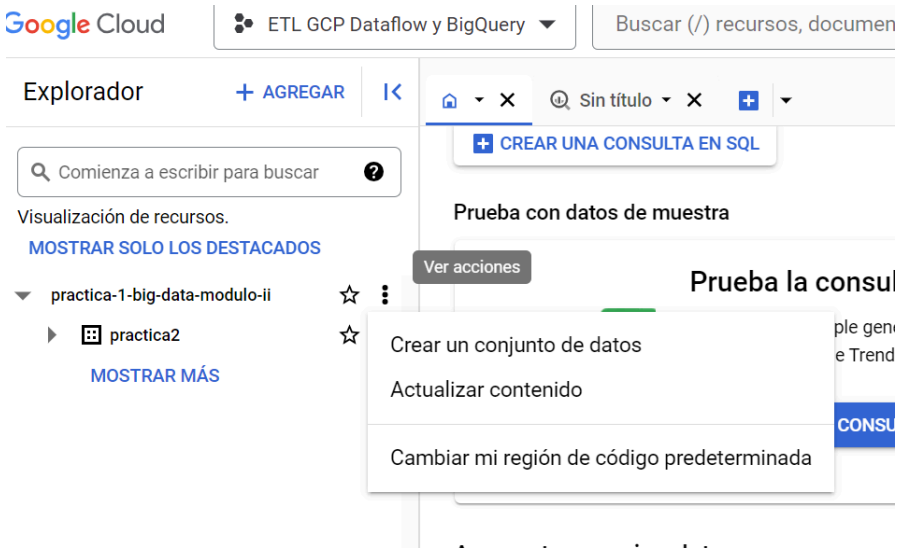
Universidad Europea

Actividad 1. Práctica Hadoop

GUILLERMO GARCÍA GONZÁLEZ

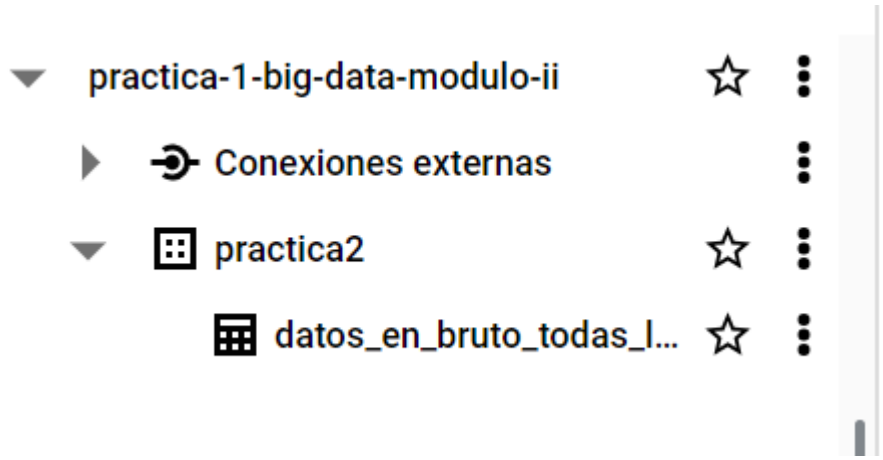
MIGUEL ÁNGEL ALCÓN GALÁN

Creamos un nuevo proyecto en Google Cloud para trabajar en la actividad y dentro de este proyecto generamos un conjunto de datos llamado “practica2”



A través de una query dentro del conjunto de nuestro proyecto creamos una tabla de los datos brutos para posteriormente enlazar con Cloud Dataprep

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



Rename

Flow Name

Ecommerce Analytics Pipeline

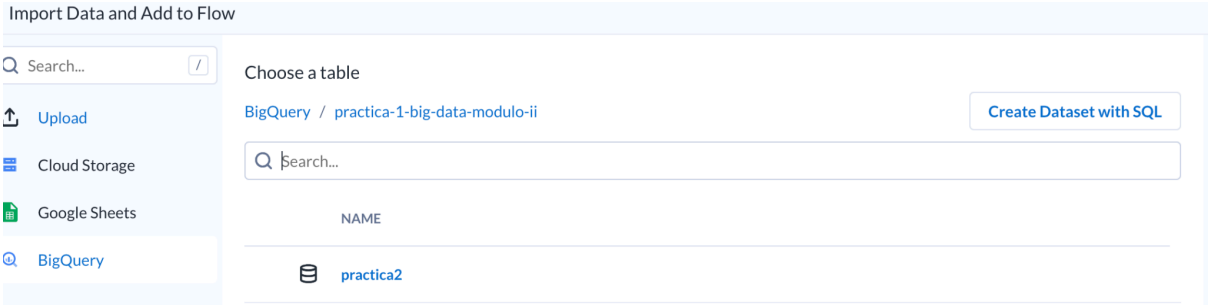
Flow Description

Tabla reporte de ingresos

Cancel OK

Ecommerce Analytics Pipeline

Tabla reporte de ingresos



Choose a table

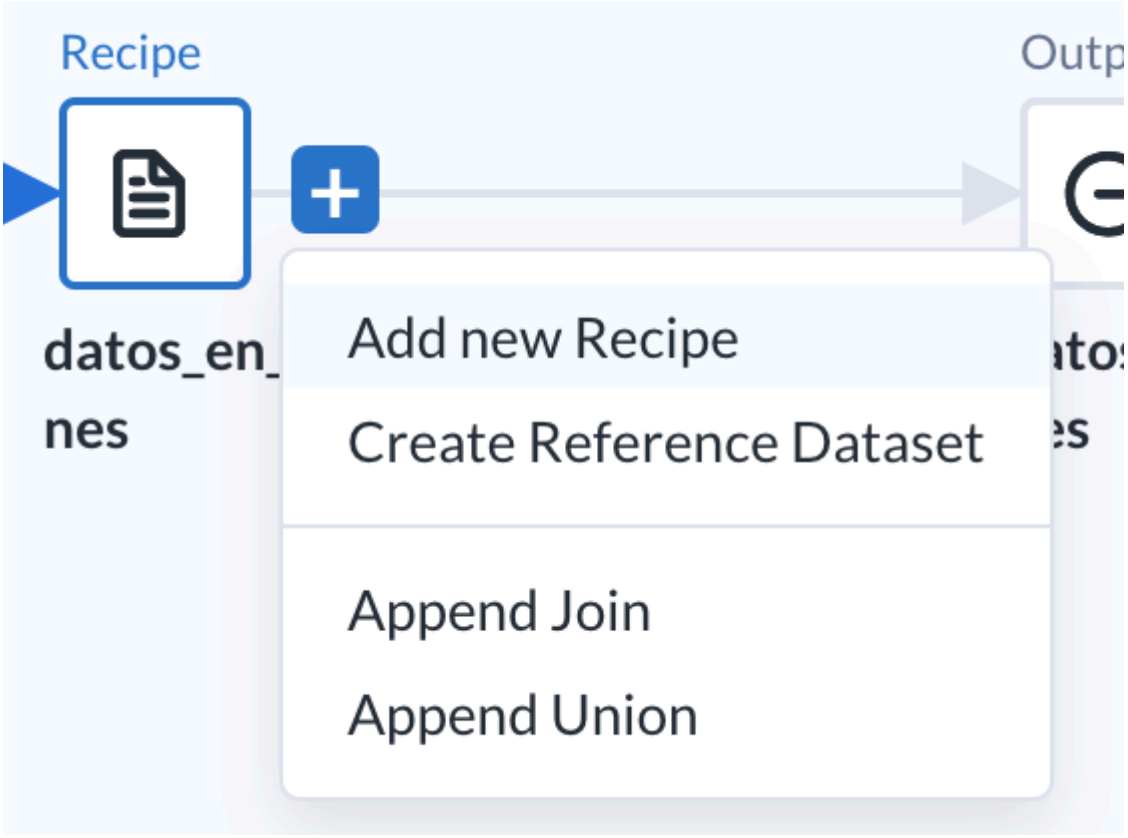
BigQuery / practica-1-big-data-mod... / practica2

Search...

