



Ingesta, E.T.L y análisis de datos empresariales con Google Cloud + BigQuery

Objetivos	2
Arquitectura y pipeline que implementaremos	3
Requisitos.....	3
PRACTICA	4
Cree un dataset de BigQuery para guardar la tabla de resultados del pipeline.....	4
Ejecute esta query.....	5
Abra Cloud Dataprep ()	6
Importe y añada el dataset practica2	7
Explore los campos de datos (datos_en_bruto_todas_sesiones).....	9
Responda:.....	10
ETL.....	12
Conversiones.....	12
Borrado	13
Duplicados.....	15
Filtrar	16
Enriquecer los datos con otras fuentes datos	17
Tabla de mapeo.....	19
Mapeo	19
Creación de nueva col. calculada.....	20
Convertir datos de INT a String.....	22
Verifique la lista completa de tareas ETL.....	23
Verifique si los trabajos de Cloud Dataprep generan los datos en BigQuery.....	29
Jobs diferidos en el pipeline (ahorrando de costes)	29
PROGRAMAR JOBS (ahorrar costes)	29
CONCLUSION.....	31



Creación de un pipeline de Ingesta, enriquecimiento y transformación de datos con Cloud Dataprep + BigQuery

Esta práctica nos permitirá explorar grandes cantidades de datos (Big Data) apoyados en la computación distribuida, concretamente en los frameworks **Hadoop** y **Spark** (que en este caso son transparentes para nosotros) conforme al apartado 1,2 y 3 de la guía de aprendizaje.

Debemos ser conscientes de que el data set que vamos a utilizar consta de **21 millones de registros por 32 columnas es decir más de 500 millones** de registros que en una máquina en local (pc /mac) por más potente que esta sea sería imposible realizar cualquier operación, haremos diferentes análisis desde el punto de vista empresarial con el objeto de dar respuesta a varias incógnitas todas ellas relacionadas con el negocio de la empresa; para ello nos apoyaremos en un producto muy famoso denominado:

[Cloud Dataprep](#) es un servicio de datos inteligente que permite explorar, limpiar y preparar visualmente los datos estructurados y no estructurados para su análisis (ETL). En esta práctica evaluada, usaremos Cloud Dataprep para compilar una pipeline de transformación de datos que se ejecuta en un intervalo programado definido por el alumno y luego transfiere los resultados a BigQuery.

Utilizará un dataset de comercio electrónico que tiene 21 millones de registros x 32 columnas de sesión de Google Analytics para [Google Merchandise Store](#) cargados en BigQuery. Tienes una copia de ese dataset para esta práctica, y explorarás los campos y las filas para responder las preguntas del apartado [RESPONDA](#)

Objetivos

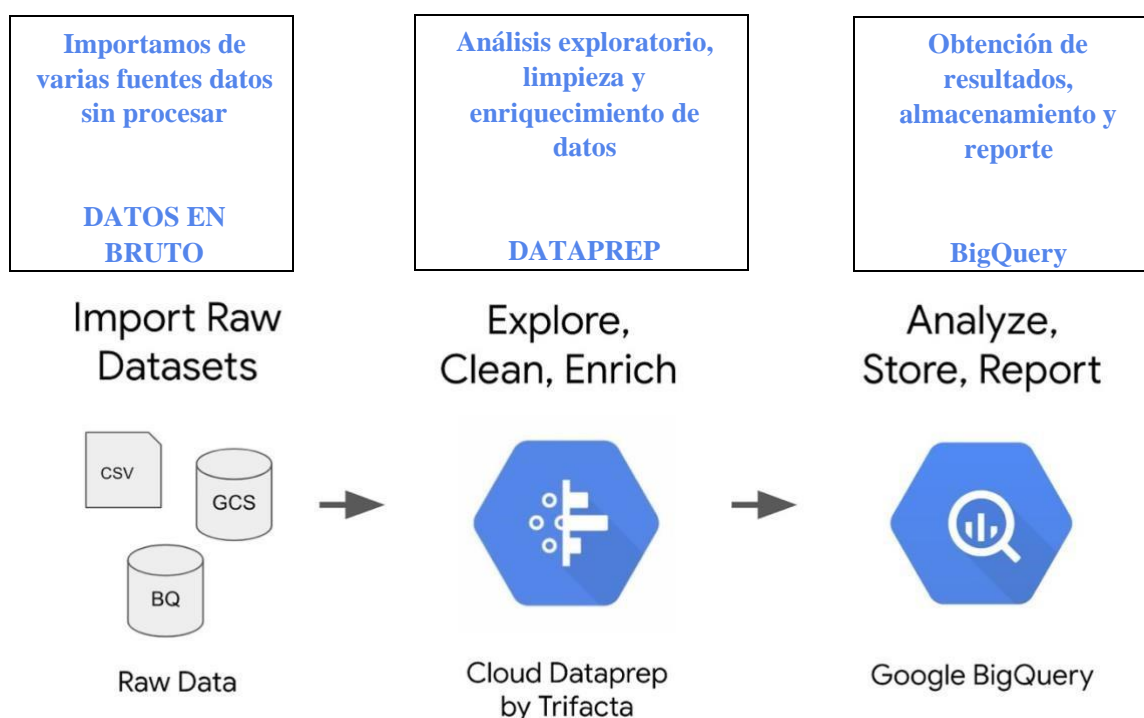
En esta práctica, aprenderá a realizar las siguientes tareas:

1. Conectar conjuntos de grandes cantidades de datos empresariales de BigQuery a Cloud Dataprep
2. Explorar la calidad del dataset con Cloud Dataprep
3. Crear un pipeline de transformación de datos con Cloud Dataprep
4. Programar resultados de los trabajos de transformación en BigQuery



5. Entender que detrás de lo anterior están trabajando tecnologías Big data y los rangos de computación distribuida que corresponde al apartado 1,2 y 3 de la guía de aprendizaje
6. Practicar todos y cada uno de los pasos dados en esta práctica, utilizando Hadoop y Spark; **los ejercicios de las practicas son similares a los que deberá entregar en la Actividad Individual 1 y 2.**

Arquitectura y pipeline que implementaremos



Requisitos

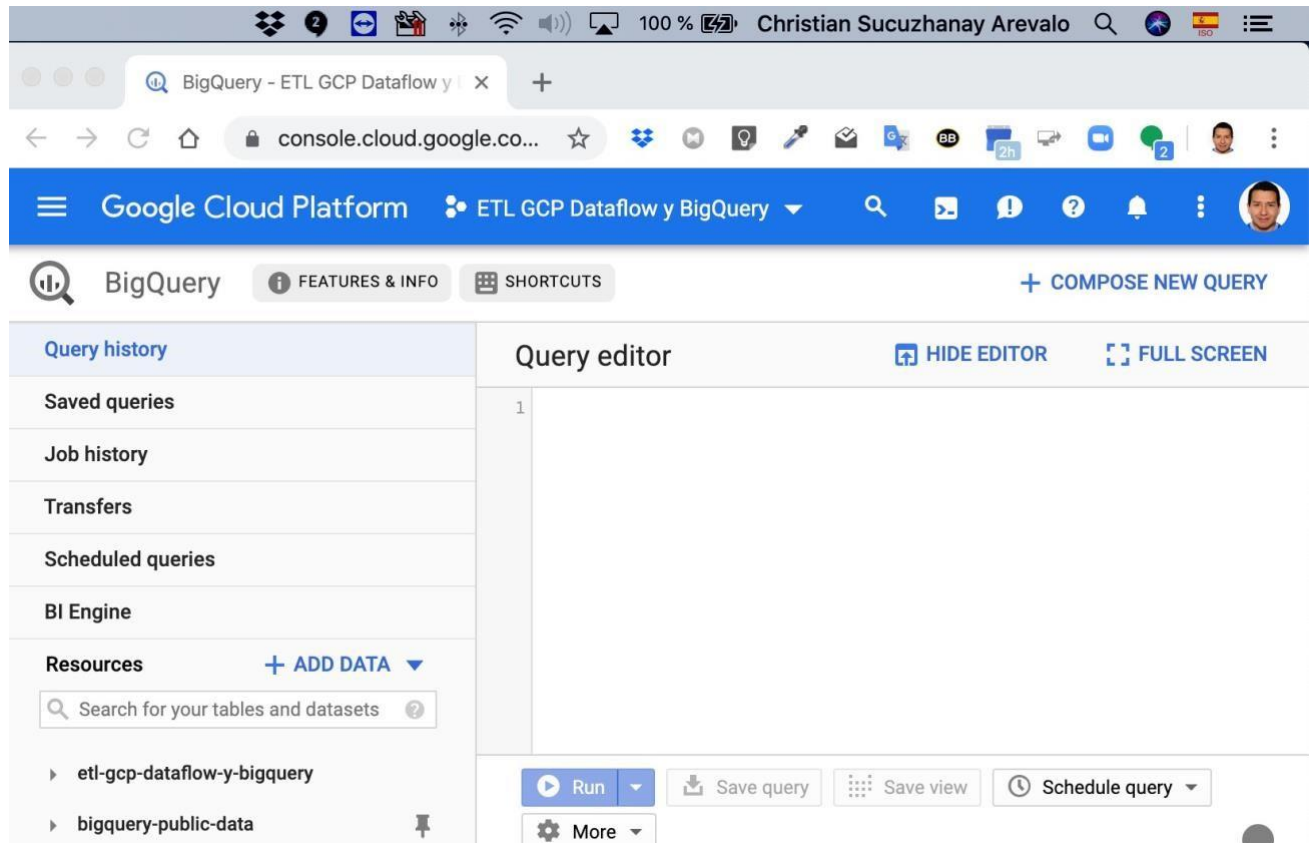
1. Google Chrome
2. Internet
3. Proyecto GCP



PRACTICA

Cree un nuevo proyecto = [ETL GCP Dataflow y BigQuery](#)

Abra BigQuery Console



Cree un dataset de BigQuery para guardar la tabla de resultados del pipeline.

1. En ID de dataset = [practica2](#), deje los otros valores predeterminados



The screenshot shows the Google Cloud Platform console with the 'Create dataset' form for 'practica2'. The form includes fields for 'Dataset ID' (practica2), 'Data location' (Default), and 'Default table expiration' (60 days). The 'Encryption' section shows 'Google-managed key' selected.

Create dataset

Dataset ID
practica2

Data location (Optional)
Default

Default table expiration
☒ 60 days (maximum for sandbox)
☐ Number of days after table creation:
60

Encryption
Data is encrypted automatically. Select an encryption key management solution.
☒ Google-managed key
No configuration required
☐ Customer-managed key
Manage via Google Cloud Key Management Service

Ejecute esta query.

The screenshot shows the Google Cloud Platform console with the 'Unsaved query' editor. The query is a SQL statement that creates a table 'practica2.datos_en_bruto_todas_las_sesiones' and inserts data from 'next-marketing-analytics.ecommerce.all_sessions_raw' for the date '20170801'.

Unsaved query Edited

```
1 CREATE OR REPLACE TABLE practica2.datos_en_bruto_todas_las_sesiones
2 OPTIONS(
3   description="Ingesta para Cloud Dataprep") AS
4 SELECT * FROM `next-marketing-analytics.ecommerce.all_sessions_raw`
5 WHERE date = '20170801'
```

Valid.

Run Save query Save view Schedule query More

This query will process 5.6 GB when run.

etl-gcp-dataflow-y-bigquery:practica2



Esta query copia un subconjunto del dataset de comercio electrónico sin procesar (con el valor de un día de datos de sesión o alrededor de 56,000 registros) en una tabla nueva llamada [datos_en_bruto_todas_las_sesiones](#), que se agregara al dataset [practica2](#).

The screenshot shows the Google Cloud BigQuery interface. On the left, the 'Resources' panel lists the project 'etl-gcp-dataflow-y-bigquery', dataset 'practica2', and table 'datos_en_bruto_todas_las_sesiones'. The main panel shows a query execution status 'Valid.' with a green checkmark. Below this, there are buttons for 'Run', 'Save query', 'Save view', and 'Schedule query'. A message indicates 'This query will process 5.6 GB when run.' with a green checkmark. The table 'datos_en_bruto_todas_las_sesiones' is highlighted. Below the table, there are tabs for 'Schema', 'Details', and 'Preview'. The 'Details' tab is active, showing a 'Description' field with the text 'Ingesta para Cloud Datastream' and a 'Labels' field with the text 'None'.

Abra Cloud Dataprep ()

1. Aceptar en todo y deje la ubicación predeterminada para el depósito de almacenamiento.

The screenshot shows the Cloud Dataprep by TRIFACTA interface. The top navigation bar includes the Google Cloud logo, the text 'Cloud Dataprep by TRIFACTA', and a blue button labeled 'Import Data'. Below this, there is a 'Create Flow' button. The main area is currently empty, showing a dashed line.

2. Conecte los datos de BigQuery a Cloud Dataprep Cree un flujo en la esquina superior derecha.
 - a. Nombre = [Ecommerce Analytics Pipeline](#).
 - b. Descripción = [Tabla reporte ingresos](#)



Cloud Dataprep by TRIFACTA

Create Flow

Flow Name

Ecommerce Analytics Pipeline

Flow Description

Tabla reporte ingresos

Cancel Create

Importe y añade el dataset [practica2](#)



Import data before wrangling in this Flow.

Import & Add Datasets



Import Data and Add to Flow

Upload

GCS

BigQuery

Choose a table

BigQuery / etl-gcp-dataflow-y-bigquery

NAME

practica2

Import Data and Add to Flow

Upload

GCS

BigQuery

Choose a table

BigQuery / etl-gcp-dataflow-y-bigquery / practica2

	NAME	SIZE	LAST UPDA
	datos_en_bru...	32 Col...	57k ... Today at...

1 New Dataset

datos_en_bruto_todas_las_s

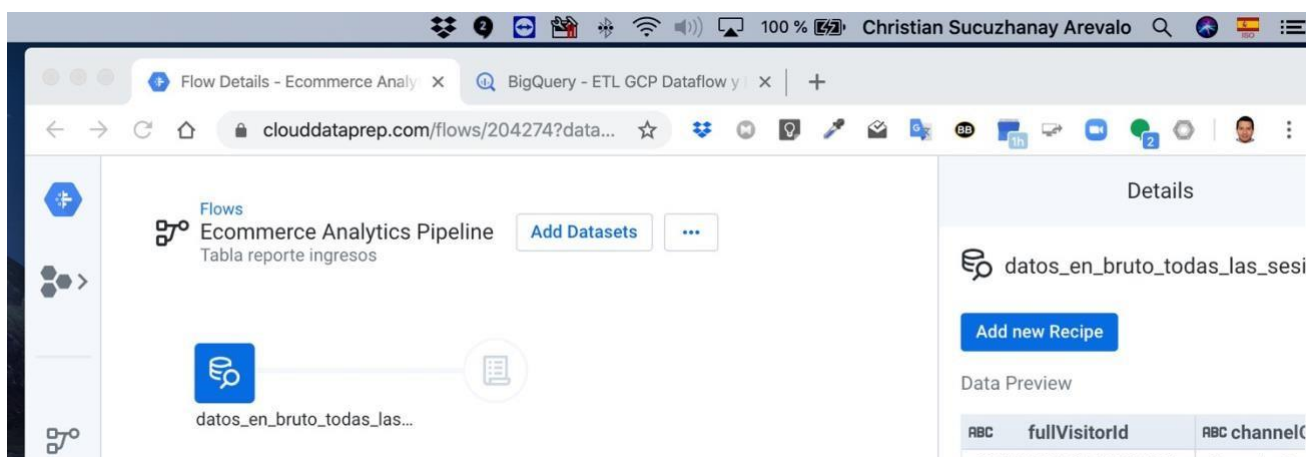
Add a Description

ABC	fullVisitorId	ABC chan
8074041050560984021		Organic
8074041050560984021		Organic
8685530477324183365		Display
3395445735354444853		Direct
3173566250804266498		Organic

Import & Add to Flow

Cancel

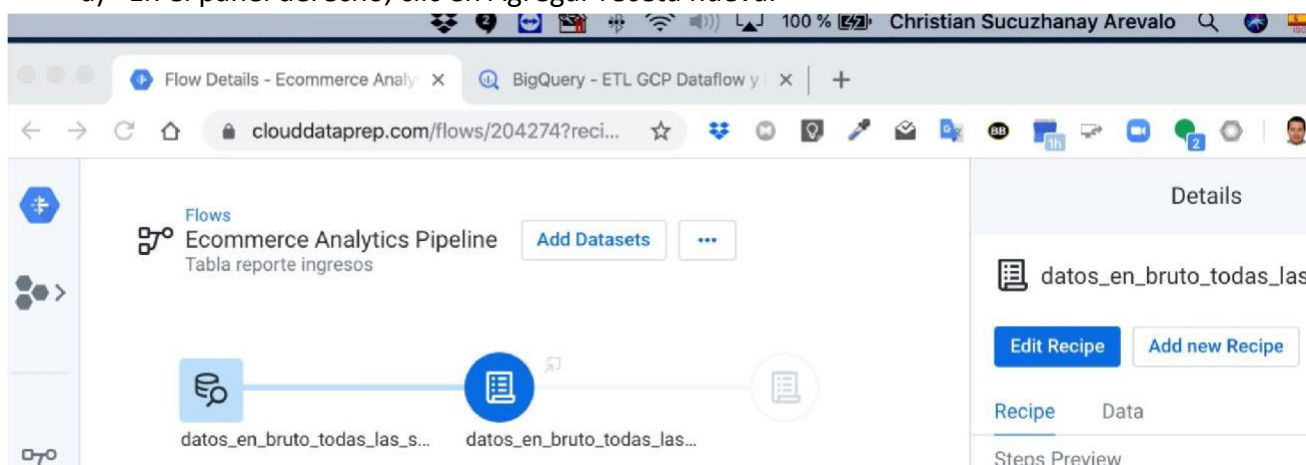
Click en Import & Add to FLOW



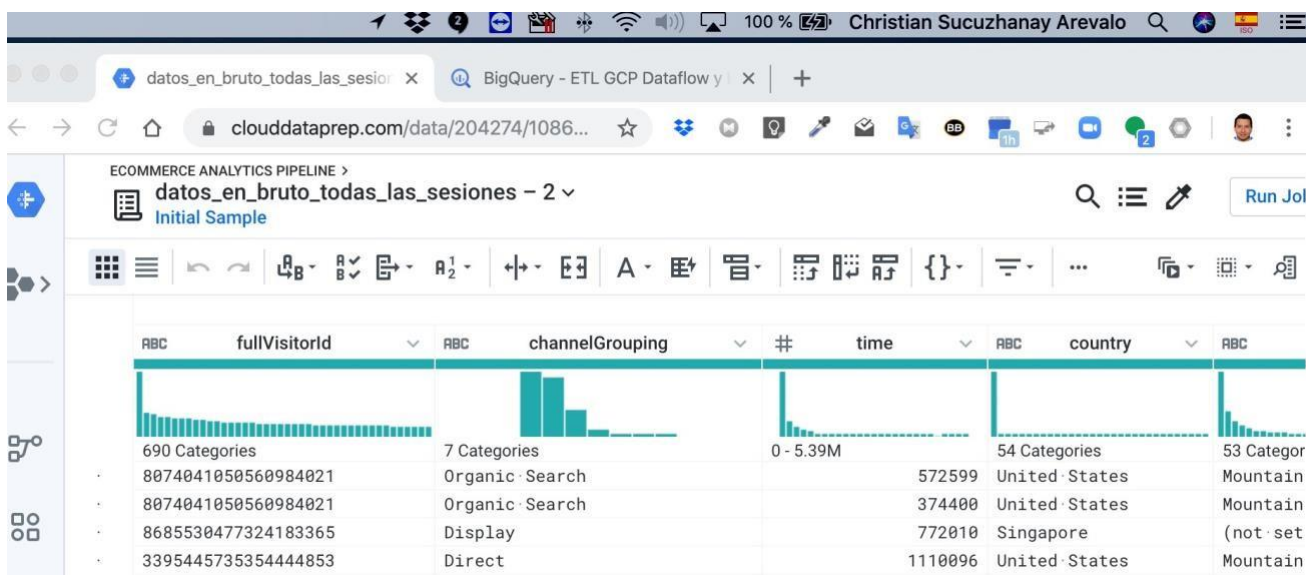
Explore los campos de datos ([datos_en_bruto_todas_las_sesiones](#))

Cargue y explore una muestra del dataset dentro de Cloud Dataprep.

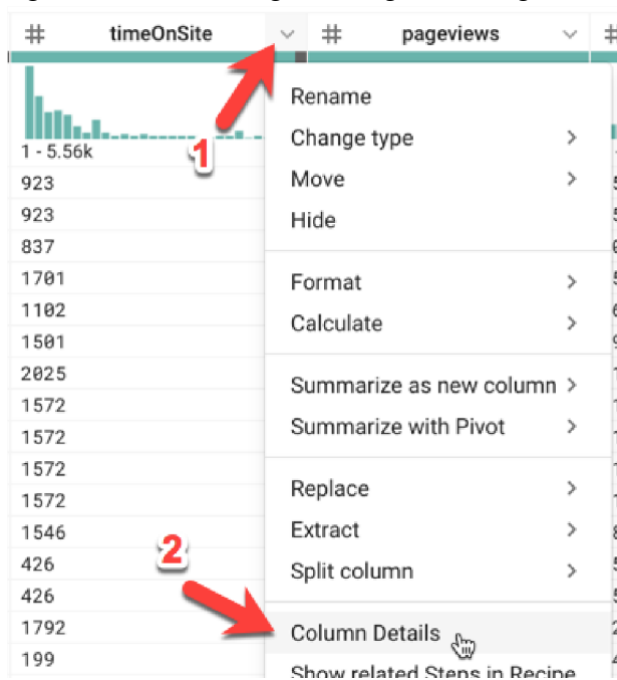
- a) En el panel derecho, clic en Agregar receta nueva.



- b) Clic en **Editar receta**, ahora Cloud Dataprep carga una muestra de tu dataset en la vista de Transformer para comenzar a explorar los datos.



Sugerencia vea los siguientes graficos siguiente:

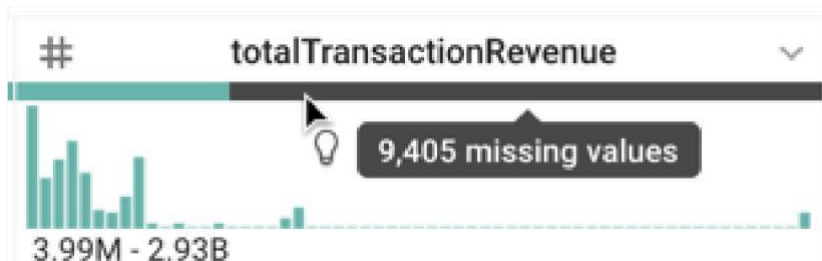


Responda:

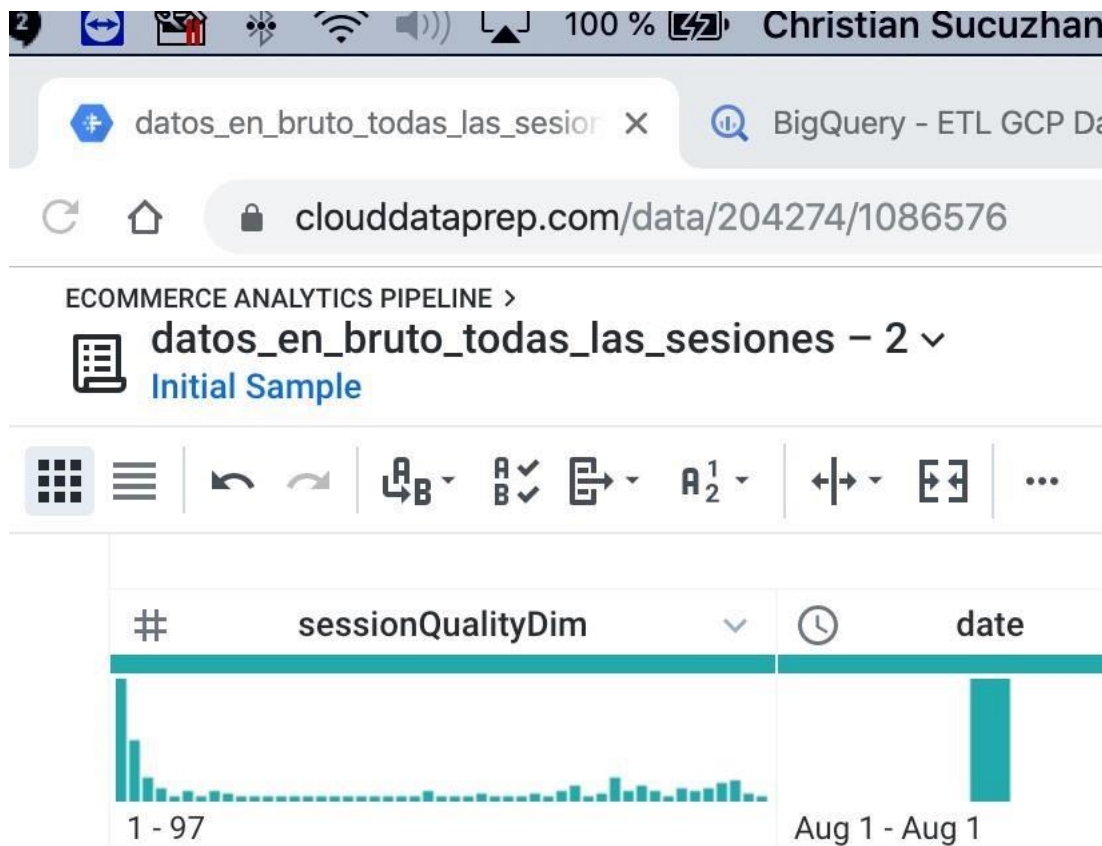
1. ¿Cuántas columnas hay en el dataset?
2. ¿Cuántas filas contiene la muestra?
3. ¿Cuál es el valor más común en la columna channelGrouping?
4. ¿Cuál es el valor máximo de timeOnSite en segundos?
5. ¿Y el valor máximo de pageviews?



- ¿Y el valor máximo de `sessionQualityDim`?
- ¿Cuáles son los tres primeros países desde donde se originaron sesiones?
- ¿Qué representa la barra gris que se encuentra debajo de `totalTransactionRevenue`?



- Si se observa el histograma de `sessionQualityDim`, ¿los valores de datos están distribuidos de manera uniforme?



- Cuál es el **período** para el conjunto de datos?
- Puede que vea una barra roja debajo de la columna `productSKU`. De ser así, ¿qué podría significar?

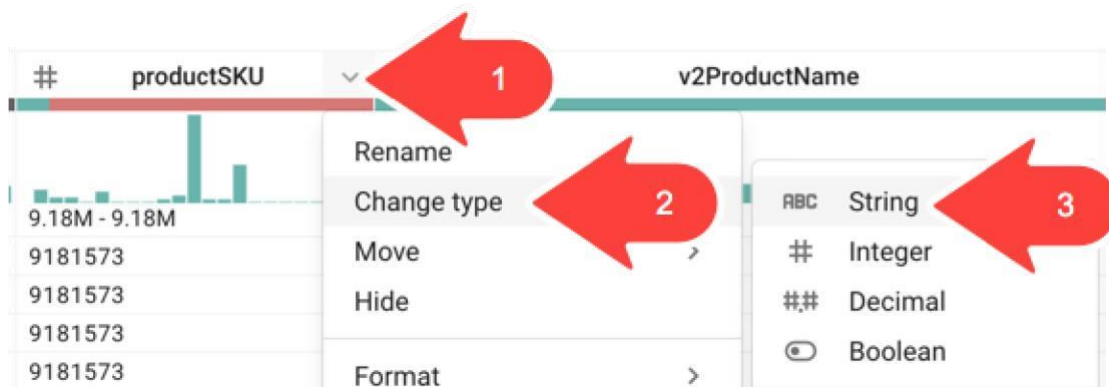


ETL

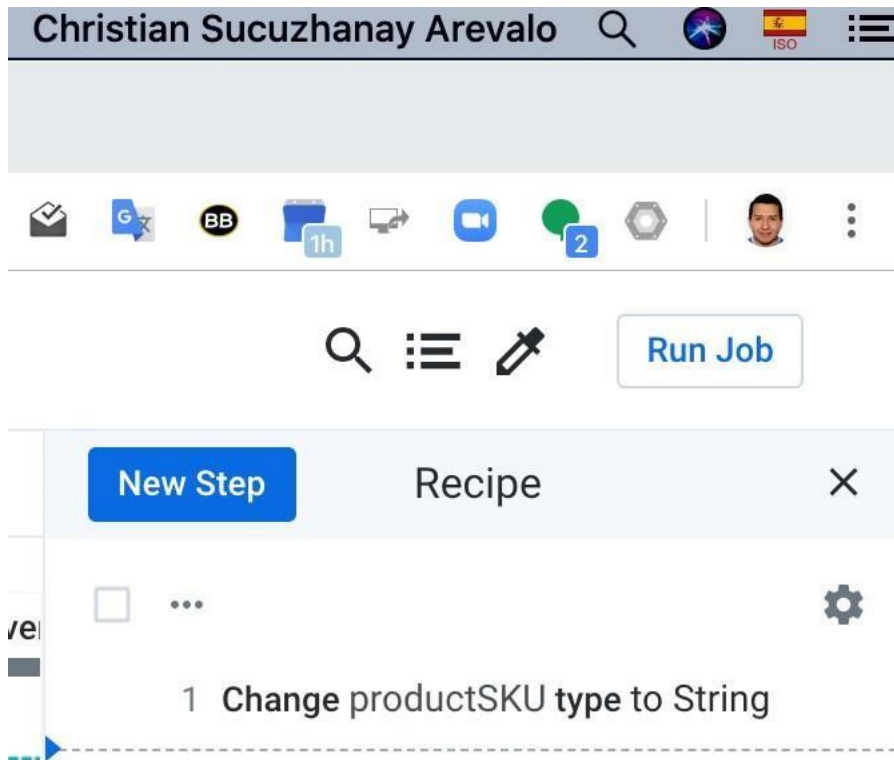
1. Limpieza de datos
2. Borrar columnas innecesarias
3. Quitar duplicados
4. Creará campos calculados
5. Filtrará las filas no deseadas para limpiar los datos.
6. Convertir datos de un tipo a otro

Conversiones

Convierta el tipo de datos de la columna **productSKU** a string. (Dado que Dataprep incorrectamente cree que los datos son de tipo numérico)



Haga clic en el ícono de Receta para verificar que se creó el primer paso en su pipeline de transformación de datos.




Borrado

Borre las columnas sin utilizar **itemQuantity** y **itemRevenue**, ya que solo contienen valores **NULL** y no son útiles **para el calculo en esta práctica.**

Abra el menú de la columna **itemQuantity** y, luego, haga clic en Delete.



ABC	itemQuantity	▼	ABC	itemRevenue	▼
No valid values.			Rename		
			Change type >		
			Move >		
			Hide		
			Format >		
			Calculate >		
			Summarize as new column >		
			Summarize with Pivot >		
			Replace >		
			Extract >		
			Split column >		
			Column Details		
			Show related Steps in Recipe		
			Lookup...		
			Delete 		
			Delete others		

Repita el proceso para borrar la columna [itemRevenue](#).

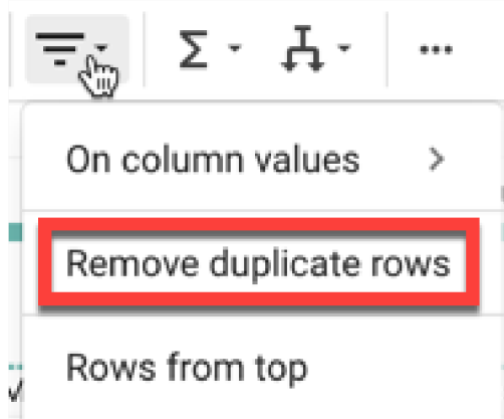


Duplicados

Anule la duplicación de filas

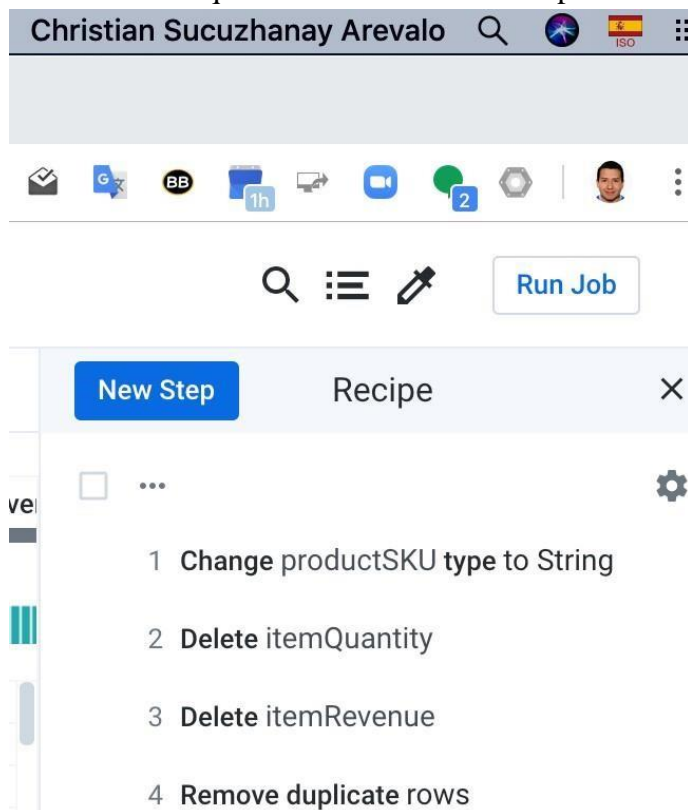
Dado que pueden existir inicio de sesión duplicados incluidos en el dataset de origen debemos eliminarlos.

Haga clic en el ícono de Filtrar filas en la barra de herramientas y, luego, haga clic en Quitar filas duplicadas.



En el panel derecho, haga clic en **Agregar**.

Revise la receta que creó hasta ahora. Debe poder ver lo siguiente:



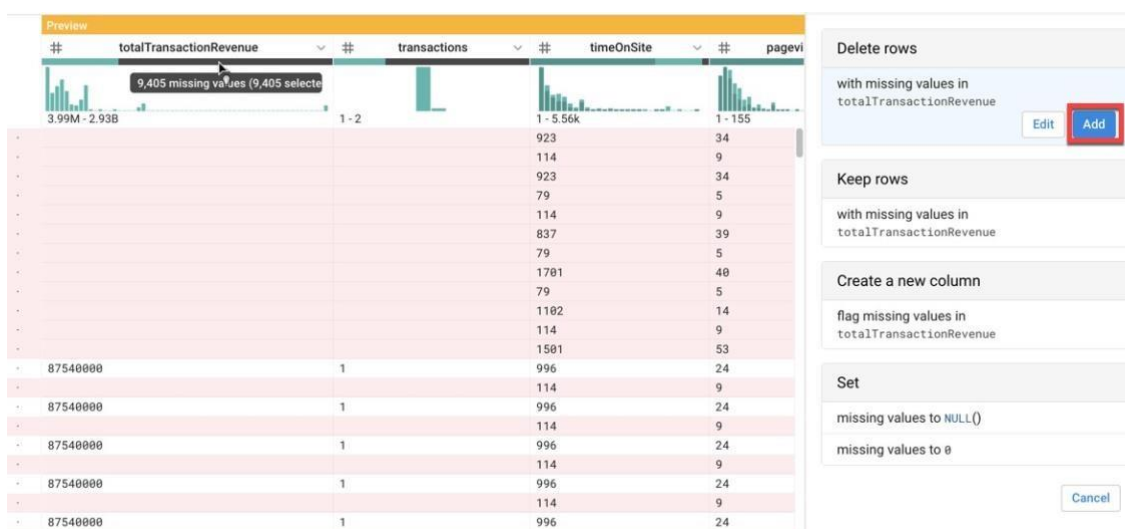


Filtrar

Filtre sesiones sin ingresos

Queremos obtener todas las sesiones de usuario que compraron **al menos un artículo** del sitio web. Filtre las sesiones de usuario con ingresos que contengan el valor NULL.

Debajo de la columna totalTransactionRevenue, haga clic en la barra gris Valores faltantes. Todas las filas con valores faltantes para totalTransactionRevenue ahora están resaltadas en rojo.

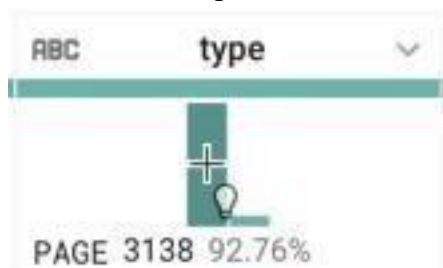


Este paso filtra su dataset para que solo incluya transacciones con ingresos (donde totalTransactionRevenue no es NULL).

Filtre sesiones que sean solo las PÁGINAS vistas

El dataset contiene sesiones de diferentes tipos, por ejemplo, PAGE (para páginas vistas) o EVENTS (para eventos activados como por ejemplo "categorías de producto vistos" o "agregados al carrito"). A fin de evitar que se cuenten **2 veces** las sesiones de páginas vistas, agregue un filtro para incluir solo hits relacionados con páginas vistas.

En el histograma que se encuentra debajo de la columna type, haga clic en la barra para PAGE. Todas las filas con el tipo PAGE ahora están resaltadas en verde.



En el panel Sugerencias, en Conservar filas, haga clic en Agregar.



id_unico_de_usuario_sesion	type
31150756085903-1501603468	PAGE
542428111966715-1501608078	EVENT

Enriquecer los datos con otras fuentes datos

Mire aquí la [documentación de esquemas](#) para **visitId** y lea la descripción para determinar si es único **en todas las sesiones** de usuario o solo es el usuario.

visitId: Es un identificador para esta sesión únicamente. Es parte del valor que se almacena en una cookie. Esto es único solo para el usuario.

Para un ID completamente único, debe utilizar una combinación de **fullVisitorId** + **visitId**.

Como podemos ver, **visitId** no es único para todos los usuarios. Necesitaremos crear un identificador único.

El dataset no tiene ninguna columna única para una sesión única de visitante. Cree un ID único para cada sesión. Para ello, concatene los campos **fullVisitorID** + **visitId** y :

Cree una columna nueva para un ID de sesión único

Haga clic en el ícono de Combinar columnas en la barra de herramientas.



En Columnas, seleccione: **fullVisitorId** y **visitId**.

En Separador, escriba un carácter: -

En Nuevo nombre de columna = **id_unico_de_usuario_sesion**



The screenshot shows the Google Cloud Data Studio interface. At the top, the user is logged in as Christian Sucuzhanay Arevalo. The main view displays a table named 's_las_sesiones' with 2 columns: 'ABC' and 'produ'. The 'ABC' column has 225 categories, and the 'produ' column has 97 categories. A 'Merge columns' dialog is open on the right, showing the columns 'fullVisitorId' and 'visitId' being merged into a new column named 'id_unico_de_usuario_sesion'. The dialog also shows a 'Separator' field and a 'New column name' field. The 'Run Job' button is visible in the top right corner of the dialog.

Haga clic en Agregar.

[id_unico_de_usuario_sesion](#) ahora es una combinación de **fullVisitorId** y **visitId**.

Los valores de la columna [eCommerceAction_type](#) son números enteros mapeados a acciones llevadas a cabo en esa sesión, es decir, para :



Tabla de mapeo

Valor para comparar	Acción
0	"Unknown"
1	"Click through of product lists"
2	"Product detail views"
3	"Add product(s) to cart"
4	"Remove product(s) from cart"
5	"Check out"
6	"Completed purchase"
7	"Refund of purchase"
8	"Checkout options"

Este mapeo no es evidente para la interpretación de nuestro CEO o persona a la que le entreguemos el REPORTE, así que **creamos un campo calculado** que incorpora el valor del **nombre de la acción** y sea fácil de entender, para ello hacemos lo siguiente :

Mapeo

Cree una declaración de caso para cada tipo de acción de comercio electrónico



En Columna para evaluar, especifique [eCommerceAction_type](#).

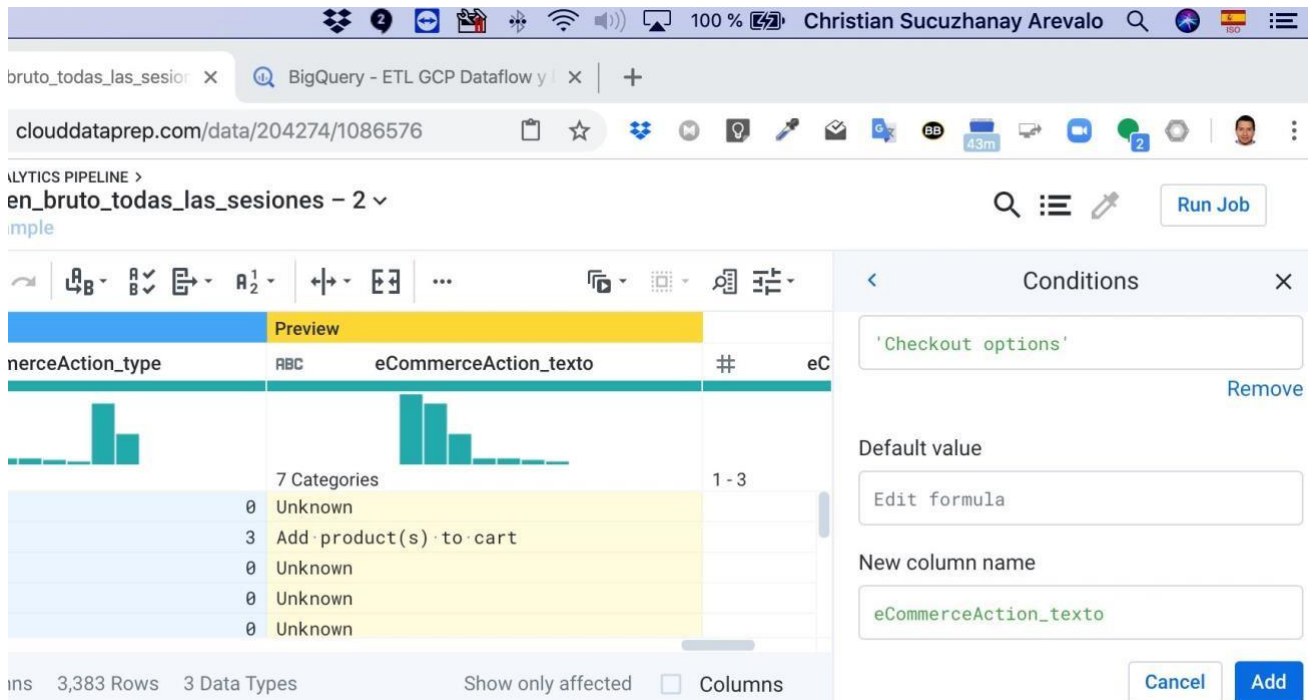
Junto a Casos (1), haga clic en Agregar **8** veces para un total de 9 casos.



Para cada Caso, especifique los siguientes valores de mapeo (incluidos los caracteres de comillas simples) [Conforme a la tabla anterior dada.](#)

Creación de nueva col. calculada

En nuevo nombre de la columna, escriba `eCommerceAction_texto`. Los demás valores dejelos en predeterminados.

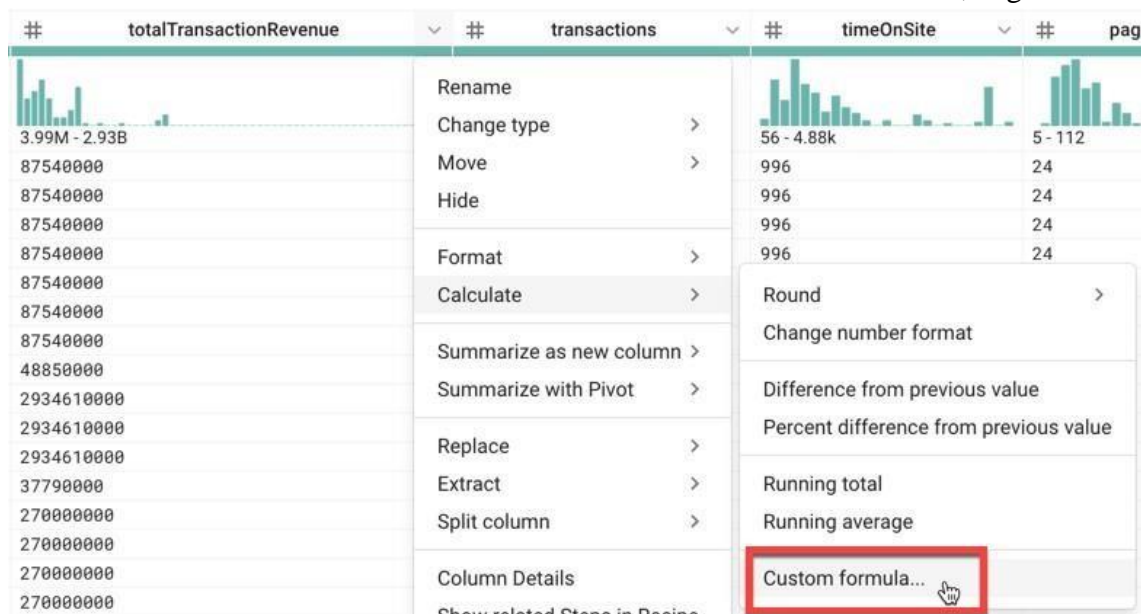


Haga clic en Agregar.

Ajuste los valores de la columna totalTransactionRevenue

Como se menciona en el [esquema](#), la columna totalTransactionRevenue contiene valores pasados a Google Analytics multiplicados por 10^6 (p. ej., 4.40 se daría como 4,400,000). Por lo tanto necesitamos dividir el contenido de esa columna por 10^6 para obtener los valores reales.

Abra el menú a la derecha de la columna **totalTransactionRevenue**. Para ello, haga :





En Fórmula, escriba: `DIVIDE(totalTransactionRevenue,1000000)` y en Nuevo nombre de columna, escriba: `totalTransactionRevenueReal`.

The screenshot shows the Google Cloud Dataprep interface. The main view displays a dataset named 'datos_en_bruto_todas_las_sesiones - 2'. A 'New formula' dialog is open on the right. The 'Formula type' is set to 'Single row formula'. The 'Formula' field contains the expression `DIVIDE(totalTransactionRevenue, 1000000)`. The 'New column name' field contains `totalTransactionRevenueReal`. The 'Run Job' button is visible in the top right corner of the interface.

Haga clic en Agregar.

Convertir datos de INT a String

Como puede ver Dataprep nos crea la nueva columna `totalTransactionRevenueReal` de tipo INT y queremos convertirla a decimal, para ello haga :

The screenshot shows the Google Cloud Dataprep interface. The main view displays a dataset named 'datos_en_bruto_todas_las_sesiones - 2'. A 'New Step' dialog is open on the right. The 'New Step' button is visible in the top right corner of the interface.



#	totalTransactionRevenueReal
ABC	String
#	Integer
0	##
0	Boolean
	48.85
	2934.61

Verifique la lista completa de tareas ETL

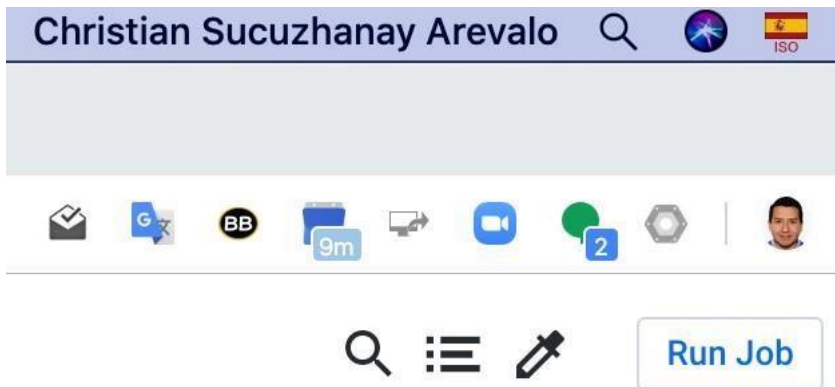
Christian Sucuzhanay Arevalo

Run Job

New Step Recipe

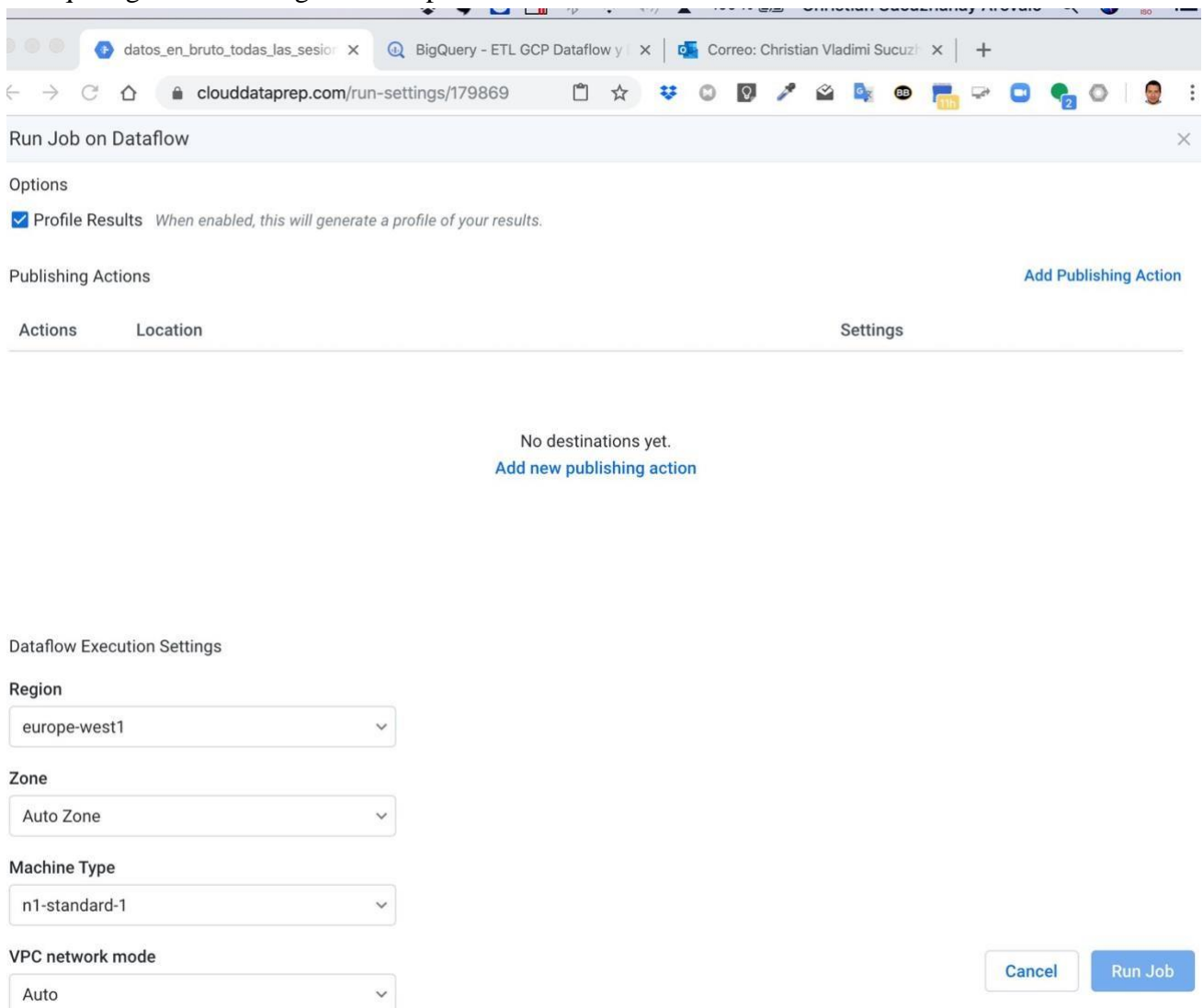
views

- 1 Change productSKU type to String
- 2 Delete itemQuantity
- 3 Delete itemRevenue
- 4 Remove duplicate rows
- 5 Delete rows where ISMISSING([totalTransactionRevenue])
- 6 Keep rows where type == 'PAGE'
- 7 Concatenate fullVisitorId, visitId separated by \
- 8 Create eCommerceAction_texto from 9 case conditions on eCommerceAction_type
- 9 Create totalTransactionRevenueReal from DIVIDE(totalTransactionRevenue, 1000000)
- 10 Change totalTransactionRevenueReal type to Decimal



Y ejecute, click en **RUN JOB**.

Cargar el resultado del trabajo en el dataset de BigQuery [practica2](#) que creó antes. **Asegúrese** de cargar el resultado en una tabla separada y asignarle el nombre [reporte_ingresos](#). **Siga** los pasos que figuran en las siguientes capturas:



Click en : [Add new publishing action](#)



Seleccionar BigQuery -> practica2

Publishing Action

GCS

BigQuery

Choose a table

BigQuery / etl-gcp-dataflow-y-bigquery

NAME

practica2

Click : Create a new table llamada = [reporteIngresosUem](#)

clouddataprep.com/run-settings/179869

Publishing Action

GCS

BigQuery

Choose a table

BigQuery / etl-gcp-dataflow-y-bigquery / practica2

	NAME	SIZE	LAST UPDA
	datos_en_bruto_t...	32 Colu...	57k ... Today at...

Create a new table [Parameterize](#)

reporteIngresosUem

Output Database

practica2

☐ Create new table every run
Create a new table with a timest
appended to the name (e.g.
reporteIngresosUem_20191021

☒ Append to this table every run
Or create it if it doesn't exist.

Cancel Add

Click en Add.



Run Job on Dataflow

Options

☒ Profile Results *When enabled, this will generate a profile of your results.*

Publishing Actions

[Add Publishing Action](#)

Actions	Location	Settings
Append-BigC	etl-gcp-dataflow-y-bigquery:practica2.reporteIngresosUem	Create table if it does not exist; Append

Dataflow Execution Settings

Region

us-central1

Zone

Auto Zone

Machine Type

n1-standard-1

VPC network mode

Auto

Cancel

Run Job

Click Run job y vera lo siguiente:

The screenshot shows the Google Cloud Dataflow console. The main view displays a flow named 'Ecommerce Analytics Pipeline' with a table report 'Tabla reporte ingresos'. Below the flow diagram, there are two nodes labeled 'datos_en_bruto_todas_las_s...'. A 'Run Job' button is visible. On the right, a 'Details' panel shows the job name 'datos_en_bruto_todas_las_sesione...', a 'Run Job' button, and a list of destinations. The job status is 'In progress' with job ID 'Job 3613876' and email 'sukuzhanay@gmail.com'.



Para verificar como va el trabajo, haga click en : **View Dataflow job**

The screenshot shows the Google Cloud Dataflow console. At the top, there is a title bar with a refresh icon and the text "datos_en_bruto_todas_las_sesiones - 2". Below this is a blue "Run Job" button and a menu icon (three dots). Underneath, there are two tabs: "Destinations" and "Jobs (1)". The "Jobs (1)" tab is active, showing a single job entry. The job is named "Job 3613876" and is in the "In progress" state. It was started by "sukuzhanay@gmail.com" and "Started Today at 8". To the right of the job entry is another menu icon (three dots). Below the job entry, there is a button labeled "View Dataflow job" with an external link icon (a square with a diagonal arrow).

Y vera los treabajos ejecutándose, siguiente fig.



[← Job details](#)[LOGS](#)

Job

Job summary

Job name	cloud-dataprep-e-commerce-analytics-pi-by-sukuzhanay
Job ID	2019-10-20_23_16_03-1405448149515
Region	us-central1
Job status	Running

[Stop job](#)

SDK version	Apache Beam SDK for Java 2.11.0
Job type	Batch
Start time	October 21, 2019 at 8:16:04 AM UTC+2
Elapsed time	4 min 31 sec
Encryption type	Google-managed key

Autoscaling

Workers	1
Current state	Worker pool started.

Current workers: 1, Target workers: 1

Resource metrics

Current vCPUs	1
Total vCPU time	0.061 vCPU hr
Current memory	3.75 GB
Total memory time	0.229 GB hr
Current PD	250 GB
Total PD time	15.278 GB hr
Current SSD PD	0 B
Total SSD PD time	0 GB hr

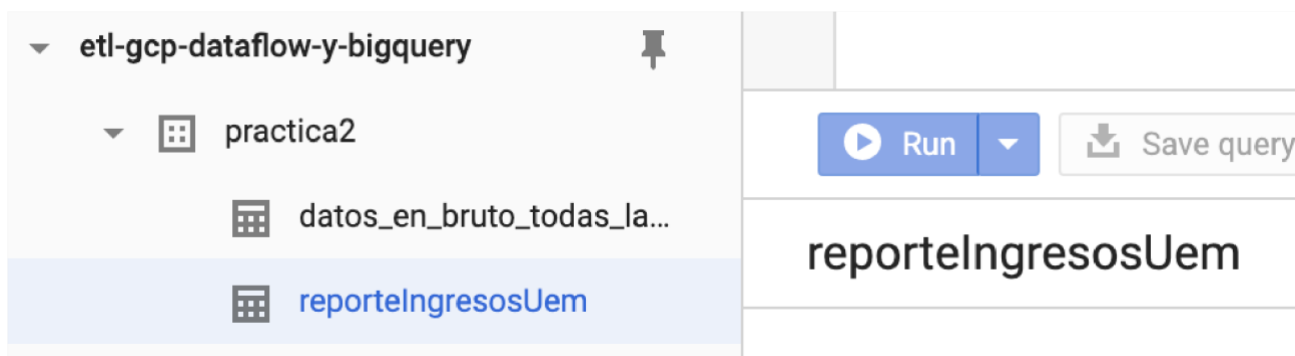
Pipeline options

workerMachineType	n1-standard-1
outputLocations	{'location1': 'etl-gcp-dataflow-y-bigquery', 'location2': 'gs://dataprep-102a-3812-4107-aaa2-6586ece5e910/gstudent@qwiklabs.net/jobrun/datos_en_s_sesiones_2_3613876/profiler/profistograms.json/file', 'location3': 'gs://d...

Ejecute y cargue el resultado del trabajo en el dataset ([practica2](#)) de BigQuery que creó antes. Cargue el resultado en una nueva tabla separada llamada = [reporteIngresosUem](#)



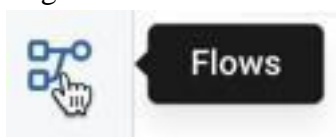
Verifique si los trabajos de Cloud Dataprep generan los datos en BigQuery



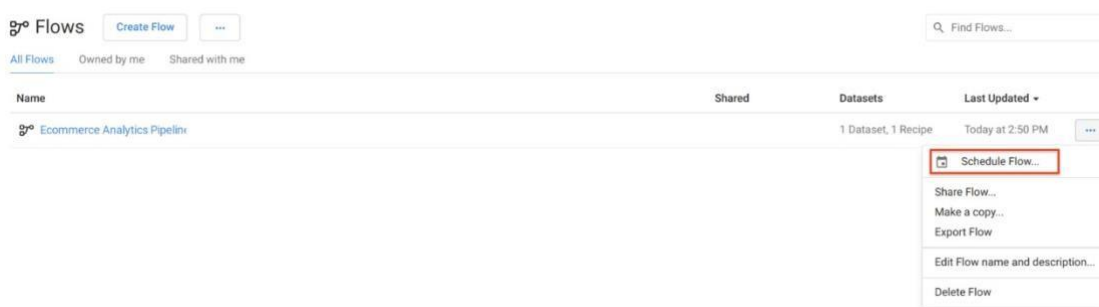
Jobs diferidos en el pipeline (ahorrando de costes)

PROGRAMAR JOBS (ahorrar costes)

Ahora, programará una ejecución recurrente del trabajo (automatizado sin control del data engineer). Haga clic en el ícono de Flujos a la izquierda de la pantalla.



A la derecha de su flujo de Ecommerce Analytics Pipeline, haga clic en el ícono de Más () y, luego, en Programar flujo.





Add Schedule

×

Scheduling Options

Timezone

Europe/Madrid

Frequency

Weekly

on

Saturday

at

03:00

AM

Add

Cancel

Save

En el diálogo Agregar programación, ponga:

- Frecuencia, seleccione Semanalmente.
- Día de la semana, Sabado (es mas barato).
- Para la hora, ingrese 3:00 y seleccione a.m.(madrugada = barato) **Clic en Guardar.**

Ahora tenemos un job programado para ejecutarse todos los sabados a las 3 a.m

Haga clic en el ícono de JOBS a la izquierda de la pantalla.



Verá la lista de trabajos; espere hasta que su trabajo esté marcado como Completado.

Job	User	Output	Status	Started
361387	sukuzhanay@gmail.com (y	datos_en_bruto_todas_las_sesiones Ecommerce Analytics Pipeline	Completed	Today at 8:15 AM Ran for 9 minutes



CONCLUSION

Exploró un dataset de comercio electrónico que contiene 21.500.000 filas y 32 columnas, creó un pipeline de ETL (transformación de datos) para que funcione de forma recurrente (Con el fin de mantener actualizado el trabajo, la información y encima ahorrando euros) usando Cloud Dataprep.+ BigQuery.

<gs://dataprep-staging-1ddb742b-13f3-4a76-b318-20480884e93c/sukuzhanay@gmail.com/> Change