ESCUDERO ERICA

TRABAJO FINAL BOOTCAM DATA ENGENEERING

DICIEMBRE 2024

LAB “CLOUD DATAPREP”

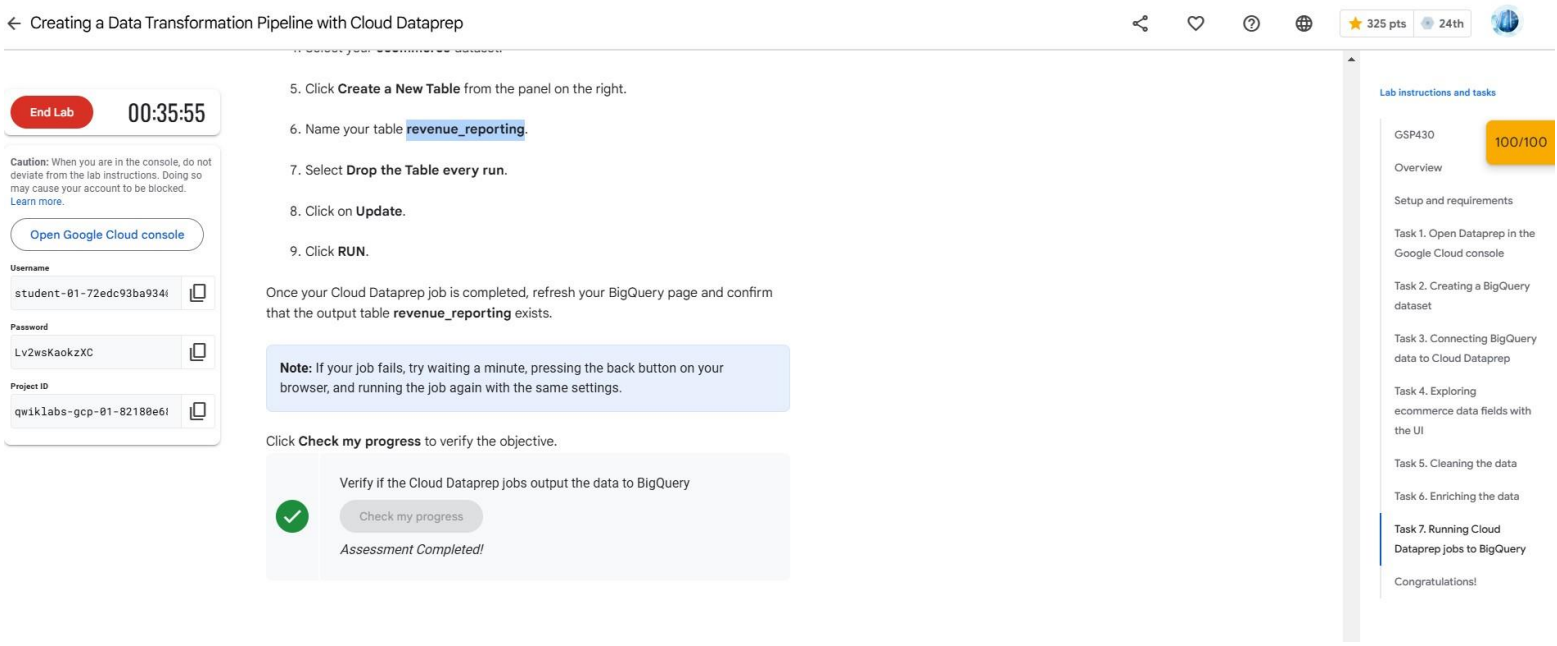
QuickLAB

Título: Creating a Data Transformation Pipeline with Cloud Dataprep

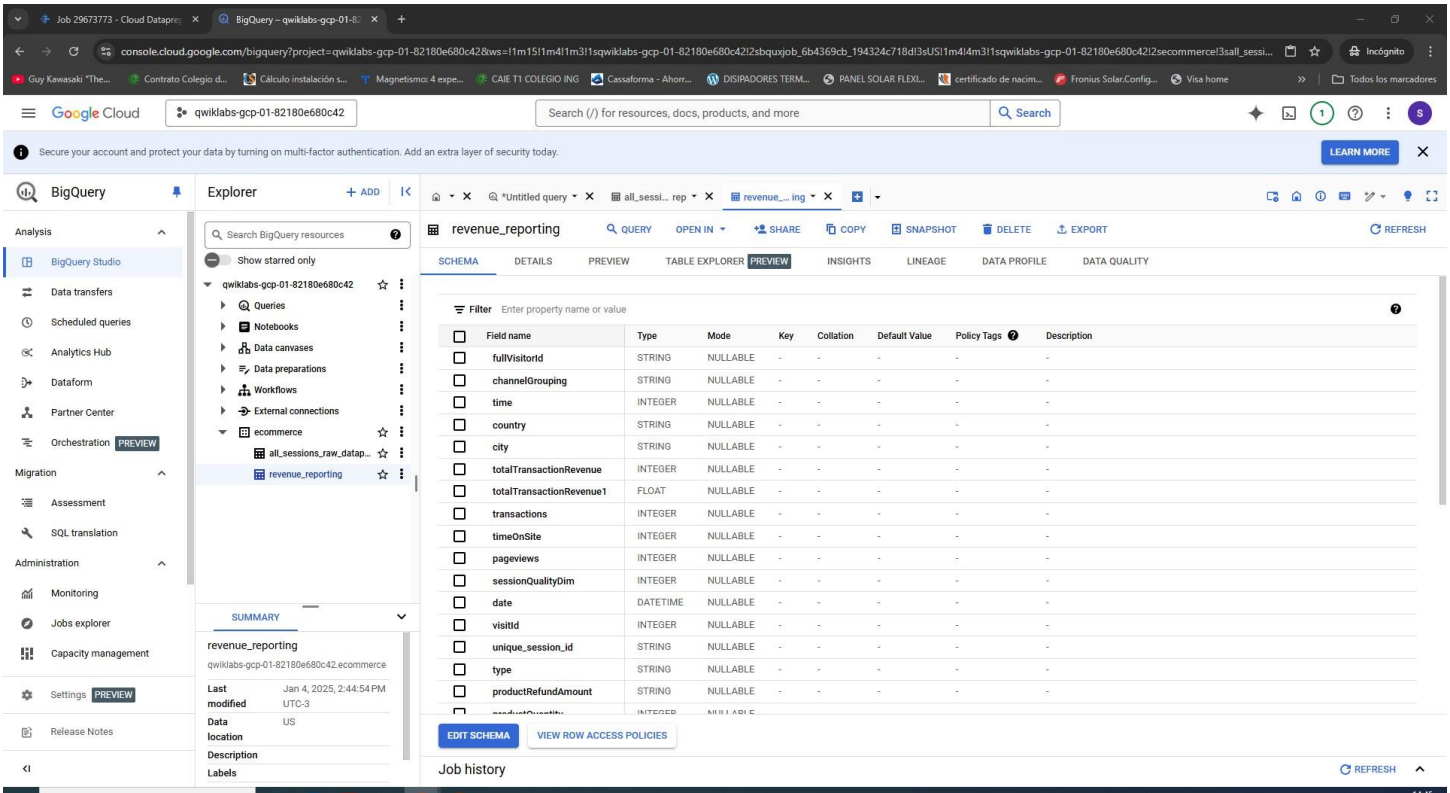
