Soporte punto 2

Rol Ingeniería de Datos

Proyecto 3 Analítica Computacional

Jorge Russi Russi

1. Buquets:

Primero se crea un buquet de entrada llamado proyecto3-jorge-entrada, donde almacenaremos nuestra data inicial y en este buquet se crea una carpeta llamada datos\_base:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

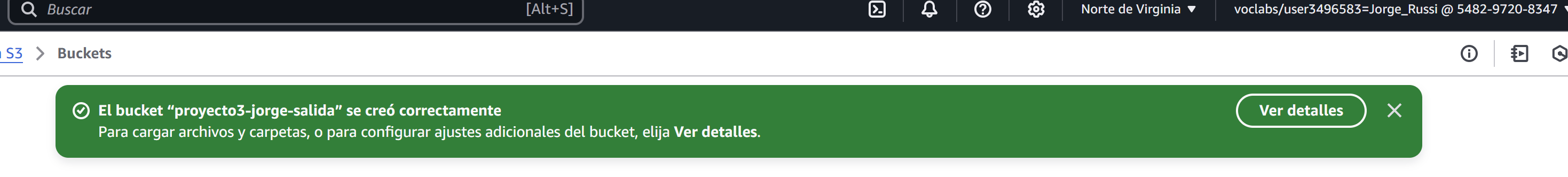
Descripción generada automáticamente

Y luego se sube el archivo “data\_proyecto3.csv” descargado con 7.1 millones de filas a la carpeta datos\_base:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ahora creamos un buquet de salida, donde almacenaremos nuestros resultados ya filtrados, este buquet se llamara proyecto3-jorge-salida el cual de momento se queda vacio



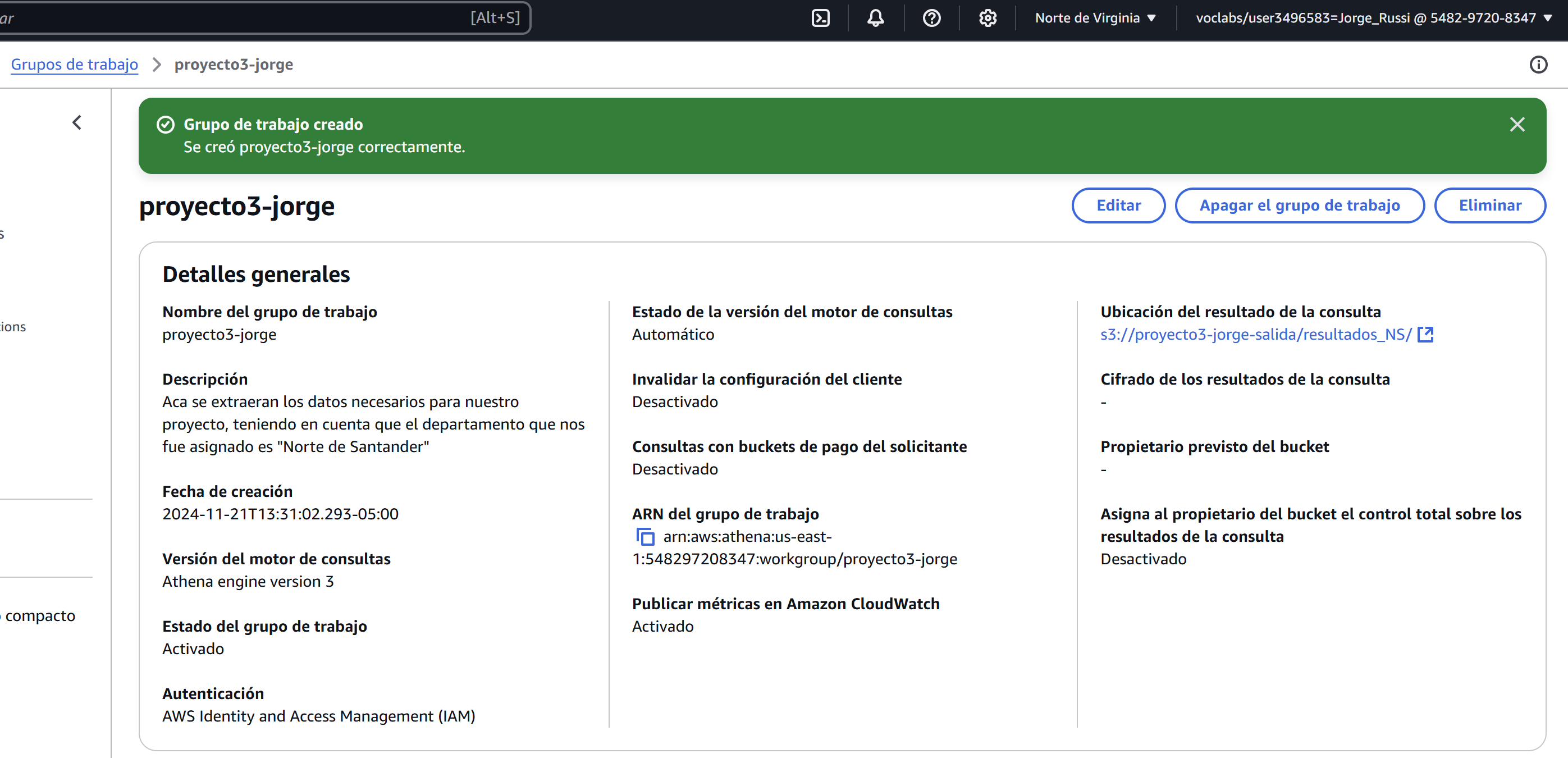
En este buquet creamos una carpeta llamada resultados\_NS, donde almacenaremos los resultados de nuestra extracción y transformación.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

1. Athena:

Primero creamos un grupo de trabajo en el cual extraeremos y transformaremos nuestra data y especificamos que los resultados de este grupo de trabajo estarán en el buquet de salida en la carpeta resultados\_NS:



1. AWS Clue:

Ahora creamos un crawler, el cual nos ayudara a leer la data de nuestro buquet de entrada, especificamos que de ese crawler debe salir la data. Y generamos una base de datos, donde queda la data de salida.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

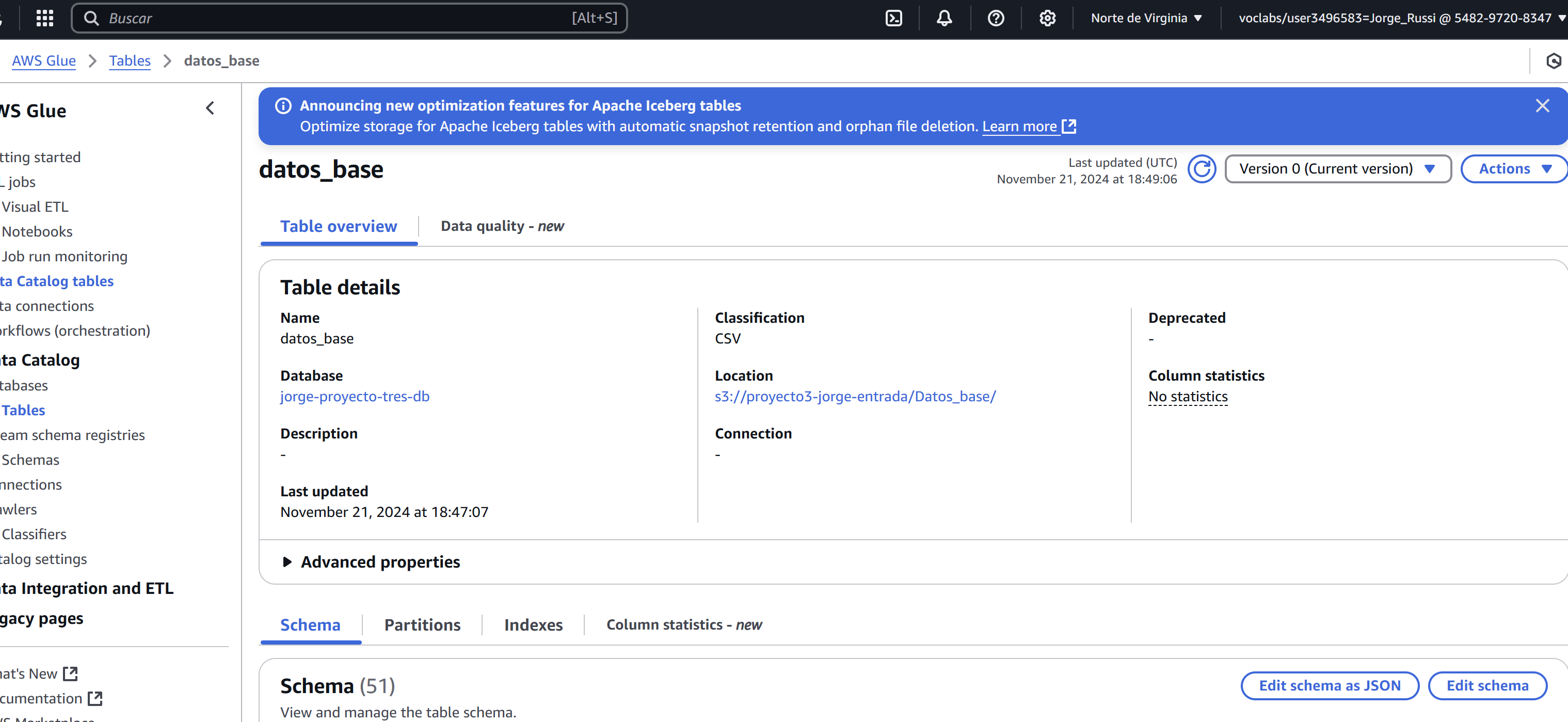
Descripción generada automáticamente

Ahora se corren el crawler proyecto3-crawler, para inicializarlo:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

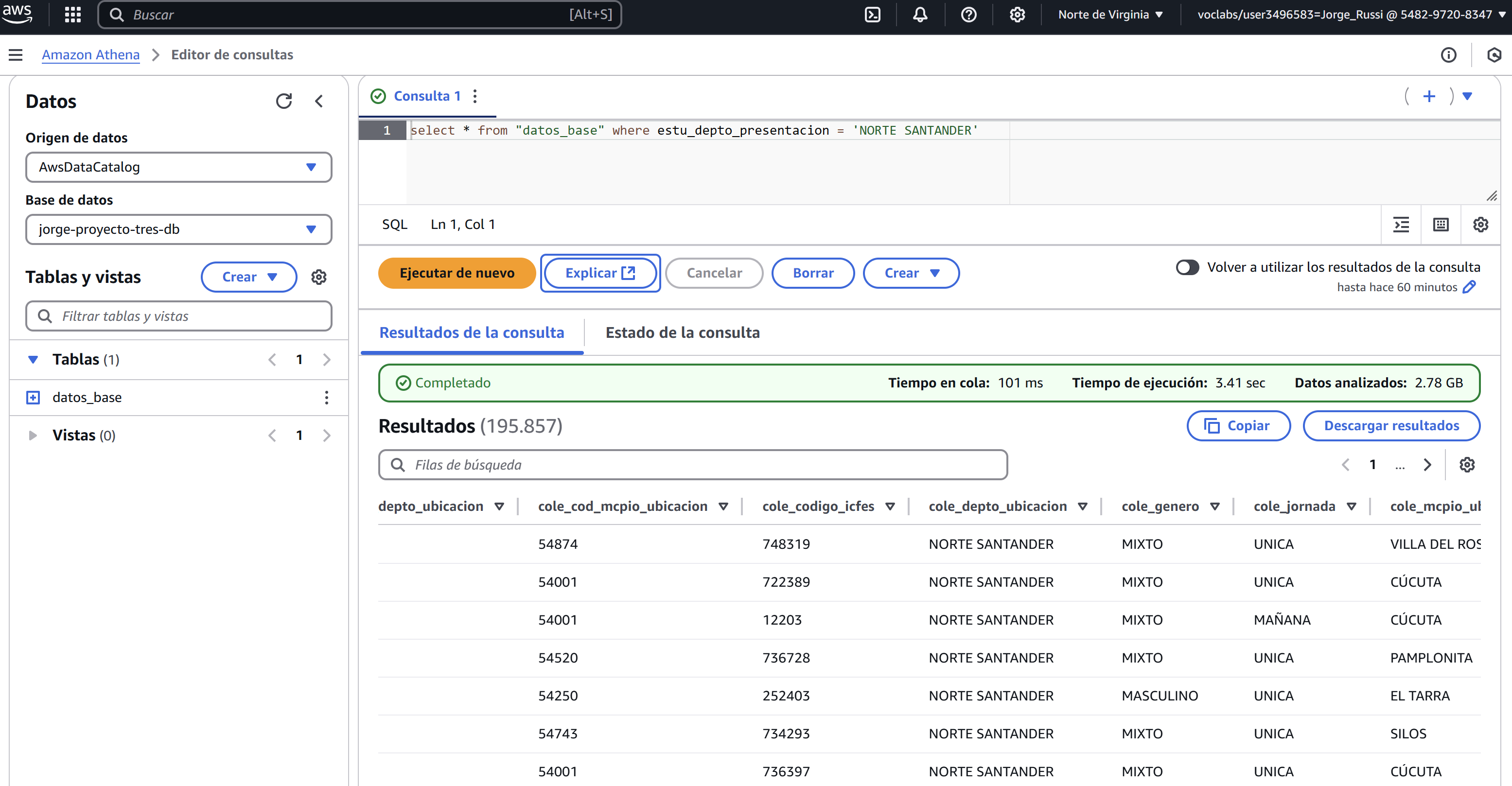
Ya quedo creado la tabla y la base de datos luego de que se termina de correr el crawler. Tiene sus 51 columnas y ya estamos listos para empezar a transformar la data:



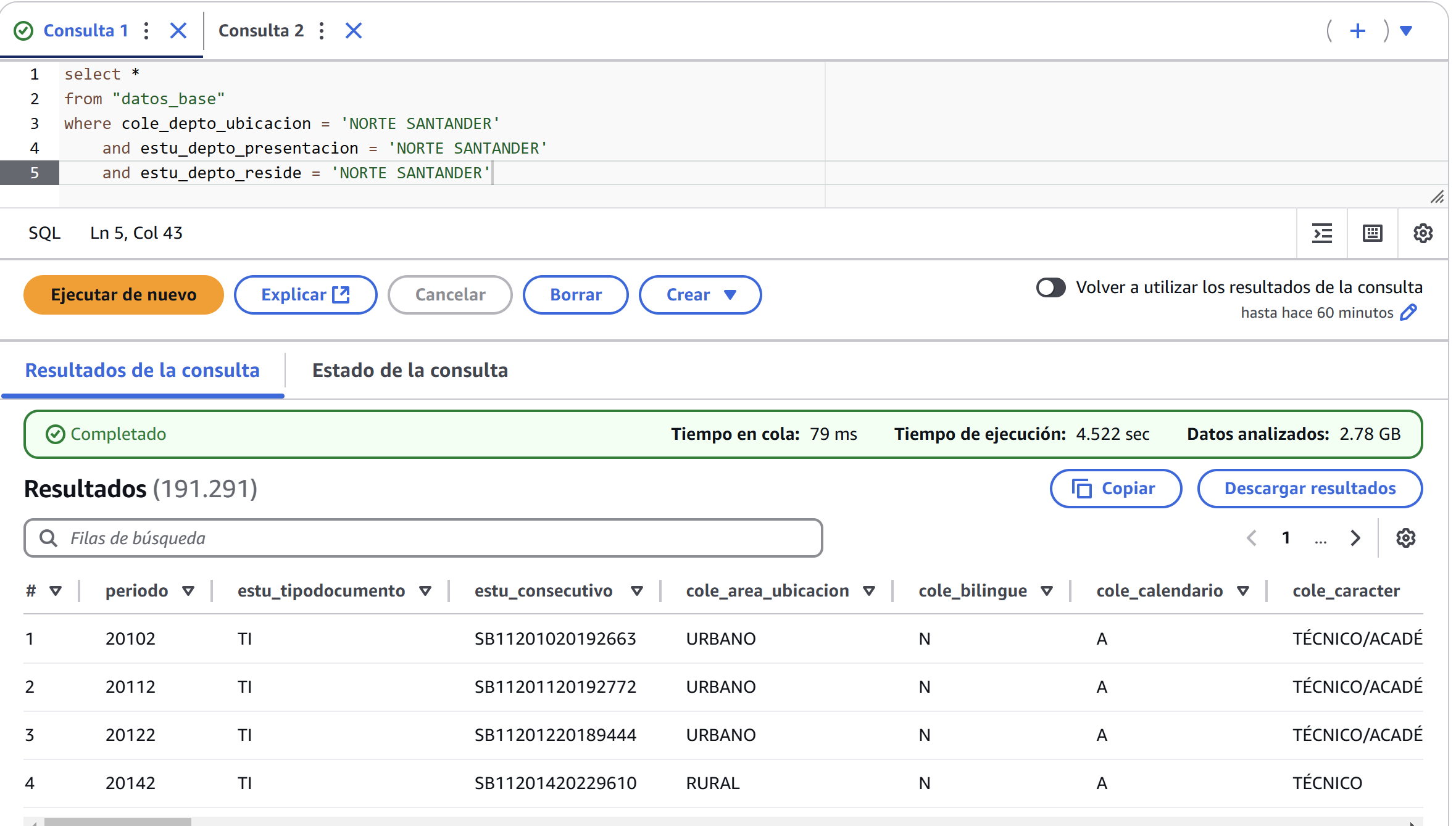
1. Volvemos a Athenas, al editor de consultas:

Acá generamos una consulta con la cual filtramos la columna estu\_depto\_presentacion para que se solo en departamentos de “Norte de Santander”, con el siguiente query:

select \* from "datos\_base" where estu\_depto\_presentacion = 'NORTE SANTANDER'



Nos damos cuenta que algunas ciudades de otros departamentos entraron a la data, entonces hacemos unos filtros extra para intentar que esto no pase:



Ya estos datos los podemos descargar resultados y tendremos un csv con los datos de solo Norte de Santander, con esto ya terminamos la primer parte del ETL, donde ya esta la Extracción y Transformación y podemos empezar a hacer la Limpieza en Python: