Contestar las siguientes preguntas:

1. ¿Para que se utiliza data prep?

Dataprep, y en particular **Google Cloud Dataprep by Alteryx**, es una herramienta de autoservicio diseñada para simplificar y acelerar el proceso de preparación de datos. Su propósito fundamental es permitir a analistas de datos, científicos de datos y usuarios de negocio, explorar, limpiar, transformar y enriquecer grandes volúmenes de datos crudos y desestructurados, dejándolos listos para el análisis, la generación de informes y la aplicación en modelos de aprendizaje automático (machine learning).

2. ¿Qué cosas se pueden realizar con DataPrep?
Exploración y perfilado visual de datos, limpieza de datos intuitiva, transformación de datos sin código, enriquecimiento de datos.

3. ¿Por qué otra/s herramientas lo podrías reemplazar? Por qué?
Dataproc es una alternativa si se trabaja con grandes volúmenes de datos,ya que está basado en Hadoop y Spark,

4. ¿Cuáles son los casos de uso comunes de Data Prep de GCP?
El caso de uso más fundamental de Dataprep es la creación de flujos de trabajo de Extracción, Transformación y Carga (ETL) o Extracción, Carga y Transformación (ELT).

5. ¿Cómo se cargan los datos en Data Prep de GCP?
Se puede cargar un archivo .csv., .json. o .txt desde la máquina local, desde Google Cloud Storage o desde Google Query.

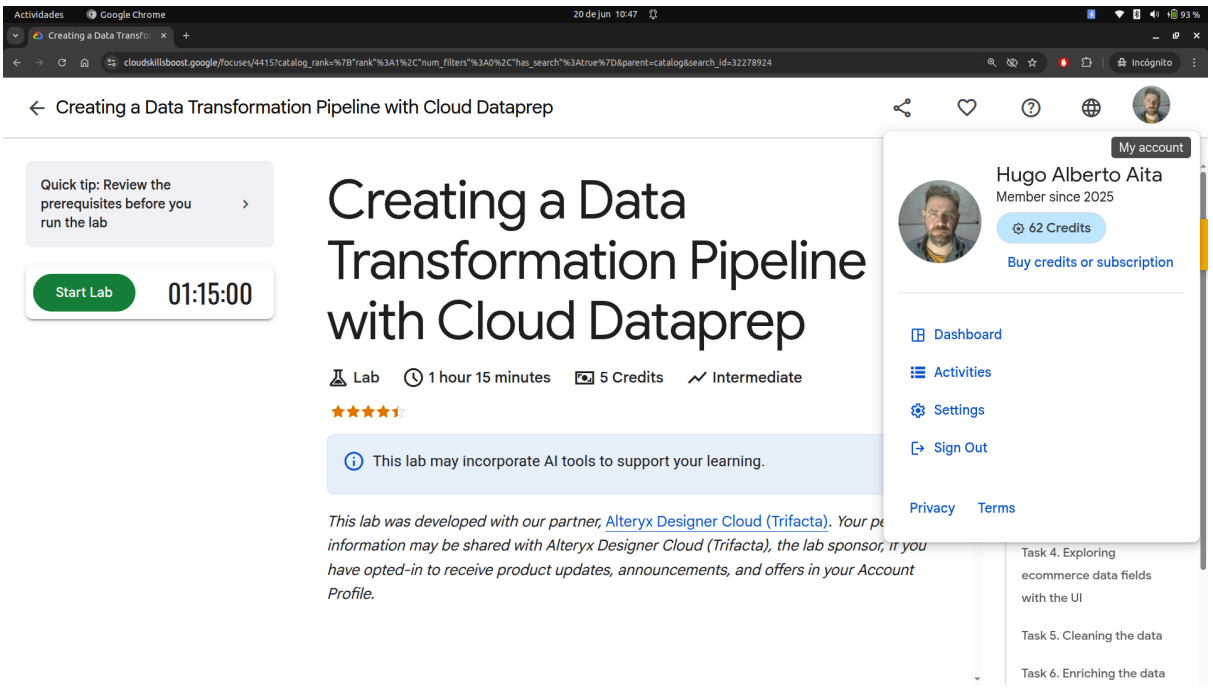
6. ¿Qué tipos de datos se pueden preparar en Data Prep de GCP?
Dataprep reconoce los tipos de datos fundamentales String, Integer, Decimal, Boolean, Date/Time, ip address y url. También reconoce archivos json, array, y tipos de datos “inteligentes” como ser email, Número de Tarjeta de Crédito, Social Security Number, Género, ZIP Code.

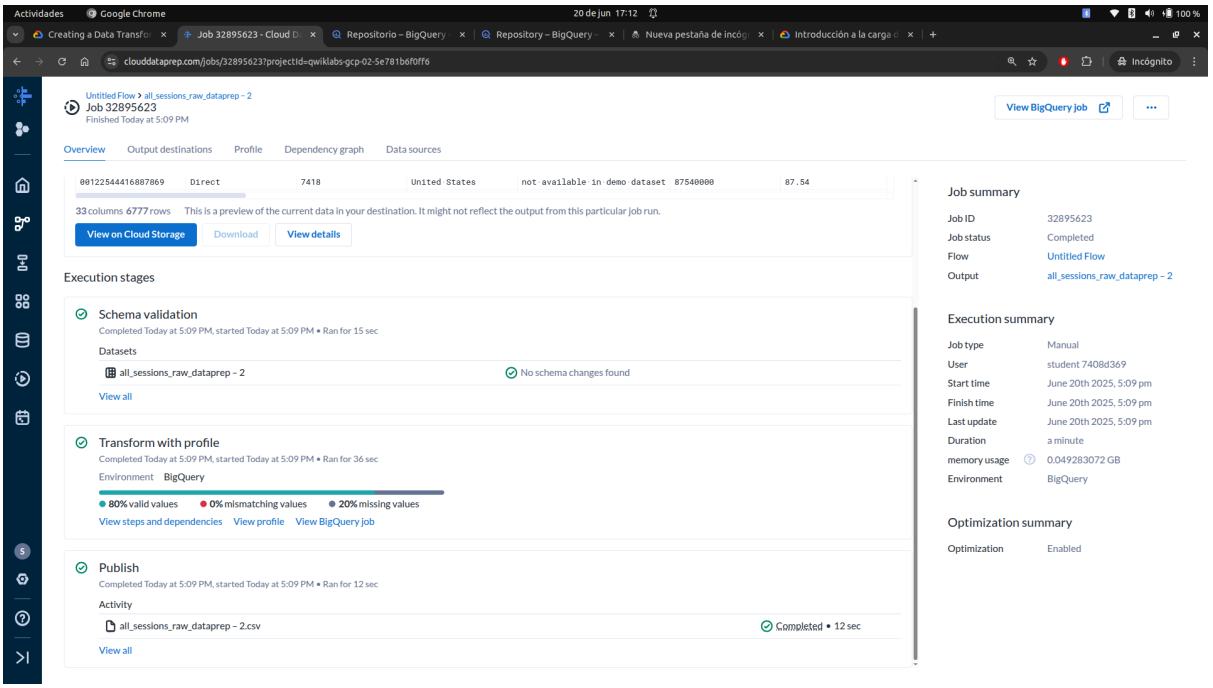
7. ¿Qué pasos se pueden seguir para limpiar y transformar datos en Data Prep de GCP?
Para limpiar y transformar datos se crea un *flujo* al cual uno le asocia uno o mas datasets. En el flujo uno define *recetas* a aplicar en el o los datasets.

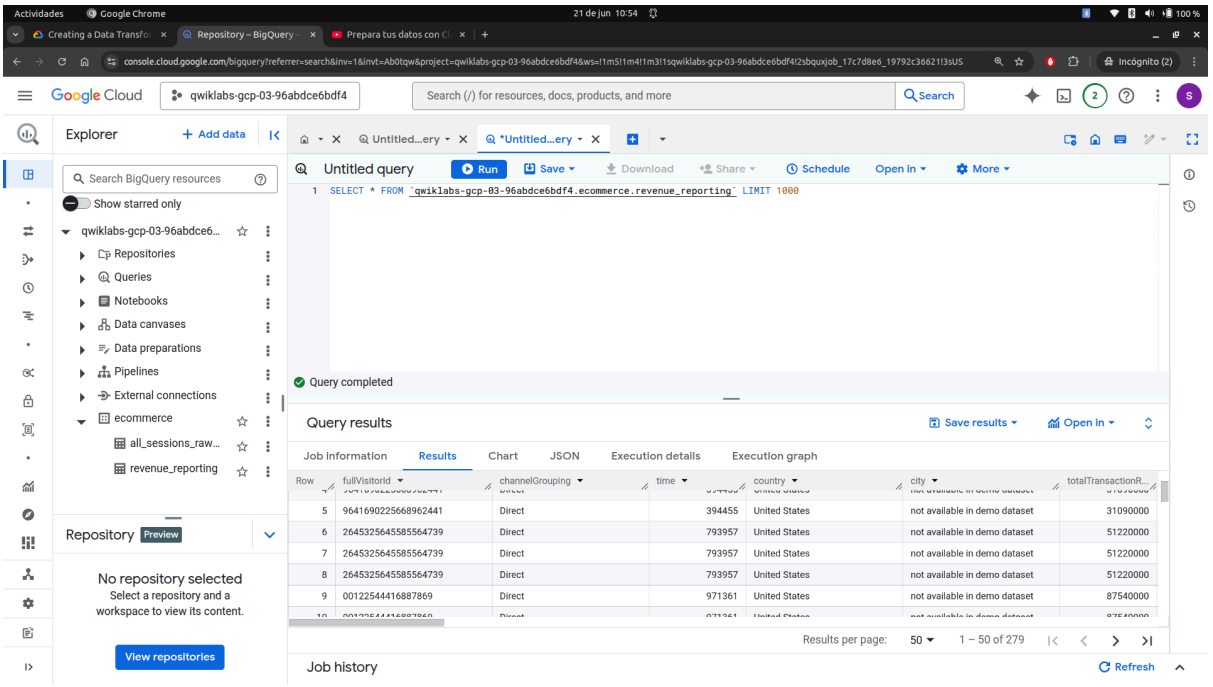
8. ¿Cómo se pueden automatizar tareas de preparación de datos en Data Prep de GCP?
Sí, es posible automatizar y programar trabajos de Dataprep utilizando Dataflow. Dataflow puede integrarse con Dataprep para ejecutar tareas de preparación de datos a gran escala y de forma automatizada. Esto permite transformar y limpiar datos utilizando la interfaz visual de Dataprep y luego ejecutar esas transformaciones a través de la infraestructura de Dataflow.

9. ¿Qué tipos de visualizaciones se pueden crear en Data Prep de GCP?
Con dataprep se pueden realizar gráficos de barra, de línea, de torta, dispersión, histogramas, gráficos de área, grafícos heatmap, gráficos de caja y otros gráficos especializados.

10.¿Cómo se puede garantizar la calidad deos datos en Data Prep de GCP?
Al abrir la tabla del dataset, Cloud Dataprep perfilará automáticamente los contenidos de tu conjunto de datos y generará histogramas de columnas, además de indicadores de calidad de los datos. Esta información de perfil puede usarse para guiar el proceso de preparación de los datos.


Actividades
Google Chrome
20 de jun 10:47
Creating a Data Transformation Pipeline with Cloud Dataprep
cloudskillsboost.google/focuses/4415?catalog_rank=%7B"rank"%3A1%2C"num_filters"%3A0%2C"has_search"%3Atrue%7D&parent=catalog&search_id=32278924
Incógnito
Creating a Data Transformation Pipeline with Cloud Dataprep
Quick tip: Review the prerequisites before you run the lab
Start Lab
01:15:00
Creating a Data Transformation Pipeline with Cloud Dataprep
Lab
1 hour 15 minutes
5 Credits
Intermediate
This lab may incorporate AI tools to support your learning.
This lab was developed with our partner, Alteryx Designer Cloud (Trifacta). Your pe
information may be shared with Alteryx Designer Cloud (Trifacta), the lab sponsor, if you
have opted-in to receive product updates, announcements, and offers in your Account
Profile.
My account
Hugo Alberto Aita
Member since 2025
62 Credits
Buy credits or subscription
Dashboard
Activities
Settings
Sign Out
Privacy
Terms
Task 4. Exploring ecommerce data fields with the UI
Task 5. Cleaning the data
Task 6. Enriching the data


Actividades
Google Chrome
20 de jun 17:12
clouddataprep.com/jobs/32895623?projectId=qwiklabs-gcp-02-5e781b6f0ff6
Incógnito
Untitled Flow › all_sessions_raw_dataprep – 2
Job 32895623
Finished Today at 5:09 PM
View BigQuery job
Overview
Output destinations
Profile
Dependency graph
Data sources
33 columns 6777 rows
This is a preview of the current data in your destination. It might not reflect the output from this particular job run.
View on Cloud Storage
Download
View details
Execution stages
Schema validation
Completed Today at 5:09 PM, started Today at 5:09 PM • Ran for 15 sec
Datasets
all_sessions_raw_dataprep – 2
No schema changes found
View all
Transform with profile
Completed Today at 5:09 PM, started Today at 5:09 PM • Ran for 36 sec
Environment BigQuery
80% valid values
0% mismatching values
20% missing values
View steps and dependencies
View profile
View BigQuery job
Publish
Completed Today at 5:09 PM, started Today at 5:09 PM • Ran for 12 sec
Activity
all_sessions_raw_dataprep – 2.csv
Completed • 12 sec
View all
Job summary
Job ID 32895623
Job status Completed
Flow Untitled Flow
Output all_sessions_raw_dataprep – 2
Execution summary
Job type Manual
User student 7408d369
Start time June 20th 2025, 5:09 pm
Finish time June 20th 2025, 5:09 pm
Last update June 20th 2025, 5:09 pm
Duration a minute
memory usage 0.049283072 GB
Environment BigQuery
Optimization summary
Optimization Enabled


Actividades
Google Chrome
21 de jun 10:54
Repository – BigQuery
Google Cloud
qwiklabs-gcp-03-96abdce6bdf4
Search (/) for resources, docs, products, and more
Search
Explorer
Add data
Search BigQuery resources
Show starred only
qwiklabs-gcp-03-96abdce6...
Repositories
Queries
Notebooks
Data canvases
Data preparations
Pipelines
External connections
ecommerce
all_sessions_raw...
revenue_reporting
Repository
Preview
No repository selected
Select a repository and a workspace to view its content.
View repositories
Untitled query
Run
Save
Download
Share
Schedule
Open in
More
1 SELECT * FROM `qwiklabs-gcp-03-96abdce6bdf4.ecommerce.revenue_reporting` LIMIT 1000
Query completed
Query results
Save results
Open in
Job information
Results
Chart
JSON
Execution details
Execution graph
Row
fullVisitorId
channelGrouping
time
country
city
totalTransactionR...
5 9641690225668962441 Direct 394455 United States not available in demo dataset 31090000
6 2645325645585564739 Direct 793957 United States not available in demo dataset 51220000
7 2645325645585564739 Direct 793957 United States not available in demo dataset 51220000
8 2645325645585564739 Direct 793957 United States not available in demo dataset 51220000
9 00122544416887869 Direct 971361 United States not available in demo dataset 87540000
Results per page: 50
1 – 50 of 279
Job history
Refresh

Arquitectura

A continuación se muestra la arquitectura propuesta: