.09	1.08	1.1	9.48
Modelín	Enfermedades del oído m...	F/M	< 5	29631	0.06	1.07	1.06	1.07	7.11
Modelín	Enfermedades del oído m...	F	5 - 14	6342	0.02	1.02	1.01	1.03	2.35
Modelín	Enfermedades del oído m...	M	5 - 14	6261	0.02	1.02	1.01	1.02	2.11
Modelín	Enfermedades del oído m...	F	> 65	1824	0	1	0.99	1	0.44
Modelín	Infecciones agudas de la...	F	< 5	154256	0.06	1.07	1.06	1.07	7.12
Modelín	Infecciones agudas de la...	M	< 5	166267	0.09	1.1	1.09	1.1	10.25
Modelín	Infecciones agudas de la...	F	5 - 14	88842	0.03	1.03	1.03	1.04	3.88
Modelín	Infecciones agudas de la...	M	5 - 14	78231	0.05	1.05	1.05	1.06	5.78
Modelín	Infecciones agudas de la...	F/M	5 - 14	159123	0.07	1.08	1.07	1.08	8.23
Modelín	Infecciones agudas de la...	F	< 5	154155	0.13	1.14	1.13	1.16	14.53
Modelín	Infecciones agudas de la...	M	< 5	166176	0.16	1.17	1.16	1.18	17.58
Modelín	Infecciones agudas de la...	F	5 - 14	88797	0.07	1.07	1.06	1.08	7.57
Modelín	Infecciones agudas de la...	M	5 - 14	78244	0.09	1.09	1.09	1.1	9.83

Figura 18. Resultados finales del proceso ETL en Redshift Spectrum

Finalmente, se muestra en la Figura 19, un resumen de los pasos y herramientas implementadas en cada una de las etapas del ciclo de vida de los datos.

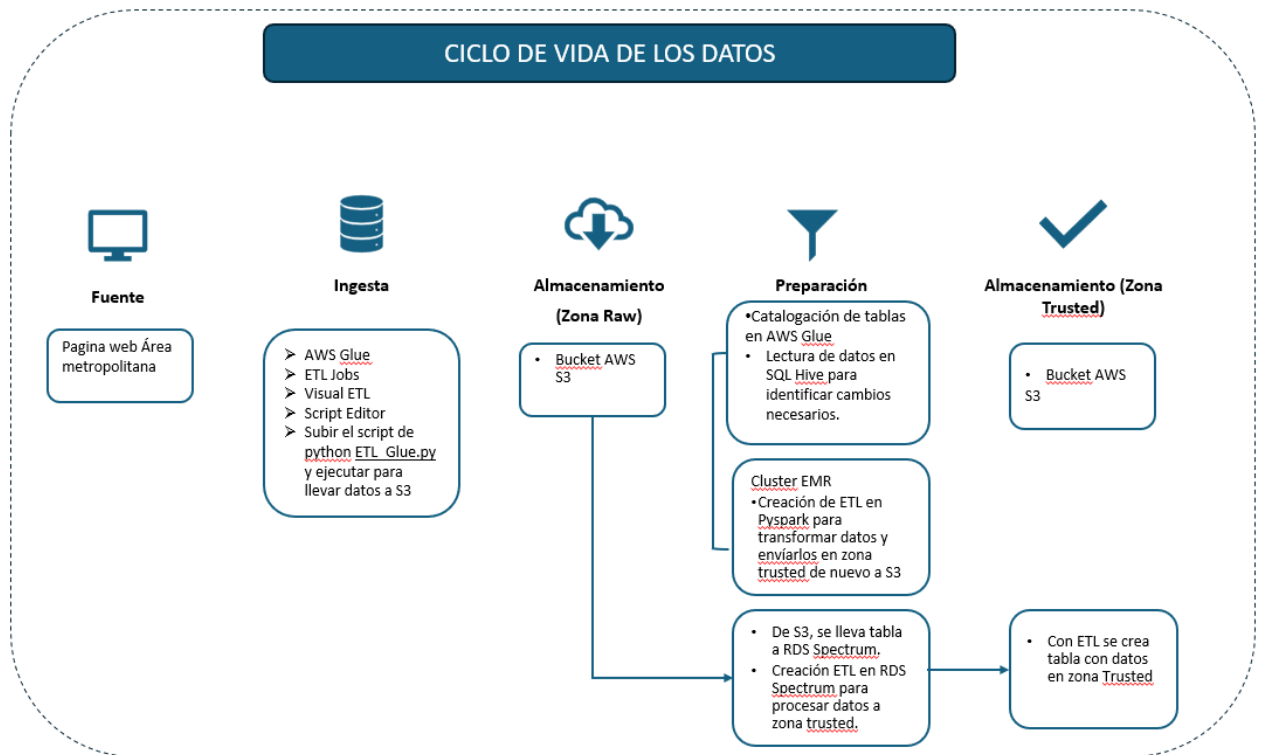


Figura 19. Esquema ciclo de vida de los datos