

Llegamos hasta la **Tarea 4: Exploración de los campos de datos de comercio electrónico con la interfaz de usuario.**

**identificación y el entendimiento de la barra roja bajo la columna `productSKU`.** Esta barra roja indica la presencia de **valores no coincidentes** dentro de la columna.

Cloud Dataprep, al muestrear los datos, intenta identificar automáticamente el tipo de cada columna. En el caso de `productSKU`, es posible que, basándose en la muestra, lo haya determinado **incorrectamente como un tipo `Integer` (entero) en lugar de `String` (cadena de texto)**. Dado que los SKU de productos no siempre son solo números (por ejemplo, "GGOEGOCD078399" es un valor correcto), Dataprep marca los valores no enteros como no coincidentes, lo que se visualiza como una barra roja.

Este punto puede ser un "atasco" porque un usuario podría no comprender inmediatamente por qué la columna está marcada como "mismatched" (no coincidente) y qué implica esta identificación de tipo incorrecta por parte de la herramienta, aunque el laboratorio indica que se corregirá más adelante.

En cuanto a lo que entendimos del ejercicio y lo que haría el pipeline:

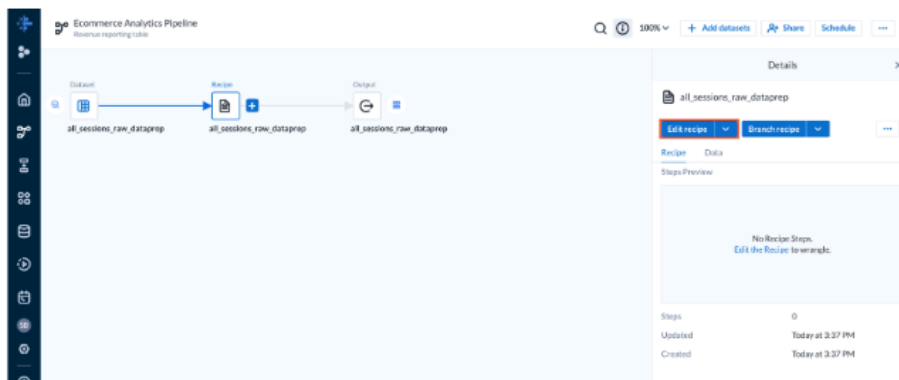
- El ejercicio tiene como objetivo principal la **creación de un pipeline de transformación de datos utilizando Cloud Dataprep**.
- Cloud Dataprep es un **servicio de datos inteligente diseñado para explorar, limpiar y preparar visualmente datos estructurados y no estructurados** con el fin de analizarlos.
- El conjunto de datos utilizado para el laboratorio es un **dataset de comercio electrónico** que contiene millones de registros de sesiones de Google Analytics para la Google Merchandise Store, cargados en BigQuery. Para este laboratorio específico, se utiliza un subconjunto de datos correspondiente a un solo día (aproximadamente 56 mil registros).
- El pipeline, en su esencia, conectaría este dataset de BigQuery a Dataprep, permitiría **explorar y mejorar la calidad de los datos**, crear una serie de pasos de transformación, y finalmente, ejecutar estos trabajos para **enviar los datos limpios y preparados de vuelta a BigQuery** para su análisis posterior.
- Hasta la Tarea 4, hemos aprendido a:
  - Configurar y acceder al entorno de laboratorio de Google Cloud.
  - Abrir y aceptar los términos de servicio de Dataprep en la Consola de Google Cloud.
  - **Crear un dataset en BigQuery** llamado `ecommerce` y cargar una submuestra de datos sin procesar (`all_sessions_raw_dataprep`) en él.
  - **Conectar Cloud Dataprep a esta fuente de datos de BigQuery**, creando un nuevo "flujo" (Flow) llamado "Ecommerce Analytics Pipeline".
  - Finalmente, en la **Tarea 4**, nos enfocamos en **explorar una muestra del dataset dentro de la vista Transformer de Cloud Dataprep**. Esto implica responder preguntas sobre el número de columnas y filas, identificar los valores más comunes en columnas clave como `channelGrouping`, los principales países de origen de las sesiones, y los productos y categorías

más populares. También se analiza la presencia de **valores faltantes** en columnas como `totalTransactionRevenue` o `itemQuantity`, se examina la distribución de valores (ej. `sessionQualityDim`), y se comprende el rango de fechas de los datos. Este proceso de exploración es crucial para **entender la calidad y las características de los datos** antes de proceder con las transformaciones necesarias para el pipeline.

## Task 4. Exploring ecommerce data fields with the UI

In this task, you will load and explore a sample of the dataset within Cloud Dataprep.

- Click the **Recipe icon** and then select **Edit Recipe**.



Cloud Dataprep loads a sample of your dataset into the Transformer view. This process might take a few seconds. You are now ready to start exploring the data!

Answer the following questions: