

Google Skills Boost - Examen Final

LAB:

← Creating a Data Transformation Pipeline with Cloud Dataprep

10 pts

End Lab

00:20:38

Caution: When you are in the console, do not deviate from the lab instructions. Doing so may cause your account to be blocked.

Learn more

Open Google Cloud console

Username

student-02-5f41b56caa

Password

x1cx5Xh1C0N6

Project ID

qsiklabs-gcp-03-f8def8

Note: If your job fails, try waiting a minute, pressing the back button on your browser, and running the job again with the same settings.

Click **Check my progress** to verify the objective.

Verify if the Cloud Dataprep jobs output the data to BigQuery

Check my progress

Assessment Completed!

Congratulations!

You've successfully explored your ecommerce dataset and created a data transformation pipeline with Cloud Dataprep.

Next steps / Learn more

Lab instructions and tasks

GSP430

100/100

Overview

Setup and requirements

Task 1. Open Dataprep in the Google Cloud console

Task 2. Creating a BigQuery dataset


Task 3. Connecting BigQuery data to Cloud Dataprep

Task 4. Exploring ecommerce data fields with the UI

Task 5. Cleaning the data

Task 6. Enriching the data

Task 7. Running Cloud Dataprep jobs to BigQuery



martin lopez

Member since 2024

1550 points

Your profile is public and accessible at https://www.cloudskillsboost.google/public_profiles/d5614013-6c6f-4136-9eb4-59367e21a766

Make profile private

Paths

Activities

Leaderboard

Badges

Course

Lab

Quiz

Game

In progress

Finished

Activity	Type	Date started	Date finished	Score	Passed
Creating a Data Transformation Pipeline with Cloud Dataprep	Lab	1 hour ago	0 minutes ago	Assessment: 100%	✓

Contestar las siguientes preguntas:

- ¿Para que se utiliza data prep?**
Dataprep es una herramienta de ETL que se utiliza para la exploración, limpieza y transformación de datos antes de su análisis.
- ¿Qué cosas se pueden realizar con DataPrep?**
Se pueden borrar columnas, crear columnas concatenando valores, crear campos calculados, eliminar duplicados, eliminar filas no deseadas, cambiar el tipo de datos, etc.
- ¿Por qué otra/s herramientas lo podrías reemplazar? Por qué?**
Se podría reemplazar por Kettle (Pentaho), Alteryx o AWS Glue. Todas estas herramientas permiten trabajar a través de una Interfaz Gráfica con Drag and Drop.
- ¿Cuáles son los casos de uso comunes de Data Prep de GCP?**
Los casos más comunes son obtener datasets de distintas fuentes; preparar los datos ajustando los tipos, eliminar duplicados y/o registros no deseados, etc. Luego cargar los datos transformados en Big Query disponiendo los datos para su análisis, generación de dashboards o ML.

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

Dataprep se integra con Big Query, Cloud Storage, Google Sheets, subidas directas, entre otros conectores para la carga de datos.

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

Estructurados, semiestructurados y no estructurados.

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

Pasos:

1-Importar los datos de distintas fuentes.

2-Visualizar Datasets Cargados. (Validar contenido)

3-Crear "Recipe". (Si es necesario pueden crearse varios "Recipe" diviendo las transformaciones).

4- Click "Edit Recipe".

5- Realizar transformaciones. (Pueden validarse los cambios con el menú de seguimiento).

6-Click en "Run".

7-Validar que los datos cargados tengan el resultado esperado.

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

"Plan" permite la ejecución de una secuencia de tareas en la plataforma, como flujos de preparación de datos. Además, cuenta con la opción de Schedule para automatizar las ejecuciones.

Otra opción es programar la ejecución del flujo seleccionado directamente desde "Flows".

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

Dataprep está diseñada para la preparación/transformación y carga de datos.

Para visualizaciones puede utilizarse una herramienta alternativa como Looker Studio.

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

Dataprep proporciona múltiples mecanismos para transformar y estandarizar datos para satisfacer las necesidades de uso, incluidas visualizaciones de perfiles y barras de calidad basadas en tipos para identificar posibles anomalías y problemas de calidad. Las comprobaciones de calidad de datos pueden aplicarse durante la importación, transformación o exportación de datos.

Algunos de ellos:

- Validación de esquemas
- "Data quality bar"
- Histogramas por columnas
- "Data quality rules"
- Estandarización

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 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

