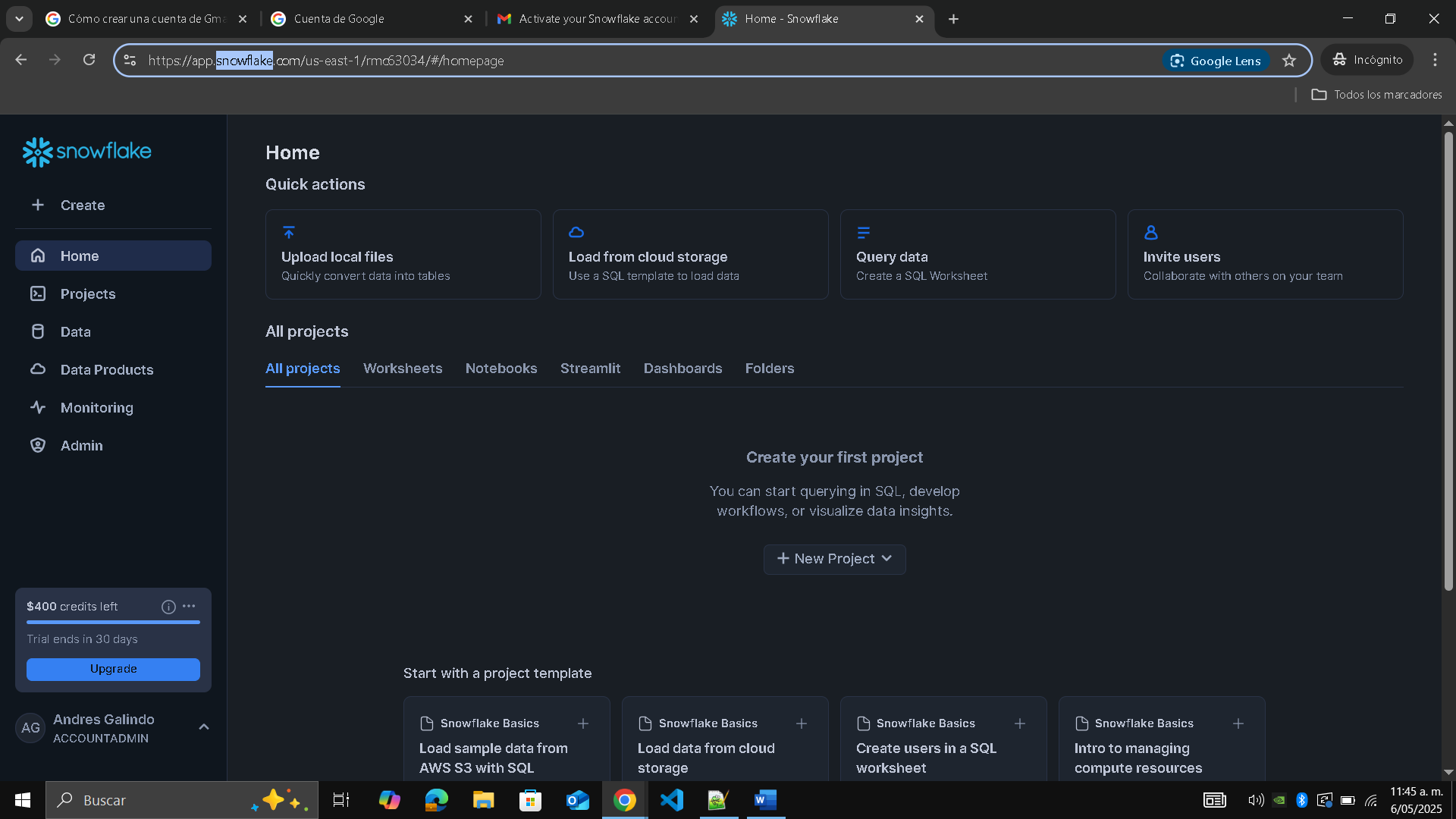
Andres Javier Galindo Vargas

Prueba Ingeniero de datos

1. Cuenta de snowflake



1. Descargo los datos

|  |  |
| --- | --- |
| Interfaz de usuario gráfica  El contenido generado por IA puede ser incorrecto. |  |
|  |  |

1. Carga de dataset
2. Para realizar la carga de los datos, lo primero que voy a realizar:

Creación de la base de datos y después importar los dataset para la creación de las tablas:

Se dejará sobre un esquema raw con el fin de informar que es la data en crudo

|  |  |
| --- | --- |
|  |  |

1. Cargar los dataset

Al cargar las tablas salen errores, esto sucede por columnas que tienen puntos en los nombres o string que Snowflake no los acepta como válidos, por lo que se les da manejo con "Autofix column names", se cargan la tablas a excepción de la tabla members, por tamaño no permite se toma un parcial de los datos

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

1. Creación de tablas auxiliares

Se realiza la revisión de las muestras de los archivos aprovisionados para el caso puntual

Análisis de los datos:

1. Analisis general de las tablas teniendo en cuenta que se cargaron en el esquema ROW

| **Nombre de la tabla** | **Descripción** |
| --- | --- |
| RAW.CATEGORIES | Contiene la lista de categorías generales disponibles en Meetup. Incluye un category\_id y su nombre. Se puede usar como tabla de referencia. |
| RAW.CITIES | Tabla con información básica de ciudades, incluyendo nombre, estado y país. Útil para enriquecer datos de ubicación. |
| RAW.EVENTS | Detalla eventos realizados, incluyendo duración, fecha de creación, descripción, URL y relación con un grupo (group\_id). Es una tabla central para análisis temporal o de participación. |
| RAW.GROUPS | Información sobre los grupos que organizan los eventos. Contiene ciudad, país, nombre, link y categoría. Se conecta con events y groups\_topics. |
| RAW.GROUPS\_TOPICS | Relación muchos-a-muchos entre grupos y los temas que representan. Funciona como tabla de unión entre groups y topics. |
| RAW.MEMBERS | Datos de los miembros registrados en la plataforma. Contiene nombre, género, edad, país, y fecha de inscripción. Es clave para análisis de usuarios. |
| RAW.MEMBERS\_TOPICS | Tabla que muestra qué temas interesan a cada miembro (member\_id, topic\_id). Permite análisis de afinidad temática o segmentación. |
| RAW.TOPICS | Lista de temas disponibles en la plataforma. Se usa como dimensión o para enriquecer relaciones con grupos y miembros. |
| RAW.VENUES | Contiene los lugares físicos donde se han realizado eventos. Incluye dirección, ciudad, país, y coordenadas. Se puede relacionar con events vía venue\_id. |

1. Ahora bien, para la creación de las auxiliares tomamos de base la anteriores y crearemos unas nuevas refinando la información de algunas de ellas como base de caso propuesto:

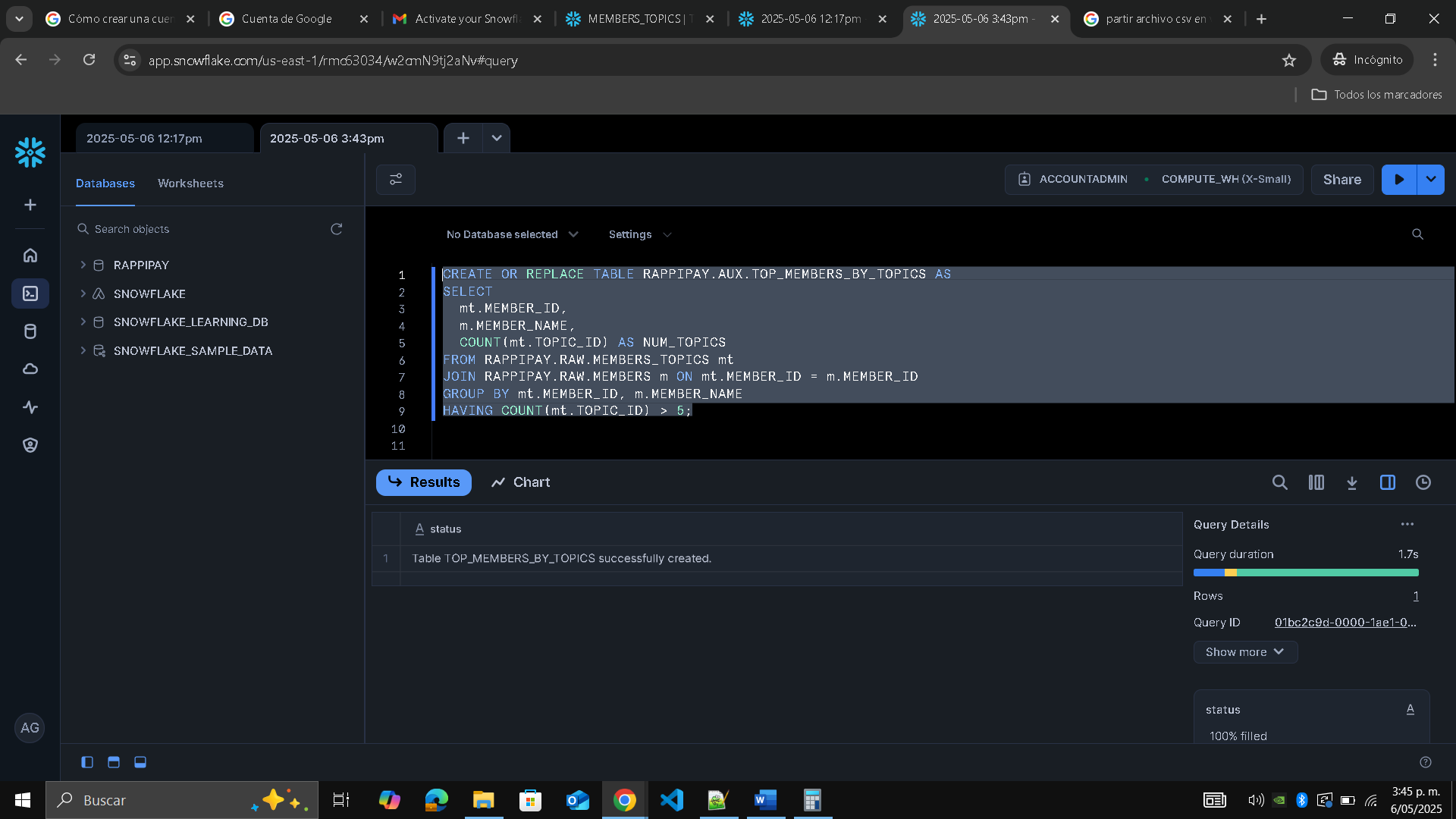
En este caso separaremos las capas se va crear una llamada AUX como parte de la solución

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Creación de tablas auxiliares

* La primera que podemos hacer es identificar miembros con mas de 5 temas de interés, en ese caso nos toca cruzar las tablas de miembros con tópicos



Nombre: AUX.TOP\_MEMBERS\_BY\_TOPICS , manteniendo el nombre de las tablas

CREATE OR REPLACE TABLE RAPPIPAY.AUX.TOP\_MEMBERS\_BY\_TOPICS AS

SELECT

mt.MEMBER\_ID,

m.MEMBER\_NAME,

COUNT(mt.TOPIC\_ID) AS NUM\_TOPICS

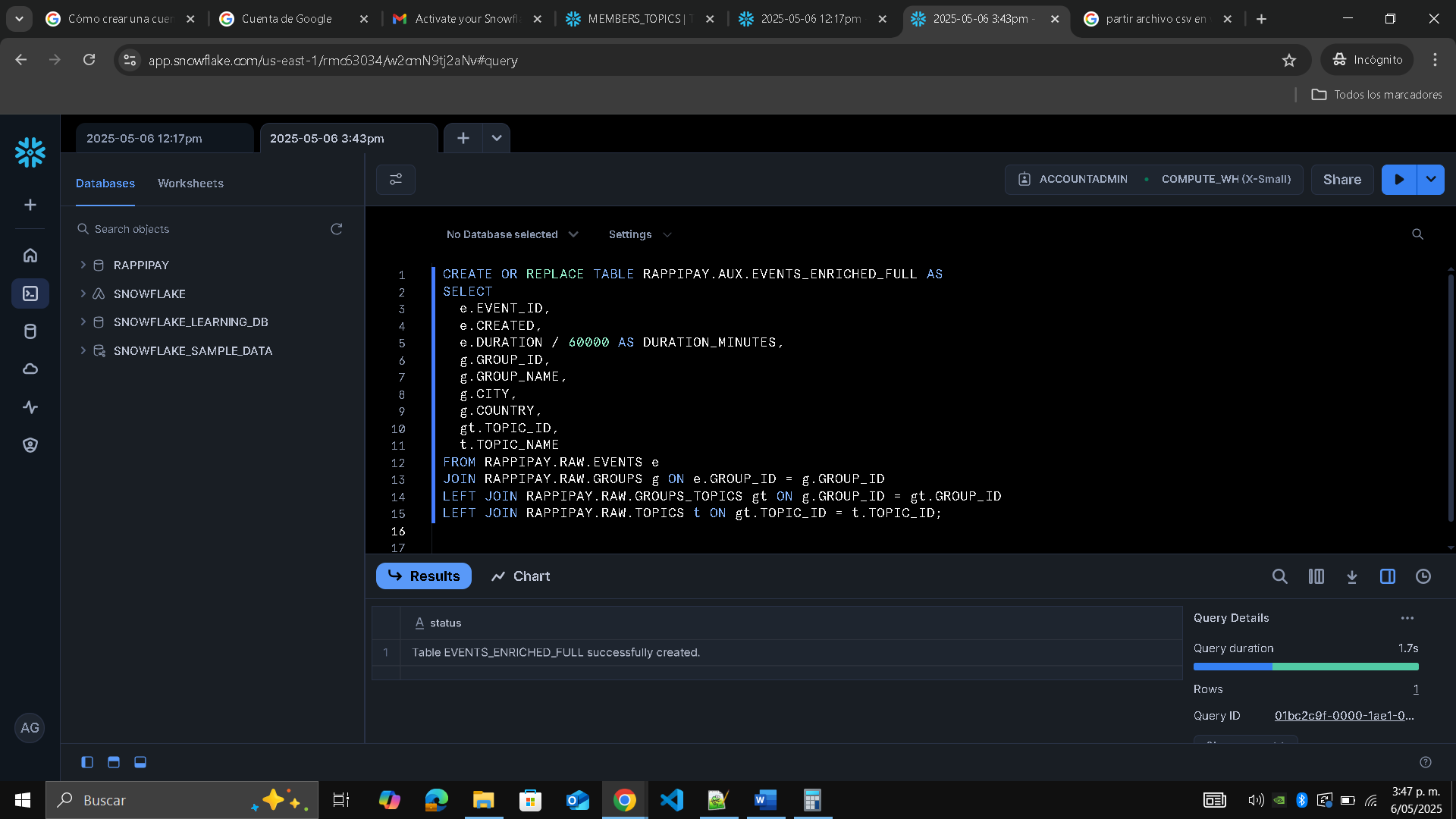
FROM RAPPIPAY.RAW.MEMBERS\_TOPICS mt

JOIN RAPPIPAY.RAW.MEMBERS m ON mt.MEMBER\_ID = m.MEMBER\_ID

GROUP BY mt.MEMBER\_ID, m.MEMBER\_NAME

HAVING COUNT(mt.TOPIC\_ID) > 5;

* La segunda combinar eventos con grupos y temas



Nombre : AUX.EVENTS\_ENRICHED\_FULL

CREATE OR REPLACE TABLE RAPPIPAY.AUX.EVENTS\_ENRICHED\_FULL AS

SELECT

e.EVENT\_ID,

e.CREATED,

e.DURATION / 60000 AS DURATION\_MINUTES,

g.GROUP\_ID,

g.GROUP\_NAME,

g.CITY,

g.COUNTRY,

gt.TOPIC\_ID,

t.TOPIC\_NAME

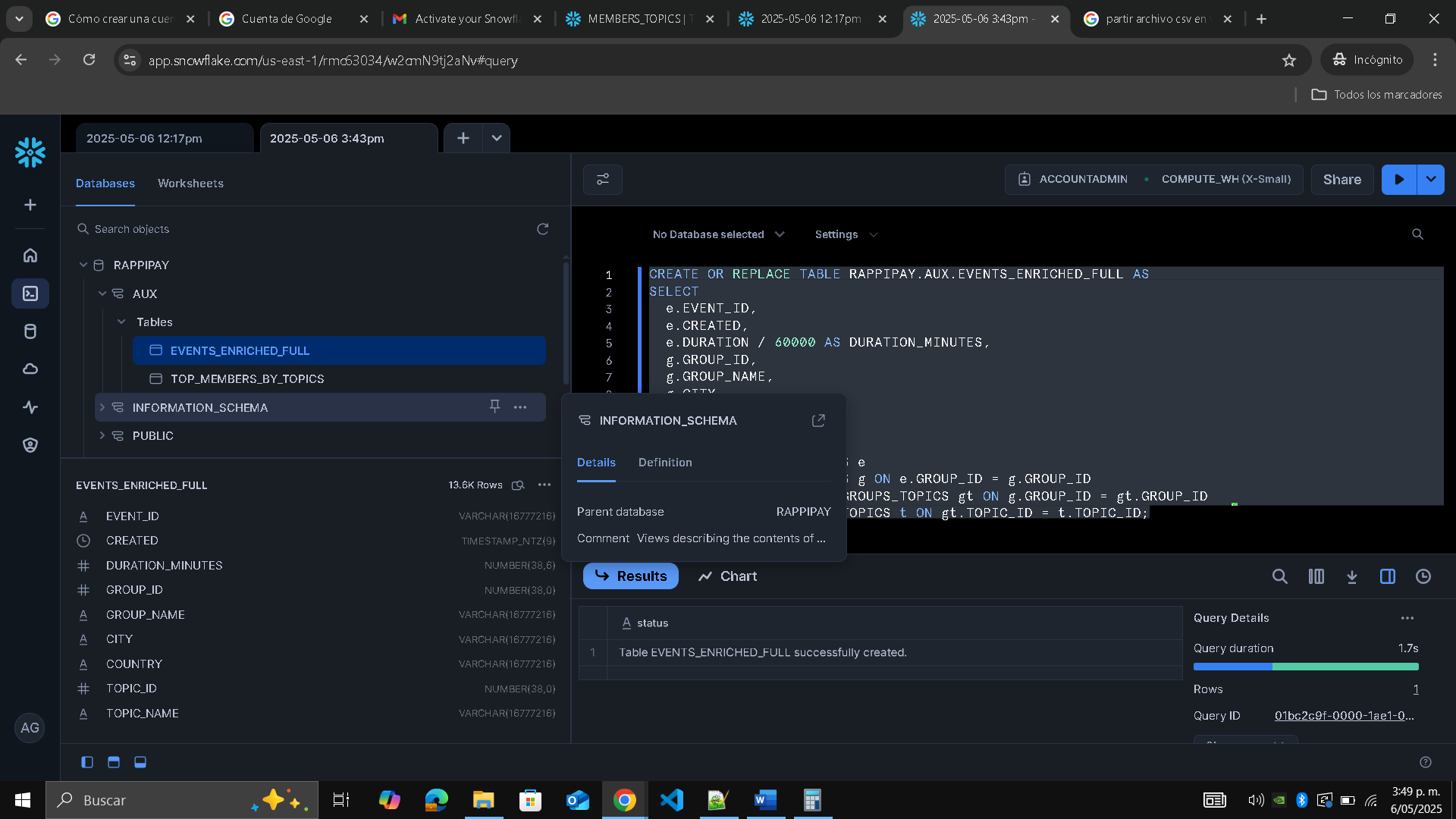
FROM RAPPIPAY.RAW.EVENTS e

JOIN RAPPIPAY.RAW.GROUPS g ON e.GROUP\_ID = g.GROUP\_ID

LEFT JOIN RAPPIPAY.RAW.GROUPS\_TOPICS gt ON g.GROUP\_ID = gt.GROUP\_ID

LEFT JOIN RAPPIPAY.RAW.TOPICS t ON gt.TOPIC\_ID = t.TOPIC\_ID;

* Tablas del nuevo esquema

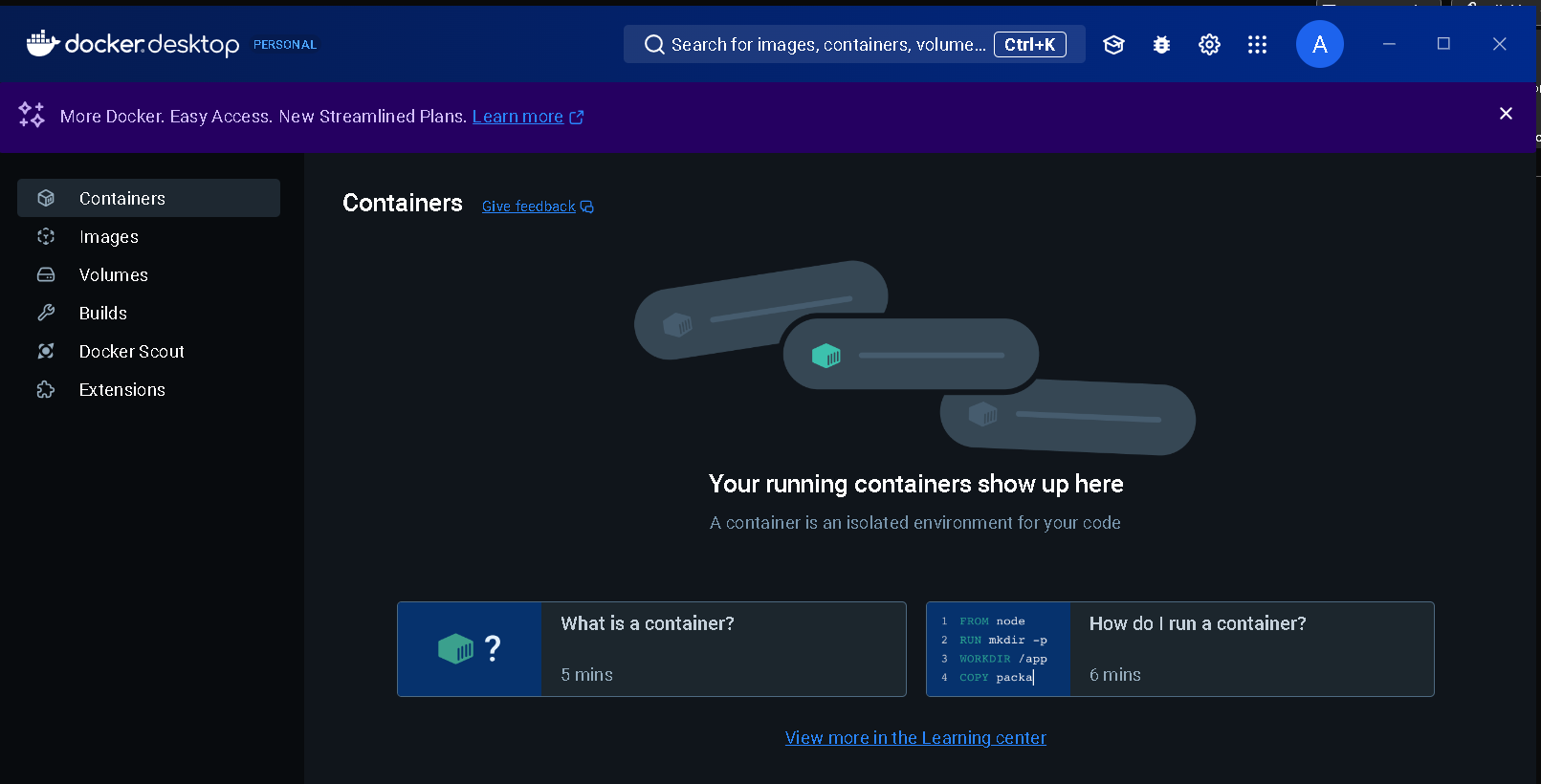


| **Tabla** | **Descripción** |
| --- | --- |
| TOP\_MEMBERS\_BY\_TOPICS | Muestra los miembros con más de 5 temas de interés |
| EVENTS\_ENRICHED\_FULL | Eventos enriquecidos con datos del grupo y temas asociados |

1. Desarrollo del DAG

De acuerdo con los costos que puede tener la nube se realiza el uso de Docker para preparar el servicio de air Flow de manera local

Se sube Docker



Hacemos la instalación y el montaje de la imagen para el contendor de AirFlow

Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.

Ingresamos al DAG

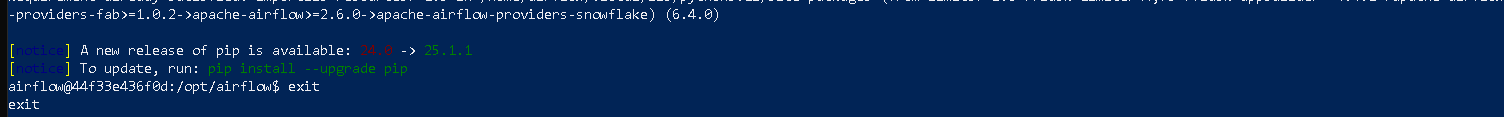
Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Y validamos que el servicio nos funcione y procedemos a configurarlo con snowflake, se instala en el contendor del apache-airflow-providers-snowflake

Interfaz de usuario gráfica, Aplicación, Word

El contenido generado por IA puede ser incorrecto.



1. Creación del DAG
2. Primero creamos la conexión

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

El contenido generado por IA puede ser incorrecto.

1. Se crea un Dag como prueba de conexión y validar que si lee los datos de las tablas de snowflake en la tabla CITIES

Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

1. Ahora vamos a desarrollar

* El objetivo en este punto es automatizar la creación de nuevas tablas auxiliares en Snowflake cada 15 minutos, integrando y limpiando la información para análisis. Usaremos funciones como MERGE, CREATE, y REPLACE.

La propuesta para crear las tablas auxiliares son las siguientes :

|  |  |  |
| --- | --- | --- |
| Nombre tabla Auxiliar | Transformación | Descripción |
| cleaned\_events | * Join entre events, groups, y venues. * Incluir ciudad, país y categoría del grupo. * Validar que duration y created estén presentes. | eventos enriquecidos con datos de grupos y locaciones |
| group\_topic\_summary | * Join entre groups\_topics y topics. * Contar número de grupos por tema. | resumen de tópicos por grupo |
| member\_interests | * Join entre members\_topics, members y topics. * Validar duplicados, y normalizar texto (topic\_name, bio). | intereses del miembro con enriquecimiento temático |
| group\_category\_stats | * Join con categories, agrupado por category\_name, cantidad de grupos, y total de eventos. | estadística de grupos por categoría |

Por lo que se debe realizar los siguientes pasos:

* Crear el DAG

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

* Crear los SQL para el desarrollo del script de las tablas auxiliares y export a bucket S3

|  |  |
| --- | --- |
|  |  |

Lo vemos en ejecución y el tiempo del file whatcher 15 minutos

Interfaz de usuario gráfica, Aplicación, Word

El contenido generado por IA puede ser incorrecto.

* Crear bucket S3

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

El contenido generado por IA puede ser incorrecto.

<https://us-east-1.console.aws.amazon.com/s3/buckets/airflow-rappy-bucket?region=us-east-1&bucketType=general&tab=objects>

Hacemos la configuración para que snowflake se conecte con S3 en la cuenta de AWS creada

Texto

El contenido generado por IA puede ser incorrecto.

1. Se valida la ejecución del DAG

Texto

El contenido generado por IA puede ser incorrecto.