Schedule: 1 hour 15 minutes

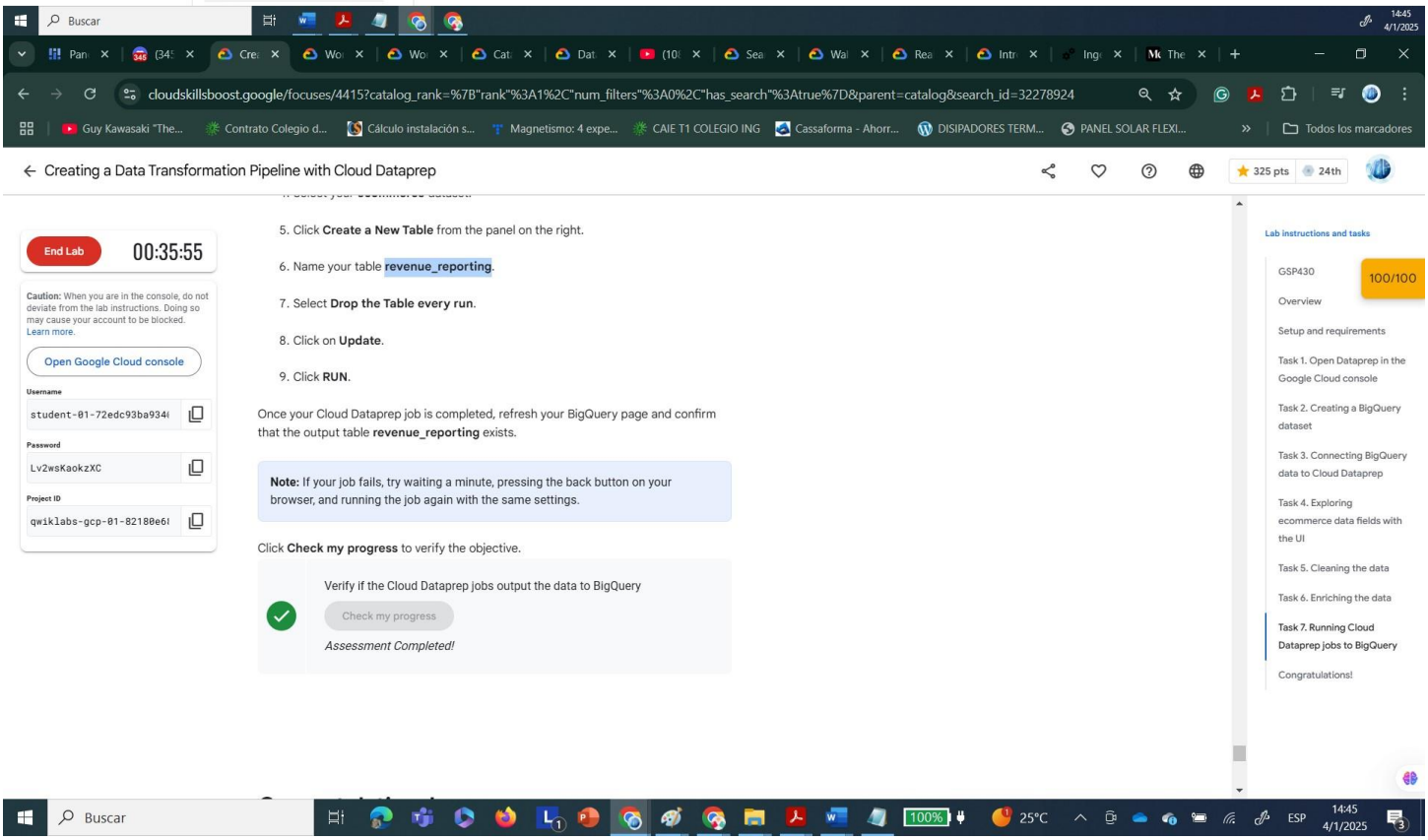
Cost: 5 Credits

Link: https://www.cloudskillsboost.google/focuses/4415?catalog_rank=%7B%22rank%22%3A1%2C%22num_filters%22%3A0%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=32278924



