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](#_Toc81206230)

[EC2 Key Pair 3](#_Toc81206231)

[Creación del bucket S3 3](#_Toc81206232)

[Particionamiento y zonas del datalake 5](#_Toc81206233)

[Subida de archivos al datalake 7](#_Toc81206234)

[Cambio de políticas del bucket 11](#_Toc81206235)

[Creación de primer Crawler 11](#_Toc81206236)

[Query de prueba en Athena 13](#_Toc81206237)

[Procesos ETL y Crawlers con Glue 14](#_Toc81206238)

[Consultas en Athena 22](#_Toc81206239)

[Creación del EMR 24](#_Toc81206240)

[Consultas con Hive 26](#_Toc81206241)

Implementación del Datalake

EC2 Key Pair

Interfaz de usuario gráfica, Texto, Aplicación, Sitio web

Descripción generada automáticamente

Primeros 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

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

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

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

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

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

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

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente

Interfaz de usuario gráfica, Sitio web

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

En las anteriores imágenes pudimos observer 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:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Creación de primer Crawler

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Query de prueba en Athena

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* 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.

Interfaz de usuario gráfica

Descripción generada automáticamente

Procesos ETL y Crawlers con Glue

* Glue ETL

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

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

Diagrama

Descripción generada automáticamente

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

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Propiedades del Job.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Se agregó AmazonS3FullAccess al rol IAM de Glue.

Interfaz de usuario gráfica, Texto, Aplicación

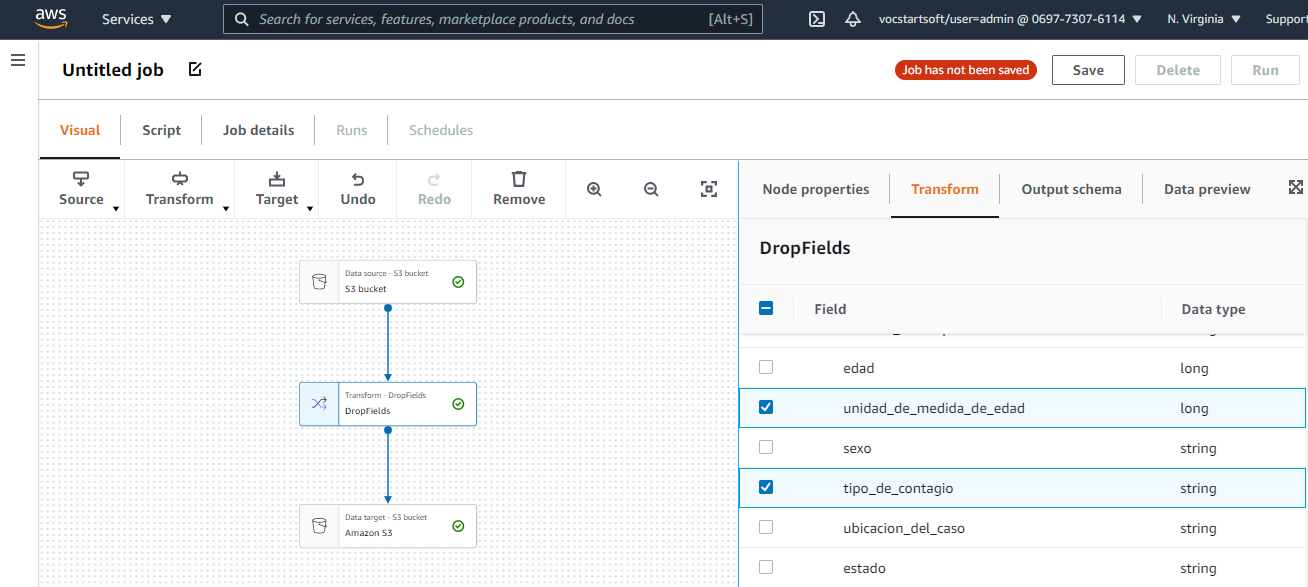
Descripción generada automáticamente

Corrida del Job exitosa.

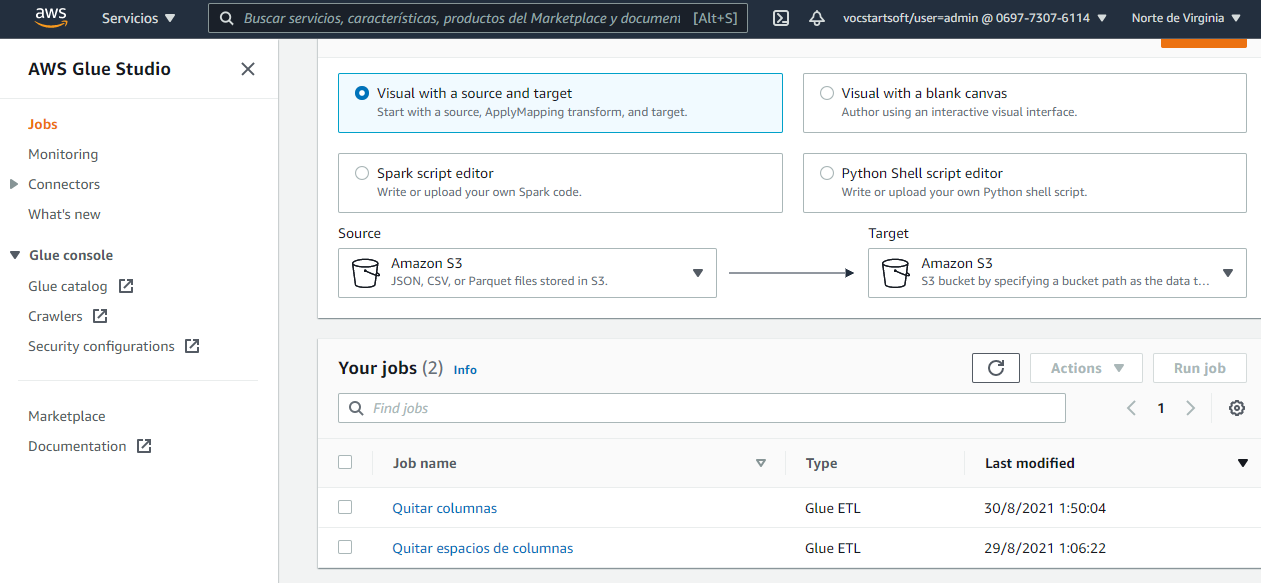
Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

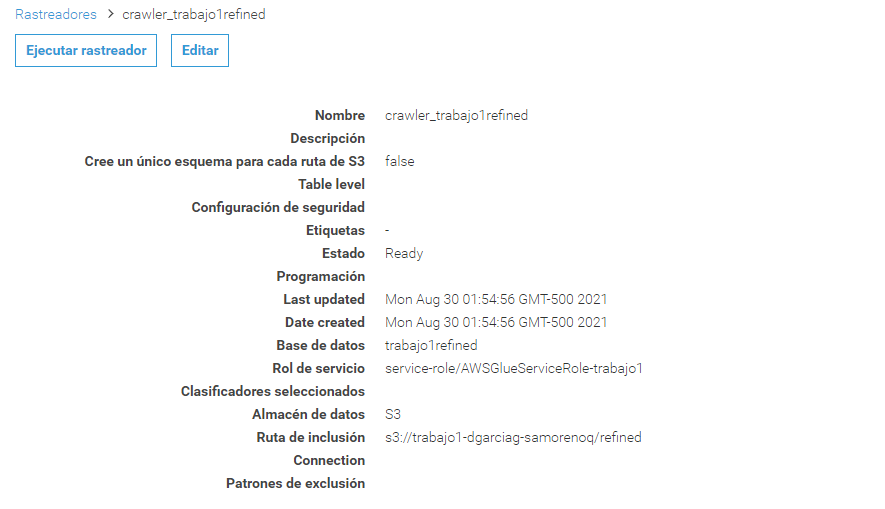
Nuevo crawler para la zona Trusted.



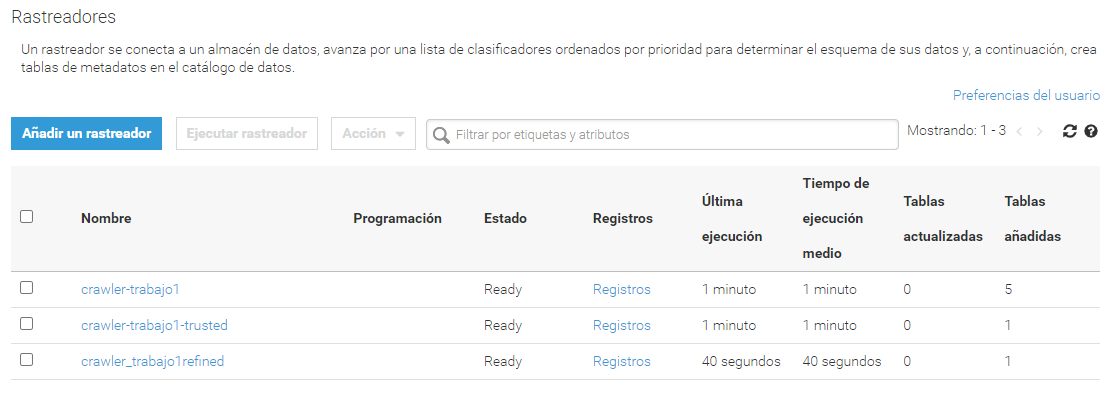
En un nuevo job eliminamos algunas columnas que para efectos de este trabajo son innecesarias y no nos brindan mucha información. Vamos a recibir la nueva tabla que tenemos en la zona trusted, hacemos la transformación y almacenamos en la zona refined.



Ahora tenemos 2 procesos ETL.



Crawler para agregar la nueva tabla a la zona refined.



Tenemos 3 crawlers en total.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

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



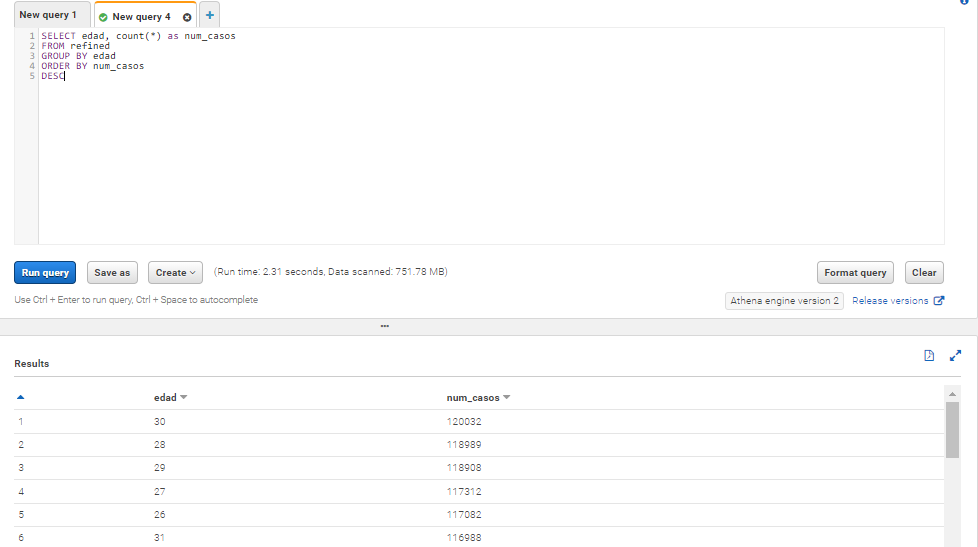
Resultado después del proceso de ETL con Glue y el crawler en refined: ya las columnas están bien nombradas y no tenemos columnas innecesarias como unidad de medida de edad, código iso del país, nombre del país (En cualquiera de los casos estamos hablando de Colombia), tipo de recuperación, etc.

Consultas en Athena

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Query en Athena usando los nuevos nombres de columnas.



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.)Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

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

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

* Nombre del cluster

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

* Key-pair del cluster

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

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

Tabla

Descripción generada automáticamente

* 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.

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente

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

Consultas con Hive

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

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

Interfaz de usuario gráfica

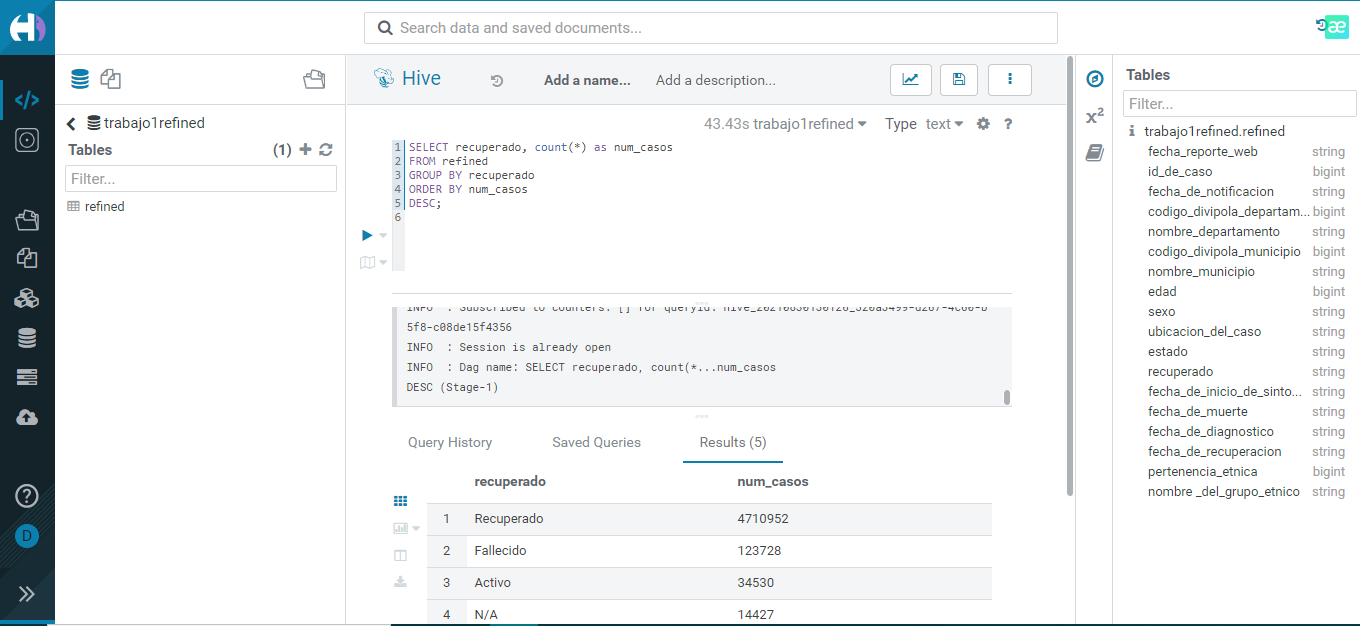
Descripción generada automáticamente con confianza media

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

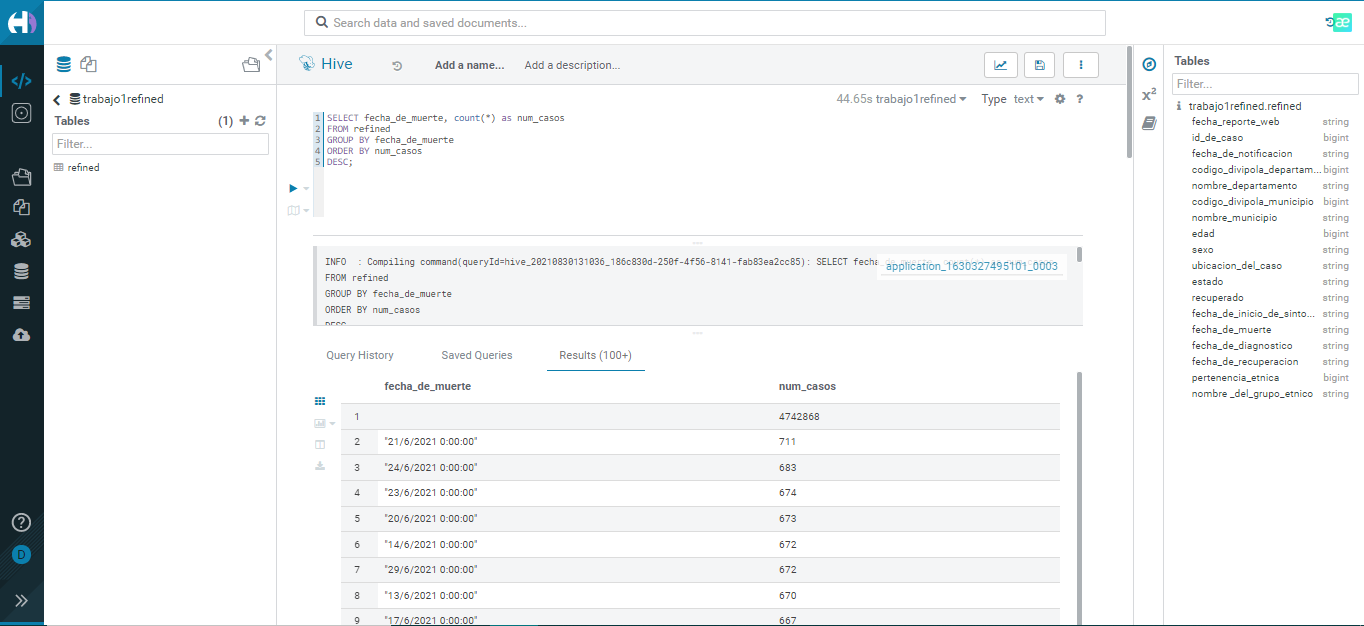
Una captura de pantalla de una computadora

Descripción generada automáticamente

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.



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.



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.