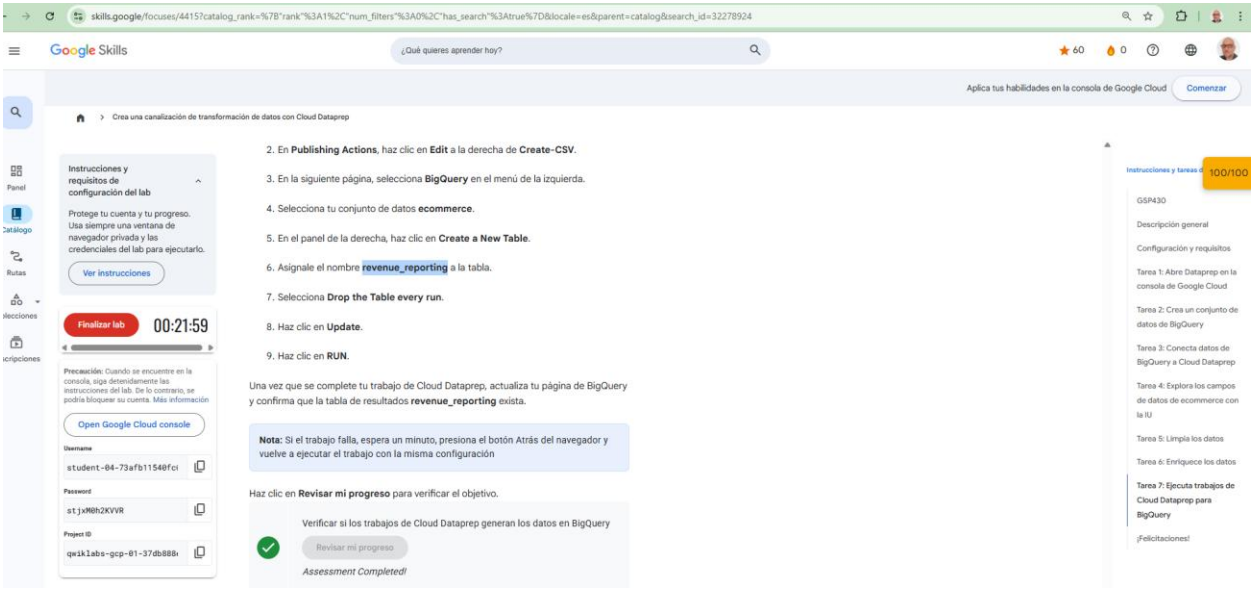
Ejercicio 3:



Cloud Dataprep *by Trifacta* en GCP (complementa el lab Creating a Data Transformation Pipeline with Cloud Dataprep)

1. ¿Para qué se utiliza Data Prep?

Data Prep se utiliza principalmente para **preparar, explorar y perfilar datos** antes de que sean utilizados en análisis, *machine learning* (ML) o cargados en un *data warehouse* (como BigQuery) o un *data lake*.

Su misión es triple:

1. **Explorar y perfilar datos** para entender su calidad y estructura.
2. **Limpiar, transformar y combinar** fuentes de datos heterogéneas.
3. **Generar *datasets* listos** para su consumo analítico.

Un detalle clave de Dataprep es su **enfoque visual e interactivo**. Permite a los usuarios **iterar rápidamente** sobre una muestra de datos (gracias al **muestreo inteligente** de Trifacta) y ver los resultados de las transformaciones **antes** de ejecutar el *job* completo sobre todo el volumen de datos. Esto acelera significativamente el ciclo de desarrollo y prueba.

2. ¿Qué cosas se pueden realizar con DataPrep?

Dataprep permite realizar todas las tareas de **Data Wrangling** o manipulación de datos de forma visual. Entre las tareas principales se encuentran:

- **Detección de tipos de datos y *data profiling*** (análisis de distribuciones, valores faltantes, *outliers*).
- **Limpieza:** Eliminación de filas/columnas, manejo de valores nulos, y estandarización de formatos (fechas, números, categorías).
- **Transformaciones:** Aplicar filtros, realizar *joins* (uniones), *unions* (concatenaciones), agregaciones, *pivots* y *unpivots*, y derivar columnas calculadas.
- **Enriquecimiento:** Combinación con otras tablas, normalización y *lookups*.
- **Orquestación de flujos:** Diseñar **"flows"** y **"recipes"** que son secuencias reproducibles de transformación.

La herramienta se distingue por su **interfaz basada en sugerencias inteligentes ("Wrangle")**. Cuando el usuario selecciona un patrón o un valor inconsistente en los datos, Dataprep sugiere automáticamente las transformaciones de limpieza de datos más probables (por ejemplo, reemplazar, dividir, extraer), lo que simplifica enormemente la creación de reglas complejas para usuarios sin experiencia en programación.

---

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

Las herramientas alternativas a Dataprep se dividen típicamente en dos categorías:

1. **Herramientas ETL/ELT visuales (Low-Code/No-Code):**
    - **Ejemplos:** Cloud Data Fusion (en GCP), Talend, Informatica, Alteryx Designer.
    - **Motivo:** También ofrecen una interfaz gráfica para diseñar *pipelines* de transformación, poseen una amplia gama de conectores y ejecutan la lógica de manera escalable. **Cloud Data Fusion** es el reemplazo más directo en GCP, siendo preferido para *pipelines* de integración de datos más robustos y empresariales (con soporte para réplica y gobierno centralizado), mientras que Dataprep se elige por su mayor facilidad de uso en la **limpieza de datos *ad-hoc***.
2. **Herramientas de código/Programación:**

- **Ejemplos:** *Notebooks* con PySpark, SQL en BigQuery, Dataflow (Apache Beam), dbt.
- **Motivo:** Ofrecen **mayor control**, facilitan el **versionado** en Git y, en muchos casos, pueden ser más **rentables** a una escala muy grande. Se utilizan cuando el equipo prioriza la estandarización en Integración Continua y Entrega/Despliegue Continuo (CI/CD) o requiere lógica de transformación muy específica que solo es posible programando.

---

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

Los casos de uso comunes aprovechan la velocidad y accesibilidad de la herramienta en el ecosistema de Google Cloud:

- **Ingesta y Limpieza de Datos:** Limpieza y estandarización de datos crudos (por ejemplo, *logs*, datos de sensores, *e-commerce*) antes de cargarlos en BigQuery o Cloud Storage.
- **Preparación para ML:** Generación de *datasets* limpios y con *feature engineering* previo para modelos de *machine learning* en Vertex AI.
- **Unificación y BI:** Unificación de datos provenientes de múltiples fuentes (marketing, finanzas, RRHH) para crear fuentes de verdad únicas que alimenten *dashboards* en Looker/Looker Studio.
- **Transformaciones delegadas:** Ejecución de transformaciones complejas delegadas directamente al motor SQL de BigQuery (*BigQuery pushdown*) para aprovechar su procesamiento masivamente paralelo.

Un caso de uso específico es la **preparación de datos para series temporales**, donde la limpieza de *gaps* (vacíos) o la interpolación de valores son pasos críticos. La capacidad de Dataprep para manejar funciones de ventana y agregaciones de tiempo de forma visual simplifica este tipo de *feature engineering* previo al ML.

---

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

Los datos se cargan importándolos a un **Flow** (Flujo de trabajo) en la consola de Dataprep.

El flujo de carga es:

1. Crear o abrir un **Flow** en Dataprep.

2. Añadir un *dataset* de entrada, importándolo desde:
    - **Cloud Storage:** Indicando la ruta gs://bucket/... de archivos (CSV, JSON, etc.).
    - **BigQuery:** Seleccionando una tabla existente, o usando una *query* SQL.
    - **Otras fuentes** compatibles.
3. Dataprep genera una **muestra de datos** (utilizada para el diseño de la *recipe* de transformación) sobre la cual se empieza a trabajar.

El proceso es asistido por el **catálogo de datos de GCP**. Dataprep utiliza la información de metadatos (esquema, formato, etc.) de las fuentes conectadas a GCP para cargar el *dataset*. Si es un archivo plano, la herramienta ayuda a **inferir su esquema** con opciones avanzadas de delimitadores o codificación, facilitando la importación inicial.

---

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

Dataprep está optimizado para trabajar con **datos tabulares**:

- **Archivos estructurados/semi-estructurados** en Cloud Storage (CSV, JSON, logs delimitados, TSV).
- **Tablas de BigQuery** o resultados de *queries* SQL.
- Datos tabulares generales (numéricos, categóricos, fechas, textos cortos).

La herramienta se especializa en la preparación de **datos tabulares anidados o jerárquicos** (típicos de archivos JSON o logs). Dataprep tiene funciones visuales clave para **"aplanar" (*un-nest*)** estas estructuras complejas en filas y columnas manejables, lo cual es un paso de limpieza y normalización fundamental en entornos de Big Data.

---

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

El trabajo se desarrolla definiendo una **"recipe"** (receta) de pasos que actúan sobre la muestra de datos:

1. **Importar el *dataset*** e iniciar el trabajo sobre la muestra.
2. **Explorar el perfil de datos** para identificar anomalías (tipos detectados, nulos, *outliers*).
3. **Definir la *recipe*** añadiendo pasos de transformación:

    - Normalizar tipos (por ejemplo, *cast* de *string* a *date*).
    - Filtrar filas inválidas o duplicadas.
    - Rellenar o eliminar nulos (*missing values*).
    - Dividir/concatenar columnas o extraer patrones con reglas y expresiones regulares.
    - Realizar *joins* y agregaciones.
4. **Previsualizar** el resultado sobre la muestra.
5. **Definir el *output*** (destino en BigQuery o Cloud Storage).
6. **Ejecutar el *job*** sobre el *dataset* completo utilizando Dataflow o BigQuery.

Cada paso de la "recipe" genera una línea en el panel lateral, actuando como un **registro de auditoría inmutable** de la transformación. Esto es fundamental para la **trazabilidad de los datos**: el usuario puede volver a cualquier paso intermedio, modificarlo, y el resto del *flow* se recalcula automáticamente, garantizando la reproducibilidad.

---

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

Una vez que un **Flow** y su **Recipe** están validados, se pueden ejecutar de forma repetible y automatizada:

- **Jobs Programados:** Se configuran **Jobs** recurrentes para que corran a ciertas horas o en respuesta a la llegada de nuevos datos (por ejemplo, si nuevos archivos llegan diariamente a un *bucket* de Cloud Storage).
- **Integración con Orquestación:** Los *jobs* de Dataprep se integran con herramientas de orquestación de GCP, como **Cloud Composer/Airflow**, llamándolos desde *DAGs* (Directed Acyclic Graphs).
- **Activadores Event-Driven:** Se pueden usar **Cloud Functions** para disparar la ejecución de un *Job* de Dataprep cada vez que se detecta un evento específico (por ejemplo, un nuevo archivo cargado en GCS).

La automatización se facilita con la **API de Dataprep** (o la línea de comandos gcloud). Esto permite que los *Jobs* sean gestionados y monitoreados desde *scripts* de DevOps o integrados en sistemas de monitoreo y despliegue continuo (CI/CD) externos a la interfaz gráfica.

---

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

Dataprep **no es una herramienta de *Business Intelligence* (BI)** como tal, sino que utiliza visualizaciones de apoyo para el proceso de *data wrangling*:

- **Visualizaciones de Distribución:** Histogramas, distribuciones y conteos por columna para entender la forma de los datos.
- **Gráficos de Calidad:** Gráficos que muestran la proporción de valores nulos, valores únicos (*cardinality*) y *outliers*.

Las visualizaciones son parte integral del **Data Profiling**. El sistema utiliza colores y gráficos de barras (el famoso *ribbon* sobre la columna) para indicar visualmente la **calidad de los datos**. Por ejemplo, un color puede indicar la proporción de valores que **no** coinciden con el tipo de dato esperado, permitiendo a los usuarios identificar problemas de calidad a simple vista y priorizar la limpieza.

---

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

La garantía de calidad de los datos se logra mediante una combinación de automatización y trazabilidad:

- **Perfilado Automático:** Detecta y resalta anomalías (*outliers*, tipos inconsistentes, nulos) al cargar la muestra, permitiendo su corrección inmediata con reglas explícitas.
- **Recetas Versionadas:** Todo el proceso de limpieza y transformación queda documentado, versionado y reproducible, lo que facilita la auditoría y evita errores de reprocesamiento.
- **Validaciones Explícitas:** Se pueden definir reglas de validación (*data quality checks*) dentro de la *recipe* que marcan o descartan valores fuera de rango o inconsistentes.
- **Motor Gestionado:** La ejecución sobre Dataflow o BigQuery reduce errores de infraestructura y escalado, asegurando que las transformaciones se apliquen de forma consistente al volumen total.

La calidad también se garantiza con las **métricas de salida de los Jobs**. Al finalizar la ejecución, el usuario recibe un resumen detallado sobre **cuántas filas se procesaron** y, más importante, **cuántas se descartaron** debido a las reglas de limpieza definidas. Esto proporciona una métrica cuantitativa de la efectividad de la limpieza del *pipeline*.

Arquitectura:

El gerente de Analitca 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 siguiente 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