Captura de Pantalla del QuickLab terminado y de la tabla del transform en BQ

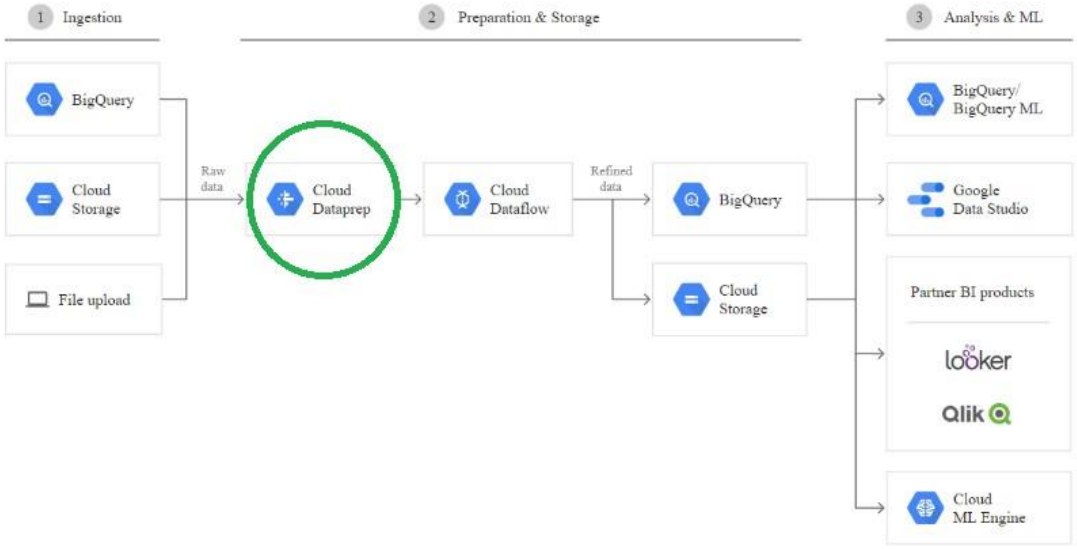




RESPUESTAS PREGUNTAS:

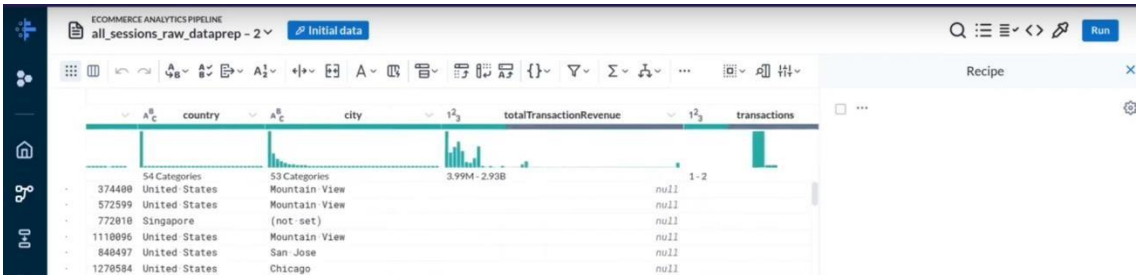
1. ¿Para qué se utiliza data prep?

Es un servicio de datos inteligente de tercero que está integrado con GCP que permite limpiar, explorar y preparar datos de forma gráfica. Permite construir un pipeline de transformación de datos, tomando los datos desde Big Querry(BQ) y luego escribe los resultados de esta transformación en Big Querry para luego hacer distinta análisis como Data Analytics o Science .



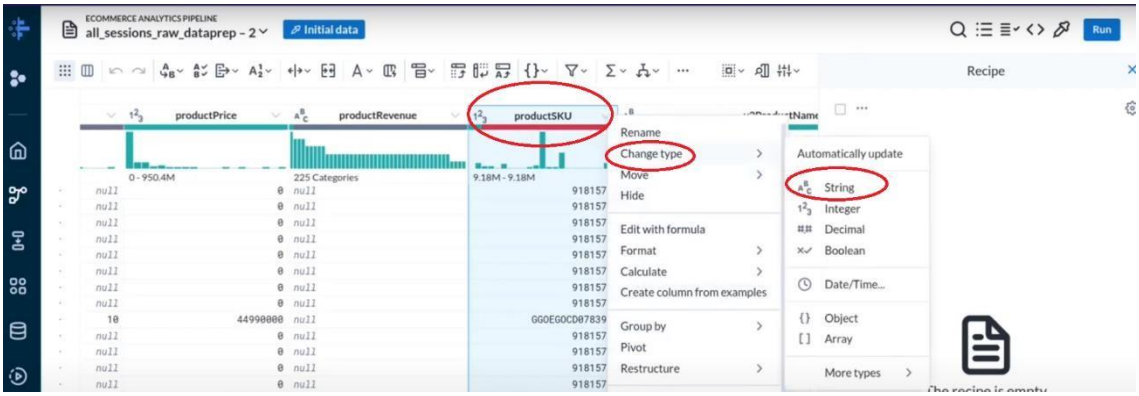
2. ¿Qué cosas se pueden realizar con DataPrep?

Se conecta a un Dataset en BQ y permite, de forma gráfica, explorar los campos a través de “recetas” pre-existentes mostrando cada columna y sus registros correspondientes. Al ser interactivo, pasando el mouse por arriba de la columna podemos ver la información de cada columna.

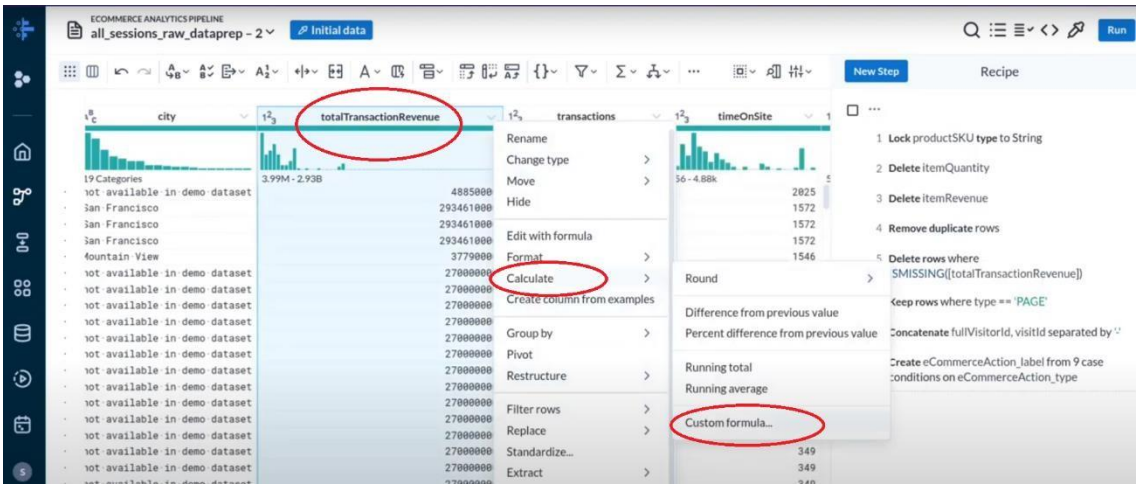


Una vez inspeccionado el dataset, se puede realizar:

- Eliminar columnas
- Cambiar el tipo de datos de las columnas
- Eliminar registros duplicados
- Crear cálculos calculados
- Y realizar distintos filtros sobre la información



Imágenes de DataPrep haciendo para transformaciones sobre los datos



3. ¿Por qué otra/s herramientas lo podrías reemplazar? Por qué?

Hay varias maneras de realizar las transformaciones de los datos, dentro de GCP.

BIGQUERY: permite realizar transformaciones directamente sobre SQL, permite hacer transformaciones complejas en la base de datos sin mover los datos y permite también escalar si el volumen de datos aumenta. A través de transformaciones basadas en consultas podemos limpiar datos duplicados, agregar columnas calculadas o crear vistas.

DATAPROC: permite realizar transformaciones de datos usando Apache Spark o Hadoop dentro de GCP a través de la ejecución de scripts en PySpark, HiveQL para grandes volúmenes de datos.

Cloud Functions o Cloud Run: se utiliza para transformaciones ligeras en Python o otru lenguaje disponible. Es fácilmente integrable con otros servicios de GCP. Se puede utilizar para la validación de datos antes de enviarlos a BigQuery.

Por qué usar estas herramientas:

Costos: Dataprep puede ser más ya que debe subcribirse un servicio de terv¿cero para poder utilizarlos y en operaciones continuas de procesamiento de datos puede elevar mucho el costo.

Flexibilidad: Herramientas como Dataflow y Dataproc ofrecen mayor personalización y control.

Integración: BigQuery y Dataflow, están mejor integradas con otros servicios de GCP.

Volumen de Datos: Dataprep tiene límites en el tamaño de los datos procesables en comparación con BigQuery o Dataflow.

4. ¿Cuáles son los casos de uso comunes de Data Prep de GCP?

Google Cloud **Dataprep** se puede usar para casos donde se necesitan realizar transformaciones, limpieza o preparación de datos de forma visual y sin escribir código. Para casos en los que los datos son desordenados, no estructurados o necesitan preprocesamiento antes de análisis o cargas en un sistema de almacenamiento como **BigQuery**.

Casos de uso: **Limpieza de datos desordenados:** incompletos con faltantes. Dataprep permite visualizar y corregir los errores rápidamente a través de la visualización.

Exploración de lo datos: para entender la calidad de los datos antes de procesarlos y anomalías que se puede ver en los gráficos estadísticos de cada columna.

Preparación para análisis: normalización de datos para análisis estadísticos del dataset. Agrupaciones y creación de columnas y filtros.

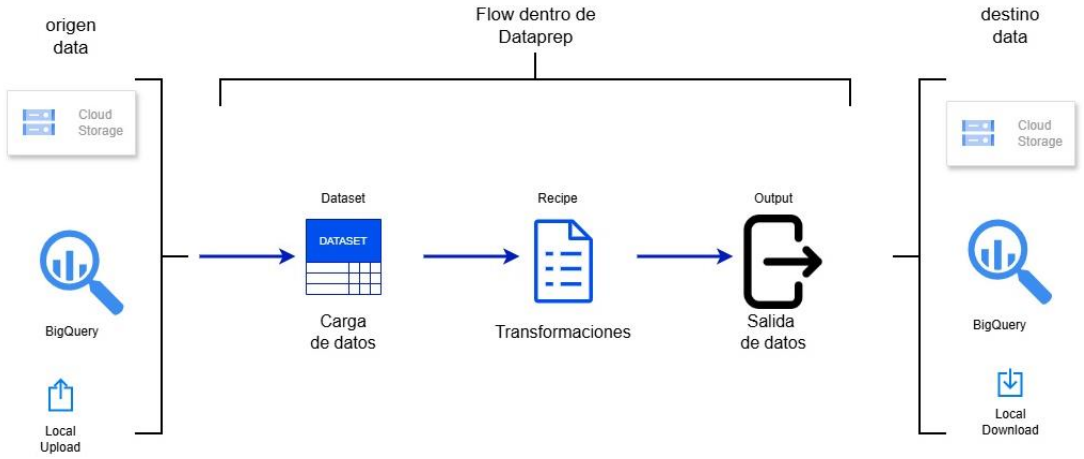
5. ¿Cómo se cargan los datos en Data Prep de GCP?

Los datos en Dataprep se pueden cargar desde:

- Cloud Storage
- BigQuery
- Subidas locales
- Google Sheets

Este sería el fujo del proceso:

Para crear un nuevo flujo se selecciona nuevo flujo en la herramienta y se el asigna un nombre y una descripción. Luego sobre el ícono de" Dataset" presionamos Add y cargamos los datasets según nuestro origen de datos. Elegimos "Create dataset", importamos y añadimos al Flow y se esa forma se cargan los datos. Luego de esto en "Recipe", editar podemos ver las opciones para transformar los datos y ya podemos ver los datos y en sus columnas el análisis estadístico. Dentro de esta sección se van agregando las diferentes transformaciones y una vez que están todas ya podemos presionar "RUN" y se abre la página de RUN JOB y elegimos el entrono de ejecución Trifacta Photon o DataFlow y también seleccionamos donde se guardaran las transformaciones, en BQ, Data Cloud etc. Y el tipo de acci´pn sobre la tabla y precionamos RUN para que se ejecute el trabajo. Una vez ejecutado podremos ver los datos trasnformados en la nueva tabla o archivo.



Esquema de flujo de trabajo en DATAPREP

6. ¿Qué tipos de datos se pueden preparar en Data Prep de GCP?

Google Cloud Dataprep está diseñada para preparar y transformar diversos tipos de datos en múltiples formatos y estructuras, facilitando su análisis y explotación. Es compatible con datos:

- estructurados,
- semiestructurados y • no estructurados.

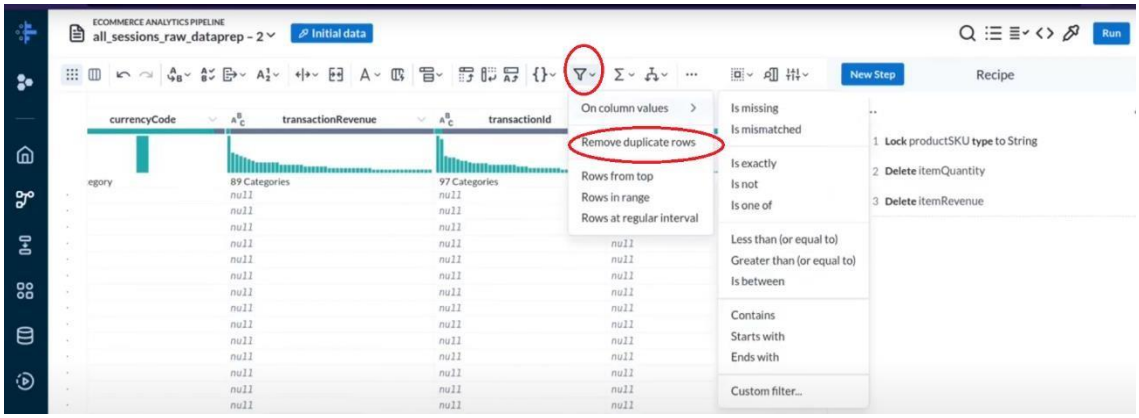
Incluye formatos comunes como CSV, JSON, Avro y Parquet, además de datos almacenados en bases de datos relacionales y sistemas de almacenamiento en la nube como BigQuery y Cloud Storage.

Dataprep maneja datos numéricos, categóricos, fechas, y texto libre, permitiendo transformaciones como la limpieza de valores faltantes, el ajuste de formatos, la eliminación de duplicados, y la corrección de inconsistencias.

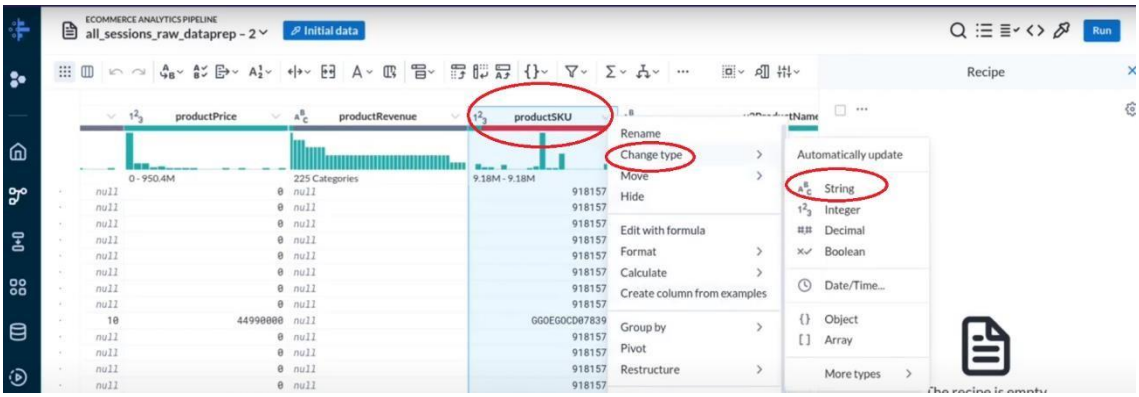
7. ¿Qué pasos se pueden seguir para limpiar y transformar datos en Data Prep de GCP?

Una vez conectado los datos, desde BQ o Cloud Storage, Dataprep nos facilita la visión de los datos y sus estadísticas sobre distribución, frecuencia y anomalías en los datos.

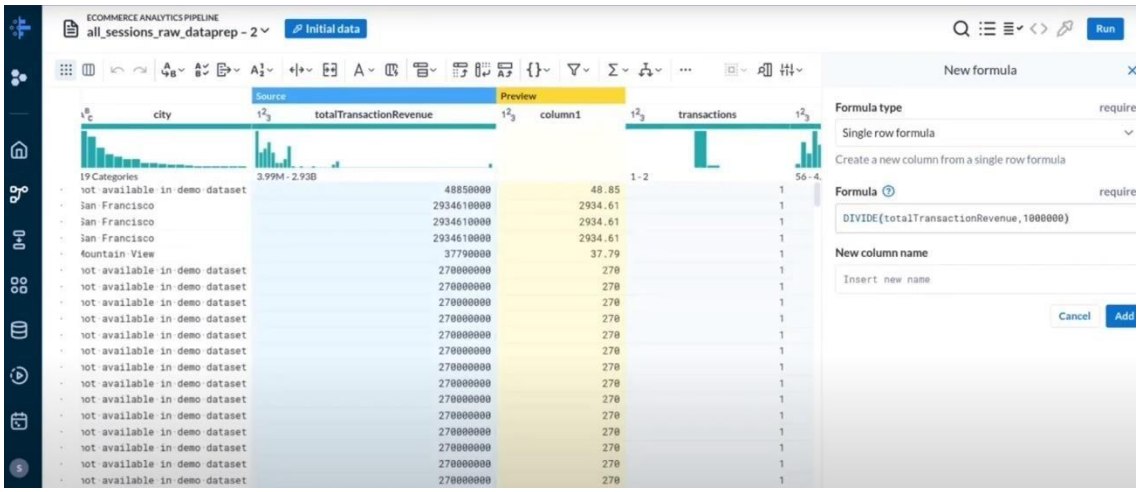
Se seleccionan los datos que sean nulos o duplicados con el ícono de filtrado y seleccionamos filas duplicadas o nulas y se eliminan.



Por ejemplo, para cambiar un tipo de dato a String podemos hacerlo de la siguiente manera:



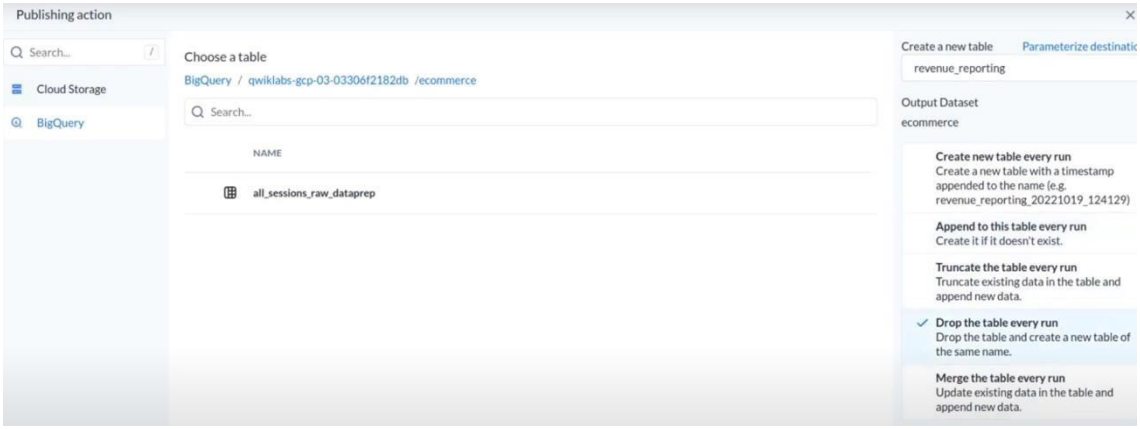
Se pueden hacer operaciones con los datos de una columna, por ejemplo, el valor de una columna se puede dividir, siempre y cuando el dato se numérico.



También se pueden hacer columnas calculadas, donde se agrega una nueva columna con un valor calculado resultado de alguna operación sobre los datos de una o varias columnas, renombrar columnas, crear nuevas columnas y ordenar.

8. ¿Cómo se pueden automatizar tareas de preparación de datos en Data Prep de GCP?

Una vez armado toda la transformación se ejecuta un JOB con Data Flow que ejecuta el pipeline sobre el dataset para que se gurade en BQ o Cloud Storage.



9. ¿Qué tipos de visualizaciones se pueden crear en Data Prep de GCP?

Dataprep permite explorar datos de manera interactiva, identificar problemas comunes como valores faltantes, duplicados, o inconsistencias, y realizar transformaciones mediante una interfaz visual. Permite ver el perfil de los datos a través de gráfico integrado proporciona estadísticas descriptivas y gráficos que destacan patrones y anomalías, lo que nos permite tomar decisiones sobre qué aspectos deben corregirse.

10. ¿Cómo se puede garantizar la calidad de los datos en Data Prep de GCP?

Dataprep permite explorar datos de manera interactiva, identificar problemas comunes como valores faltantes, duplicados, o inconsistencias, y realizar transformaciones mediante una interfaz visual. Permite ver el perfil de los datos a través de gráfico integrado proporciona estadísticas descriptivas y gráficos que destacan patrones y anomalías, lo que nos permite tomar decisiones sobre qué aspectos deben corregirse a través de la definición del pipeline de transformación.

Al ver en cada paso de la transformación como se grafican los gráficos estadísticos se puede ir supervisando el resultado de la transformación y detectar anomalías a corregir. Esto garantiza la calidad final de los datos.

Arquitectura:

El gerente de Analítica te pide realizar una arquitectura hecha en GCP que contemple el uso de esta herramienta ya que le parece muy fácil de usar y una interfaz visual que ayuda a sus desarrolladores ya que no necesitan conocer ningún lenguaje de desarrollo.

Esta arquitectura debería contemplar las siguientes etapas:

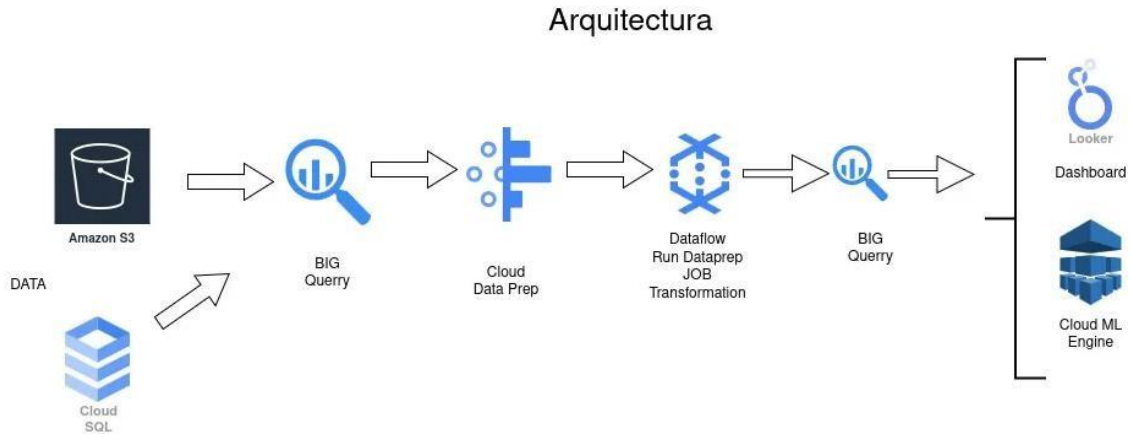
Ingesta: datos parquet almacenados en un bucket de S3 y datos de una aplicación que guarda sus datos en Cloud SQL.

Procesamiento: filtrar, limpiar y procesar datos provenientes de estas fuentes

Almacenar: almacenar los datos procesados en BigQuery

BI: herramientas para visualizar la información almacenada en el Data Warehouse

ML: Herramienta para construir un modelo de regresión lineal con la información almacenada en el Data Warehouse



Descripción de la solución propuesta:

La arquitectura diseñada para análisis de datos en GCP es una solución modular, escalable y eficiente que integra múltiples herramientas y servicios para procesar datos de diversas fuentes y obtener insights avanzados para el área que ha solicitado esta arquitectura.

El flujo comienza con la ingesta de datos desde fuentes externas como **Amazon S3** y **Cloud SQL**, que se consolidan en **BigQuery** como almacenamiento centralizado de data.

En **BigQuery**, los datos se almacenan en su forma bruta, donde están disponibles para consultas iniciales y preparación. Posteriormente, **Cloud Dataprep** se utiliza para realizar transformaciones visuales e interactivas, como limpieza de datos, cambios de formatos, creación de nuevas columnas calculadas según los requerimientos del negocio y creación de nuevas estructuras. Las reglas definidas en Dataprep se implementan mediante **Dataflow**, lo que permite procesar grandes volúmenes de datos de forma automatizada y escalable, cargando los datos transformados nuevamente en BigQuery.

Una vez que los datos están preparados, se guardan en nuevas tablas en BigQuery y permite habilitar consultas analíticas avanzadas y prepara los datos para ser consumidos por herramientas como **Looker**, que proporciona visualizaciones interactivas y dashboards dinámicos.

Adicionalmente, los datos transformados pueden alimentar **Cloud ML Engine**, donde se desarrollan y despliegan modelos de machine learning para análisis predictivo y automatización inteligente.

Esta arquitectura garantiza:

- **Eficiencia** en el manejo de grandes volúmenes de datos mediante BigQuery y Dataflow.
- **Flexibilidad** al integrar múltiples fuentes de datos externas e internas.
- **Automatización** del proceso de transformación con pipelines repetibles y confiables.
- **Escalabilidad** para soportar necesidades de datos crecientes y análisis complejos.

Es una propuesta de solución técnica robusta diseñada para optimizar el flujo de datos y habilitar análisis para la obtención de información relevante para la emrpesa, permitiendo a los equipos técnicos manejar datos de extremo a extremo de manera eficiente.