NAME

datos_en_bruto_todas_las_sesiones

Responda:


Recipe
Outp
datos_en_
nes
Add new Recipe
Create Reference Dataset
Append Join
Append Union

1 New Dataset Clear All

datos_en_bruto_todas_las_se

Add a Description

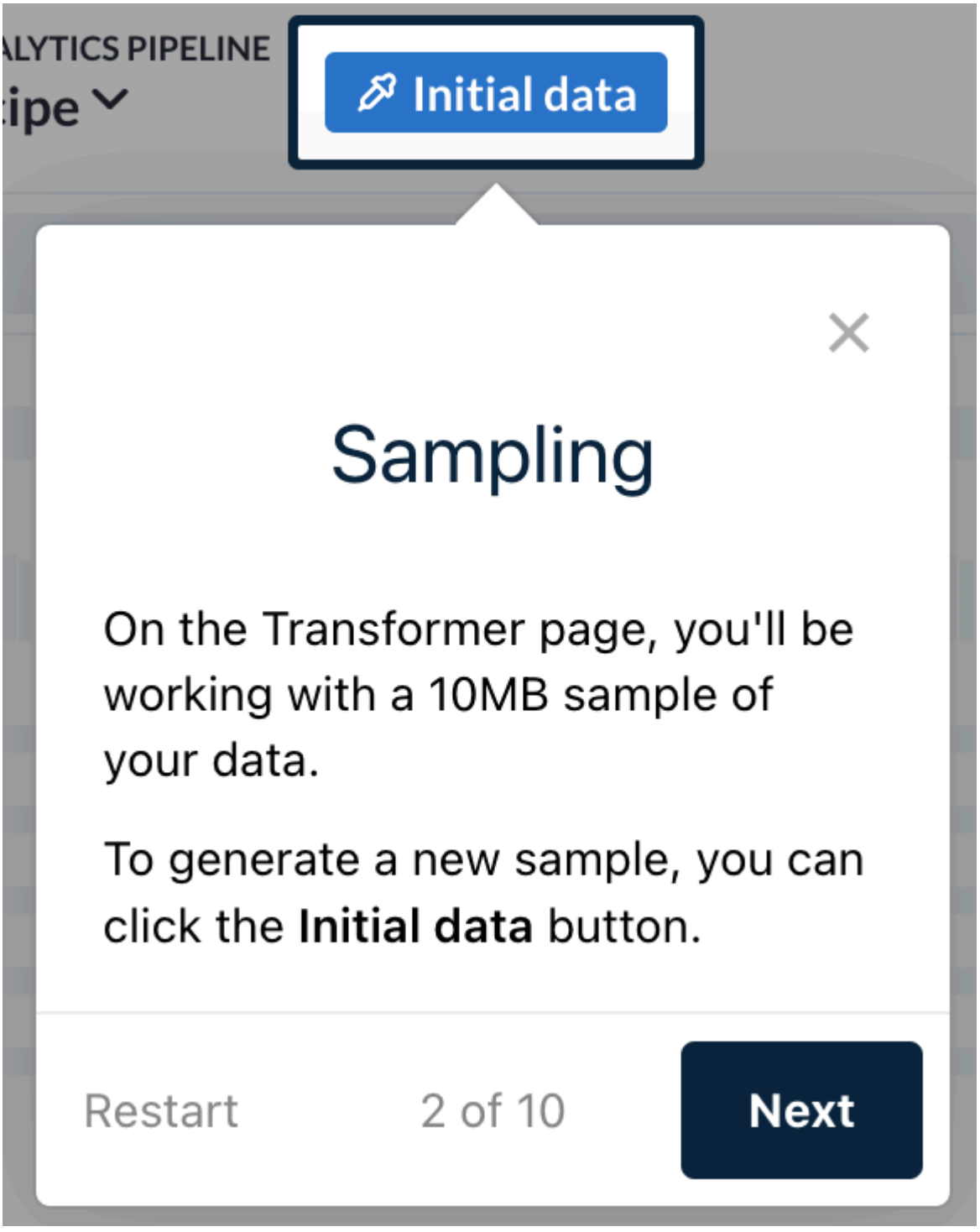
fullVisitorId	channe
8074041050560984021	Organic Se
8074041050560984021	Organic Se
8685530477324183365	Display
3395445735354444853	Direct
3173566250804266498	Organic Se

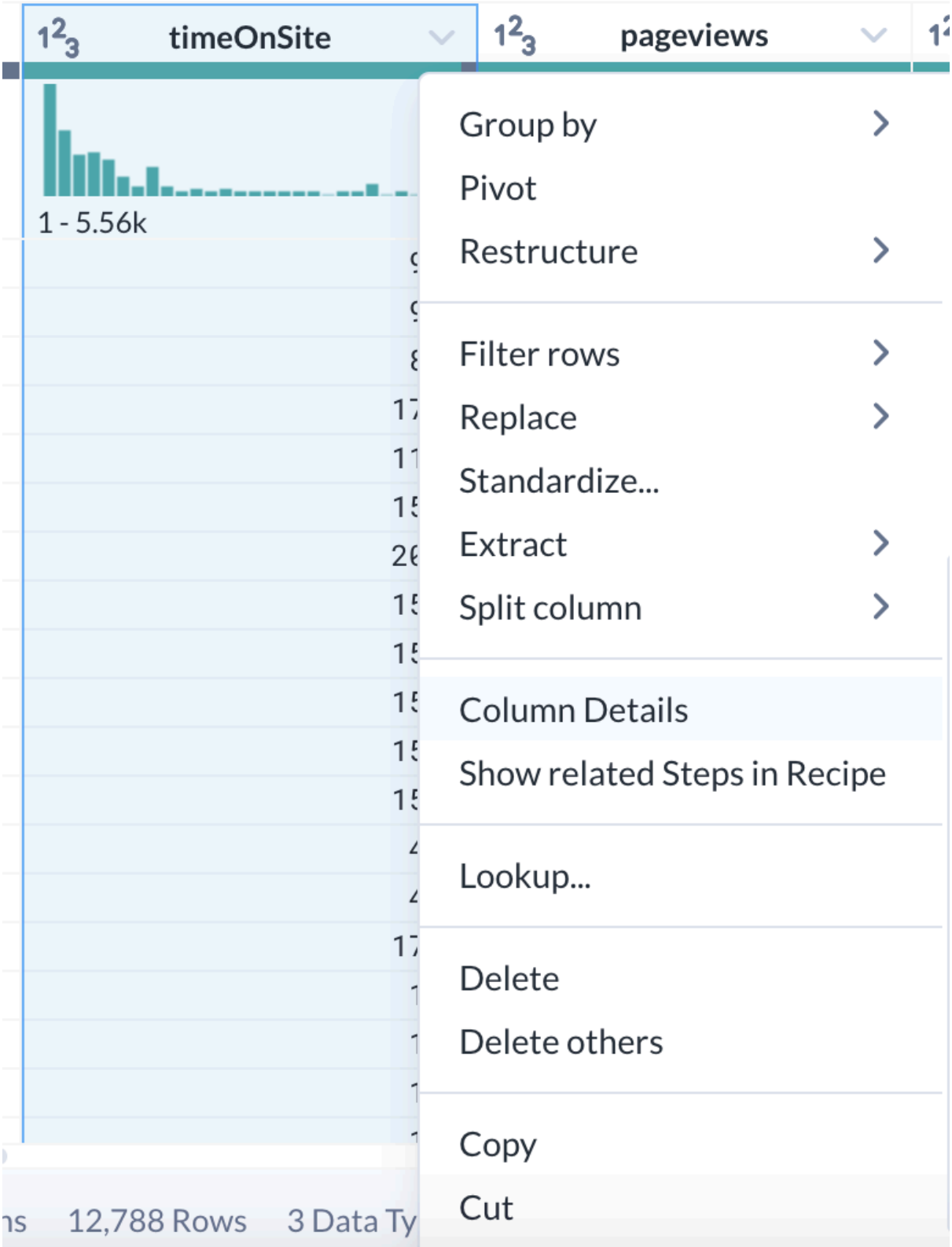
Edit settings

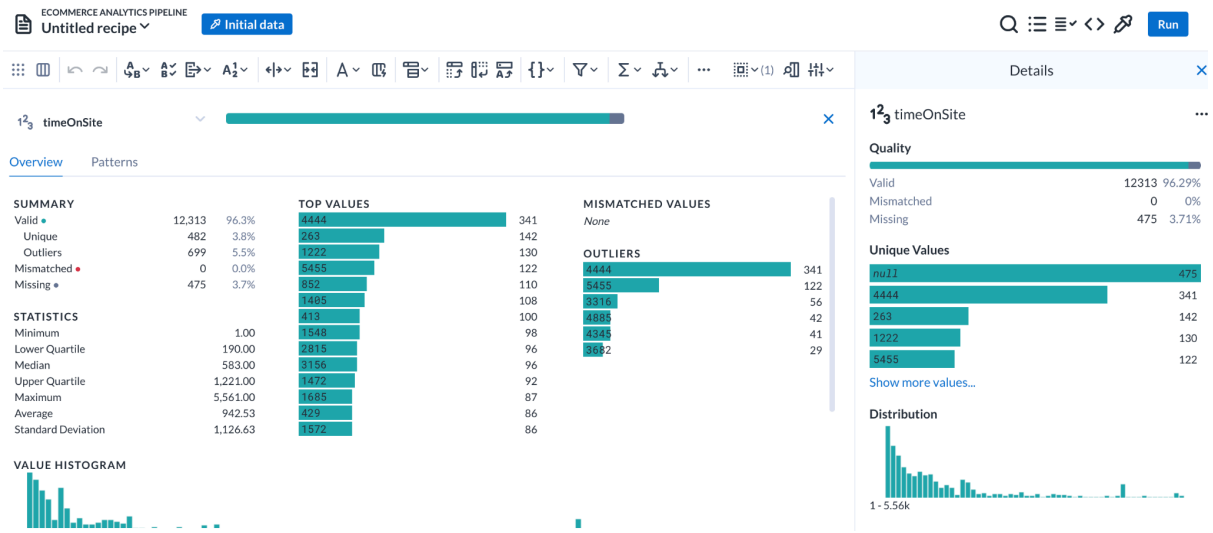
Untitled recipe

Welcome to the Transformer

Let's take a look around.


ALYTICS PIPELINE
ipe
Initial data
Sampling
On the Transformer page, you'll be working with a 10MB sample of your data.
To generate a new sample, you can click the **Initial data** button.
Restart
2 of 10
Next

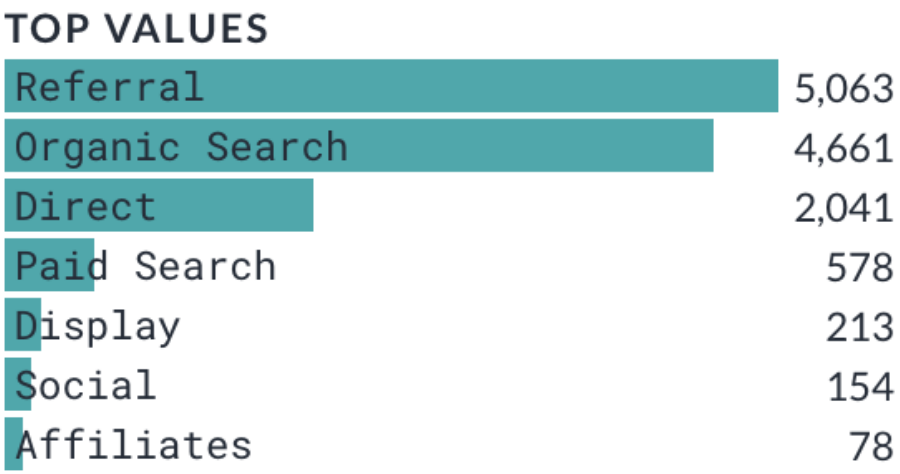

timeOnSite
pageviews
1 - 5.56k
Group by
Pivot
Restructure
Filter rows
Replace
Standardize...
Extract
Split column
Column Details
Show related Steps in Recipe
Lookup...
Delete
Delete others
Copy
Cut
12,788 Rows
3 Data Ty



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?

STATISTICS	
Minimum	1.00
Lower Quartile	190.00
Median	583.00
Upper Quartile	1,221.00
Maximum	5,561.00
Average	942.53
Standard Deviation	1,126.63

5. ¿Y el valor máximo de pageviews?

STATISTICS	
Minimum	1.00
Lower Quartile	7.00
Median	14.00
Upper Quartile	25.00
Maximum	155.00
Average	20.43
Standard Deviation	22.03

6. ¿Y el valor máximo de sessionQualityDim?

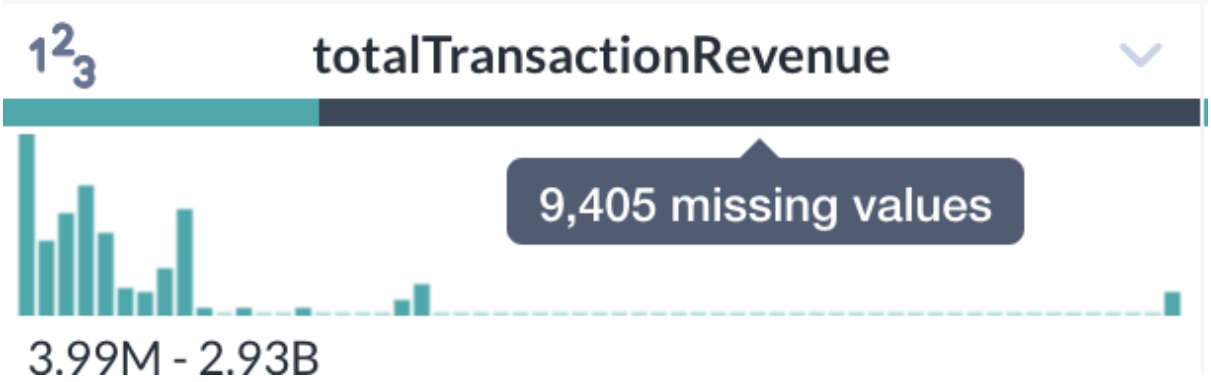
STATISTICS	
Minimum	1.00
Lower Quartile	2.00
Median	27.00
Upper Quartile	75.00
Maximum	97.00
Average	38.35
Standard Deviation	35.94

7. ¿Cuáles son los tres primeros países desde donde se originaron sesiones?

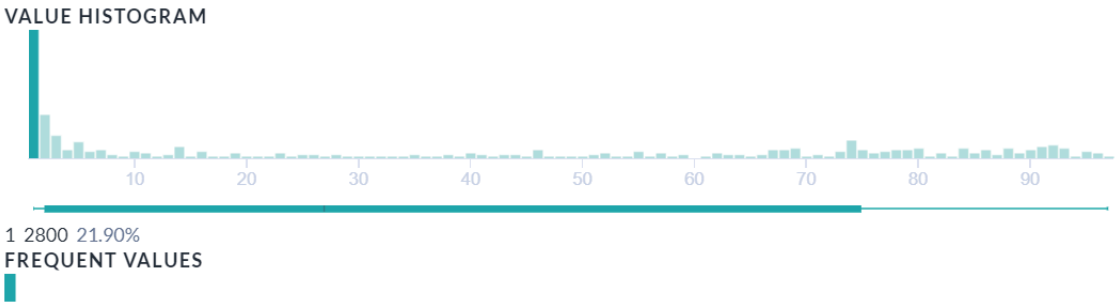
TOP VALUES

United States	10,226
India	344
United Kingdom	292
Canada	260
France	210
Germany	136
Australia	113
Japan	108
Singapore	95
Malaysia	73
Italy	68
Austria	59
Brazil	59
Dominican Republic	53

8. ¿Qué representa la barra gris que se encuentra debajo de totalTransactionRevenue?



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



No, ya que tal y como se puede observar, la distribución de los datos no responde a un orden ascendente o descendente. Sino que los datos se encuentran desordenados, viendo picos de subida y de bajada constantes.

10. Cuál es el período para el conjunto de datos?

STATISTICS	
Minimum	Aug 1 2017 00:00
Lower Quartile	Aug 1 2017 00:00
Median	Aug 1 2017 00:00
Upper Quartile	Aug 1 2017 00:00
Maximum	Aug 1 2017 00:00
Average	Aug 1 2017 00:00

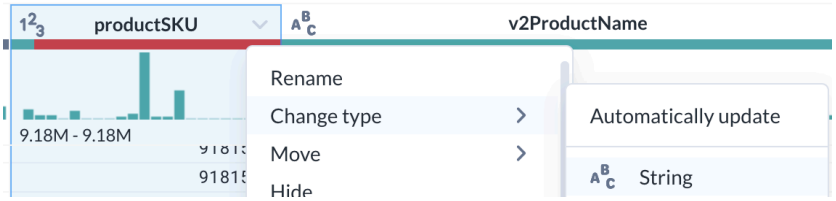
11. Puede que vea una barra roja debajo de la columna productSKU. De ser así, ¿que podría significar?

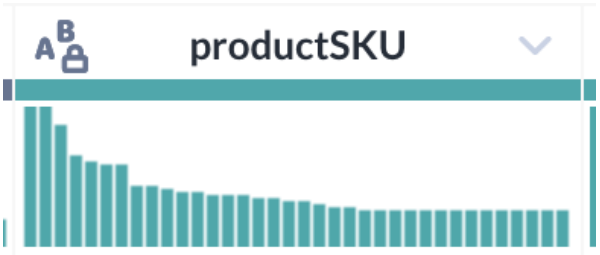
Son todos aquellos valores que no coinciden. Discrepancia entre los datos almacenados en diferentes ubicaciones o campos. Por ejemplo, si dos campos deberían contener la misma información pero tienen valores diferentes

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

1. Limpieza de datos

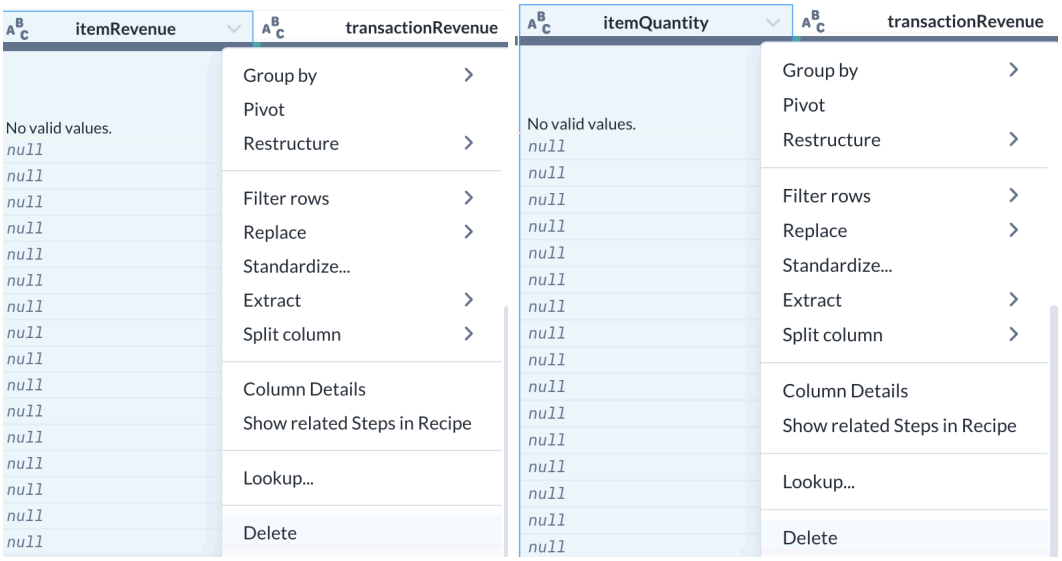
A. CAMBIAMOS EL TYPE DE LA COLUMNA PRODUCTSKU a string





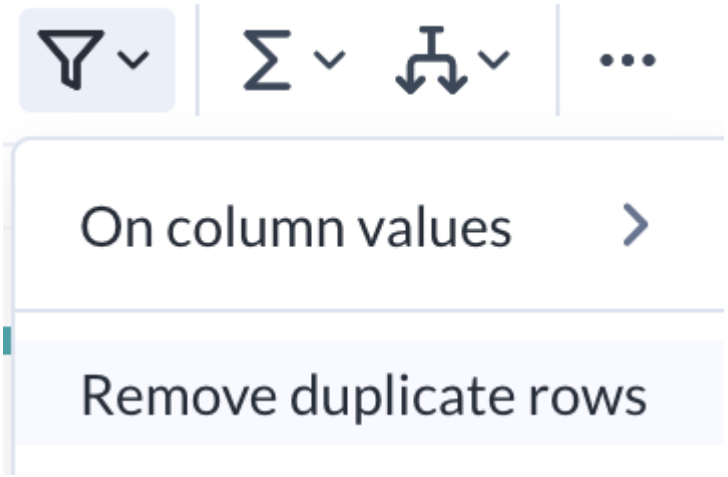
Observamos el historial de las acciones realizadas en la recipe y confirmamos el cambio realizado.

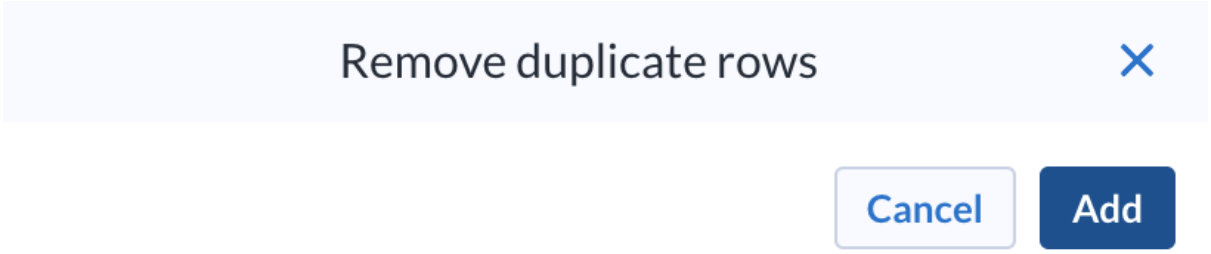
2. Borrar columnas innecesarias
Se procede al delete de las columnas innecesarias, siendo estas itemrevenue y ItemQuantity respectivamente.



3. Quitar duplicados

Se procede a la eliminación de aquellos valores duplicados.



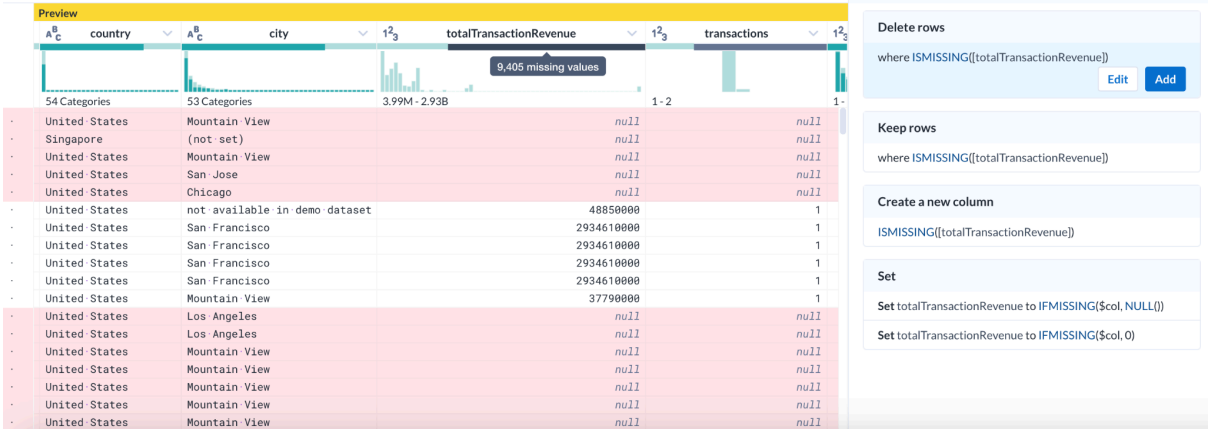


1 **Lock** productSKU **type** to String

2 **Delete** itemRevenue

3 **Delete** itemQuantity

4 **Remove duplicate** rows



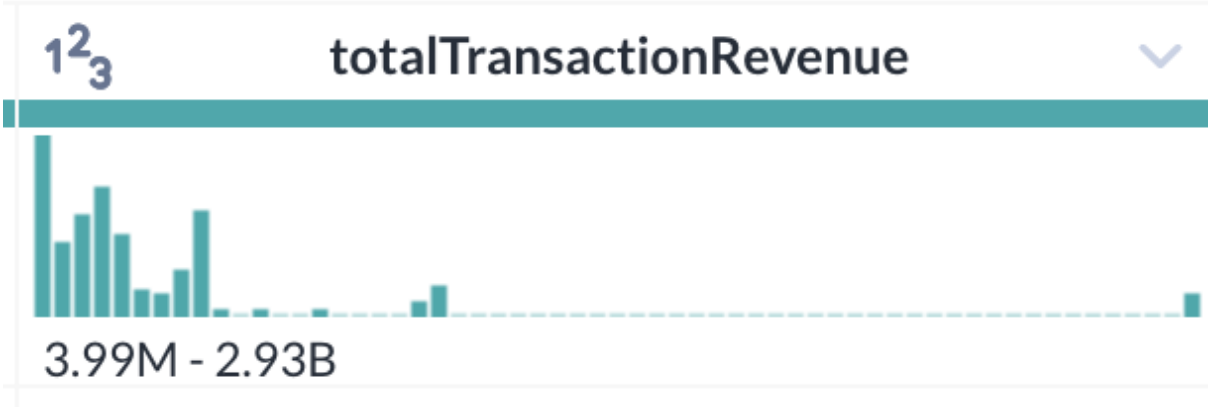
1 **Lock** productSKU **type** to String

2 **Delete** itemRevenue

3 **Delete** itemQuantity

4 **Remove duplicate** rows

5 **Delete rows** where ISMISSING([totalTransactionRevenue])

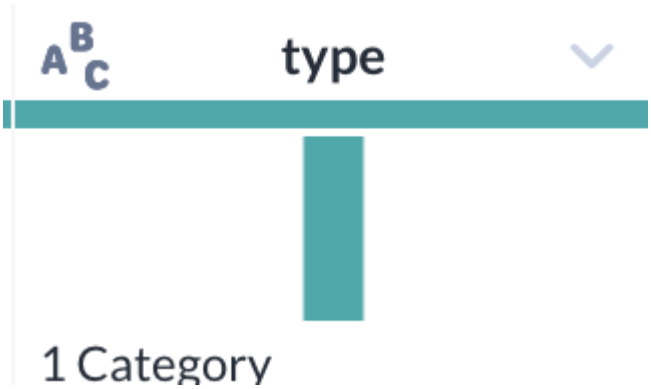


Confirmamos que las acciones quedaron registradas en el propio historial de la recipe.

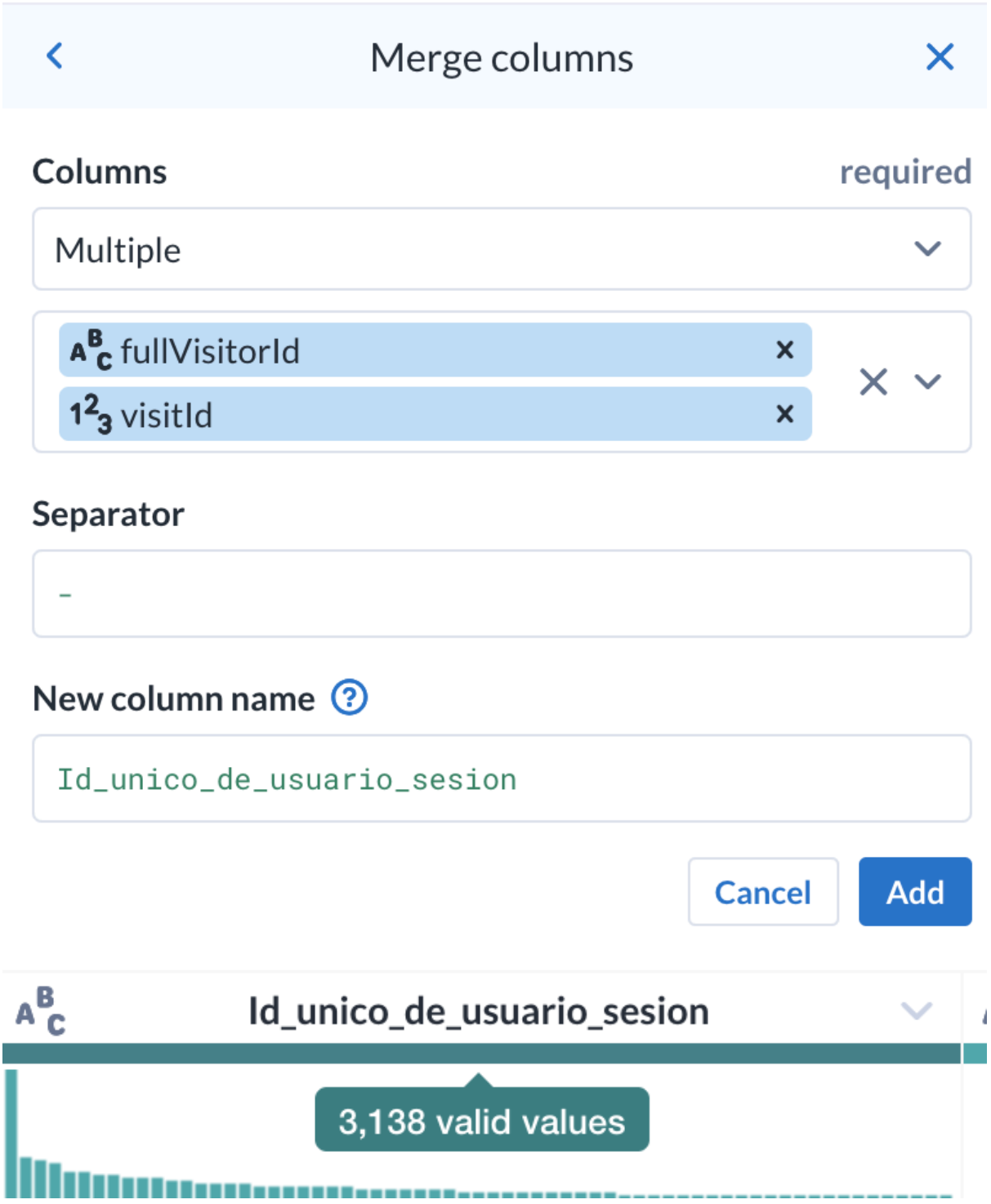
Keep rows

where type == 'PAGE'

Edit Add



1 Category


Merge columns
Columns
required
Multiple
fullVisitorId
visitId
Separator
_
New column name
Id_unico_de_usuario_sesion
Cancel
Add
Id_unico_de_usuario_sesion
3,138 valid values

1 **Lock** productSKU **type** to String

2 **Delete** itemRevenue

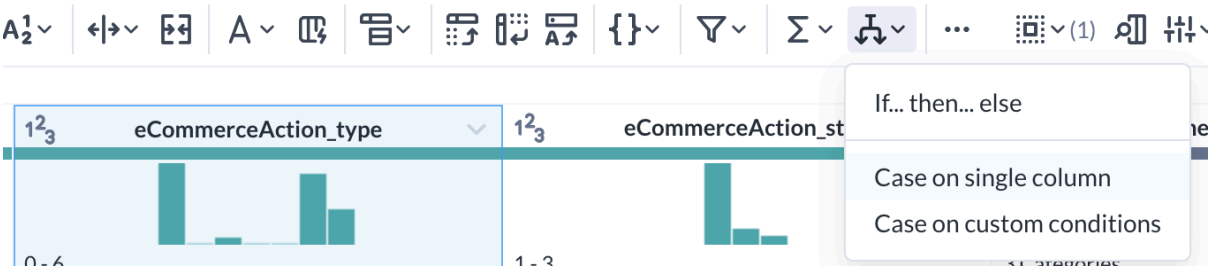
3 **Delete** itemQuantity

4 **Remove duplicate** rows

5 **Delete rows** where ISMISSING([totalTransactionRevenue])

6 **Keep rows** where type == 'PAGE'

7 **Concatenate** fullVisitorId, visitId separated by '-'



Condition type required

Case on single column

Specify multiple conditions on a single value or formula, using the case statement

Column to evaluate required

123 eCommerceAction_type

Cases (10) Add

Comparison

Enter a value or formula

New value

Enter a value or formula

Comparison

Comparison

```
0
```

New value

```
"Unknown"
```

Comparison

```
1
```

New value

```
"Click through of product lists"
```

Comparison

```
2
```

New value

```
"Product detail views"
```

Comparison

3

New value

"Add product(s) to cart"

Comparison

4

New value

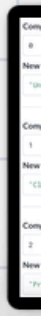
"Remove product(s) from cart"

Comparison

5

New value

"Check out"



Comparison

6

New value

"Completed purchase"

Comparison

7

New value

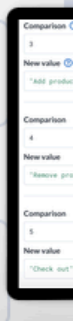
"Refund of purchase"

Comparison

8

New value

"Checkout options"



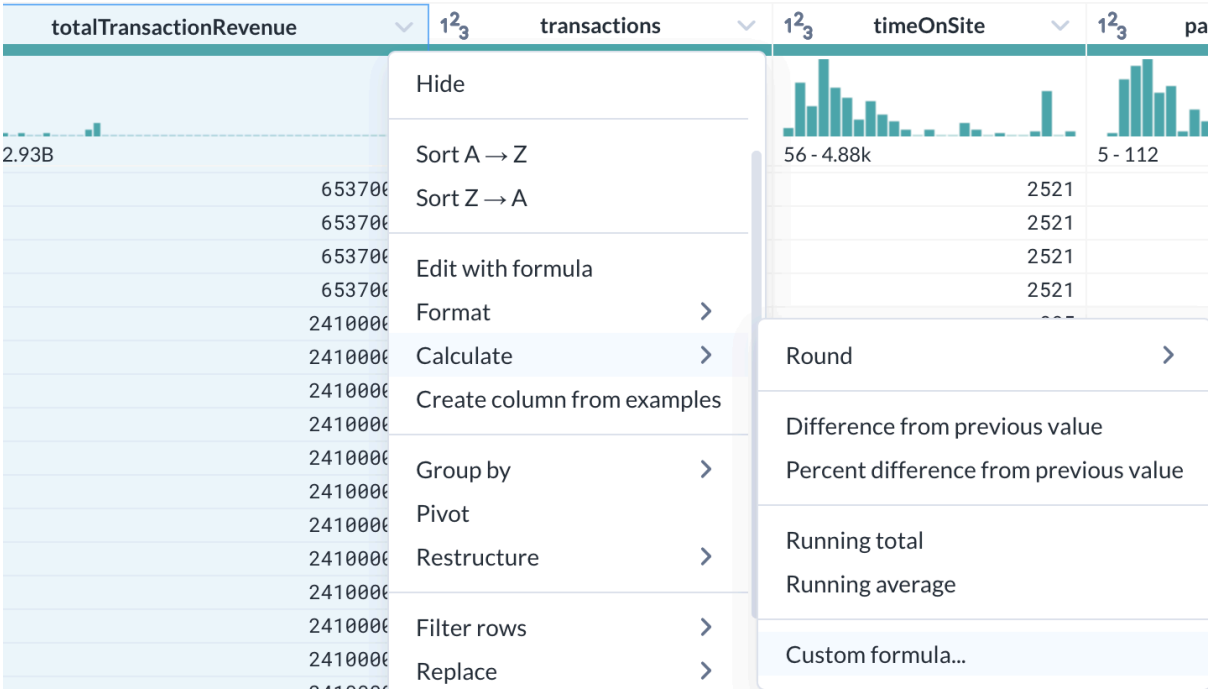
Default value

```
Edit formula
```

New column name

```
eCommerceAction_texto
```

1. **Lock** productSKU **type** to String
2. **Delete** itemRevenue
3. **Delete** itemQuantity
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



New formula

Formula type required

Single row formula

Create a new column from a single row formula

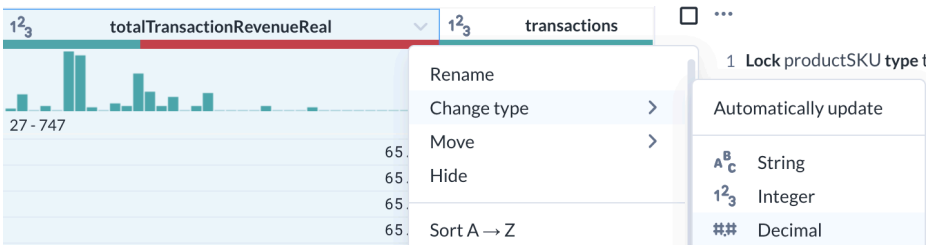
Formula required

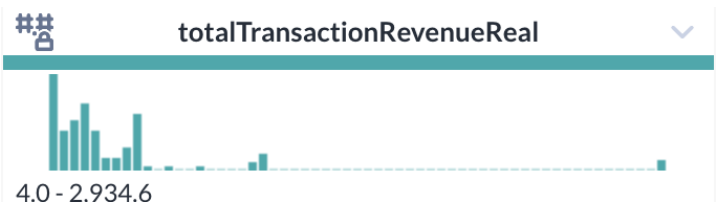
```
DIVIDE(totalTransactionRevenue, 1000000)
```

New column name

totalTransactionRevenueReal

Cancel Add





1 **Lock** productSKU **type** to String

2 **Delete** itemRevenue

3 **Delete** itemQuantity

4 **Remove duplicate** rows

5 **Delete rows** where ISMISSING([totalTransactionRevenue])

6 **Keep rows** where type == 'PAGE'

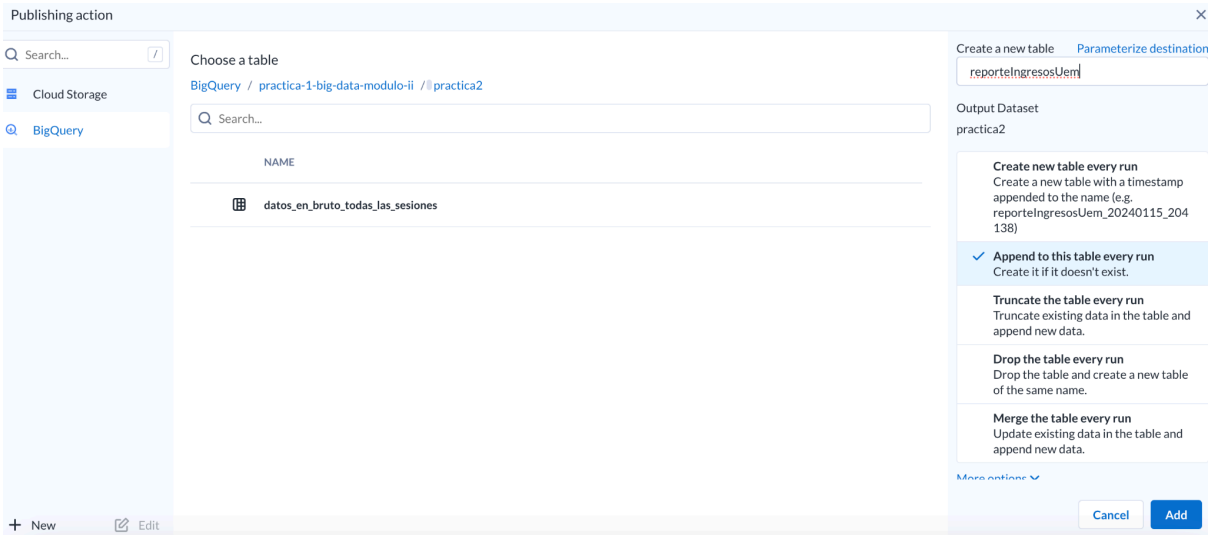
7 **Concatenate** fullVisitorId, visitId separated by

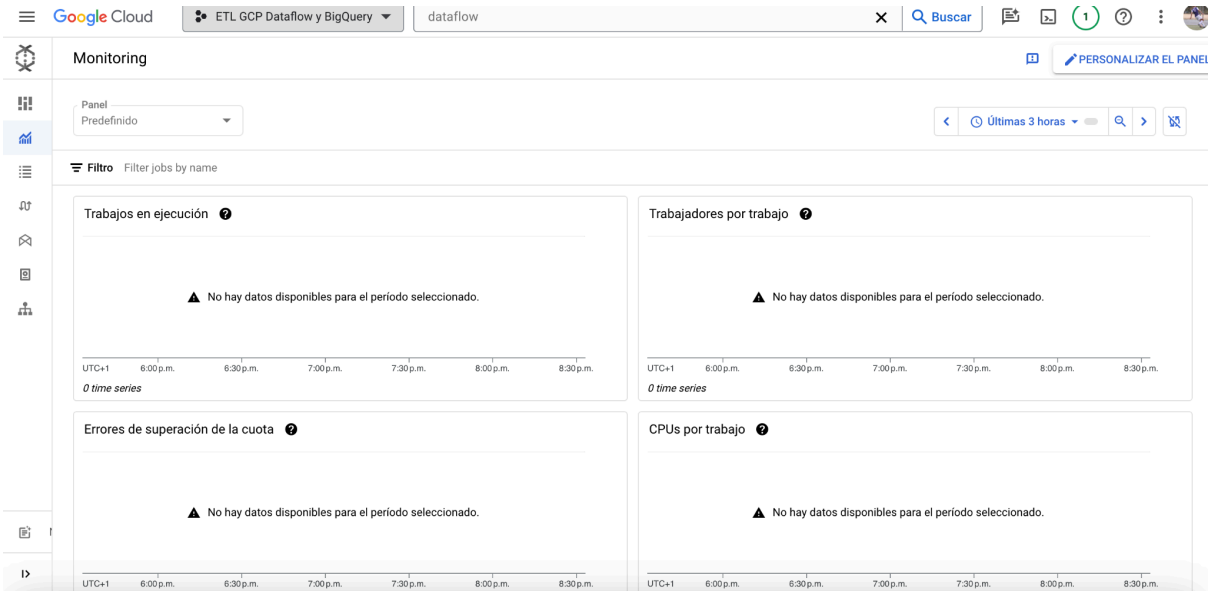
8 **Create** eCommerceAction_texto from 9 case conditions on eCommerceAction_type

9 **Create** totalTransactionRevenueReal from DIVIDE(totalTransactionRevenue, 1000000)

10 **Lock** totalTransactionRevenueReal **type** to Decimal

CORREMOS EL TRABAJO





Thea from Trifacta

🎉 Congratulations on your first job run!

Your job is running and your data will be ready once all the transformations are applied. You will then be able to access your final data as well as useful metadata.

If you'd like, you can leave this page while the job is running. A notification will be sent when the job is completed and you can always get back to this page.

Meanwhile, let me show you where to find all the details about your job and your final data.

Next

Sin título 2 | EJECUTAR | GUARDAR | DESCARGAR | COMPARTIR | PROGRAMACIÓN | MÁS | Esta secuencia de

```
SAFE_SUBTRACT(RANGE_BUCKET(t8.`_f29`, t4.`_f29`.buckets), 1)), SAFE_SUBTRACT(t4.`_f29`.numBins, 1)), 20)) AS eCommerceAction_type,
FORMAT_CATEGORICAL_PROFILE_RESULTS(ANY_VALUE(t4.`_f54`), 20) AS eCommerceAction_texto, FORMAT_NUMERICAL_PROFILE_RESULTS(ANY_VALUE(t4.
`_f34`.min), ANY_VALUE(t4.`_f34`.max), ANY_VALUE(t4.`_f34`.quartiles), ANY_VALUE(t4.`_f34`.roundMin), ANY_VALUE(t4.`_f34`.roundMax),
ANY_VALUE(t4.`_f34`.numBins), APPROX_TOP_COUNT(LEAST(GREATEST(0, SAFE_SUBTRACT(RANGE_BUCKET(t8.`_f31`, t4.`_f34`.buckets), 1)),
SAFE_SUBTRACT(t4.`_f34`.numBins, 1)), 20)) AS eCommerceAction_step, FORMAT_CATEGORICAL_PROFILE_RESULTS(ANY_VALUE(t4.`_f58`), 20) AS
eCommerceAction_option
14 FROM (SELECT APPROX_TOP_COUNT(`_f0`, 21) AS `_f0`, APPROX_TOP_COUNT(`_f1`, 21) AS `_f1`, COMPUTE_BUCKET_INFO(MIN(SAFE_CAST(`_f2` AS
FLOAT64)), MAX(SAFE_CAST(`_f2` AS FLOAT64)), APPROX_QUANTILES(SAFE_CAST(`_f2` AS FLOAT64), 4), 20) AS `_f2`, APPROX_TOP_COUNT(`_f3`, 21)
AS `_f5`, APPROX_TOP_COUNT(`_f4`, 21) AS `_f6`, COMPUTE_BUCKET_INFO(MIN(SAFE_CAST(`_f5` AS FLOAT64)), MAX(SAFE_CAST(`_f5` AS FLOAT64)),
APPROX_QUANTILES(SAFE_CAST(`_f5` AS FLOAT64), 4), 20) AS `_f7`, COMPUTE_BUCKET_INFO(MIN(SAFE_CAST(`_f6` AS FLOAT64)), MAX(SAFE_CAST(`_f6`
AS FLOAT64)), APPROX_QUANTILES(SAFE_CAST(`_f6` AS FLOAT64), 4), 20) AS `_f8`, COMPUTE_BUCKET_INFO(MIN(SAFE_CAST(`_f7` AS FLOAT64)), MAX
(SAFE_CAST(`_f7` AS FLOAT64)), APPROX_QUANTILES(SAFE_CAST(`_f7` AS FLOAT64), 4), 20) AS `_f9`, COMPUTE_BUCKET_INFO(MIN(SAFE_CAST(`_f8` AS
```

Ubicación de procesamiento: US

Presiona Option+F1 para ver las opciones de accesibi

Todos los resultados

Tiempo transcurrido	Declaraciones procesadas	Estado del trabajo
7 s	4	SUCCESS

Estado	Hora de finalización	SQL	Acción
✓	8:29 p.m. [2:1]	CREATE TEMP FUNCTION COMPUTE_BUCKET_INFO(min FLOAT64, max FLOAT64, quartiles ARRAY<FLOAT64	VER RESULTADOS
✓	8:29 p.m. [9:1]	CREATE TEMP FUNCTION COMPUTE_BUCKET_INFO(min FLOAT64, max FLOAT64, quartiles ARRAY<FLOAT64	VER RESULTADOS
✓	8:29 p.m. [19:1]	CREATE TEMP FUNCTION COMPUTE_BUCKET_INFO(min FLOAT64, max FLOAT64, quartiles ARRAY<FLOAT64	VER RESULTADOS
✓	8:29 p.m. [29:1]	CREATE TEMP FUNCTION COMPUTE_BUCKET_INFO(min FLOAT64, max FLOAT64, quartiles ARRAY<FLOAT64	VER RESULTADOS

Ecommerce Analytics Pipeline > Untitled recipe

Job 23334707

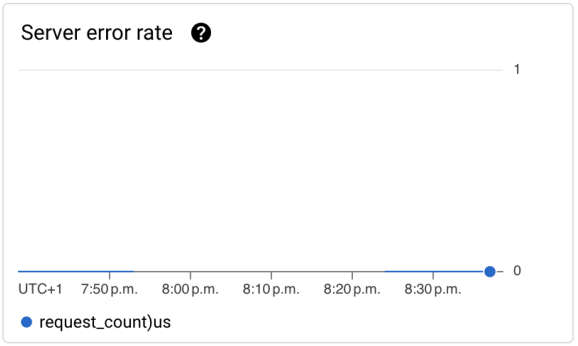
Finished Today at 8:30 PM

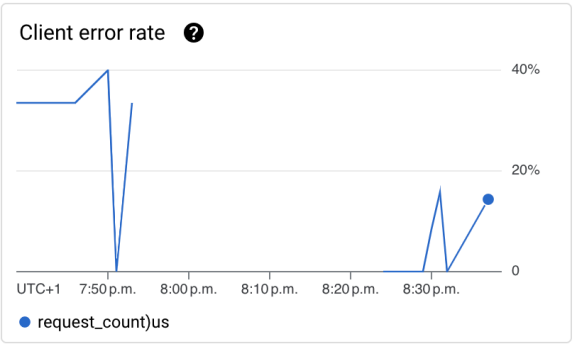
View BigQuery job

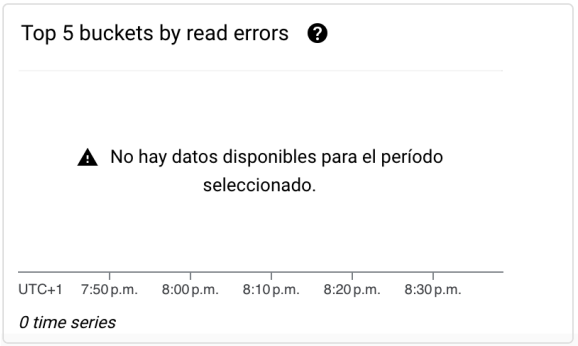
Overview | Output destinations | Profile | Dependency graph | Data sources

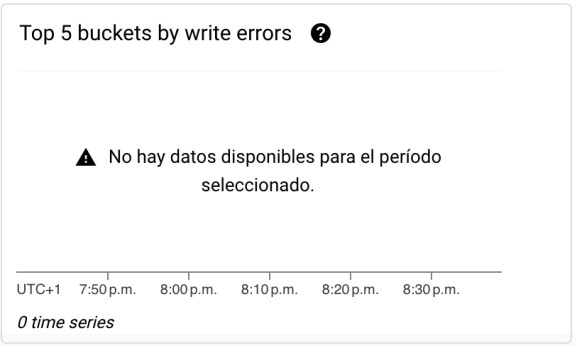
1. Habilita la API de Dataflow — ✓ LISTO
2. Habilita la API de Data Pipelines — EN EJECUCIÓN
3. Habilita la API de Cloud Scheduler — EN EJECUCIÓN

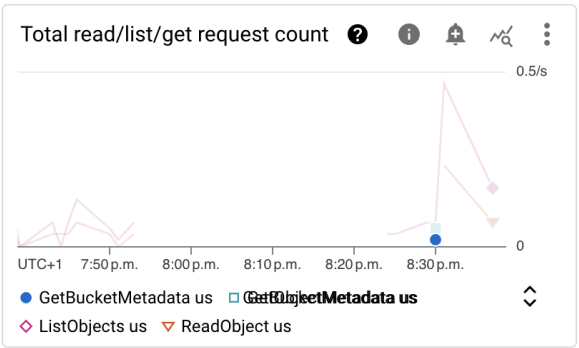
FIX MISSING STEPS

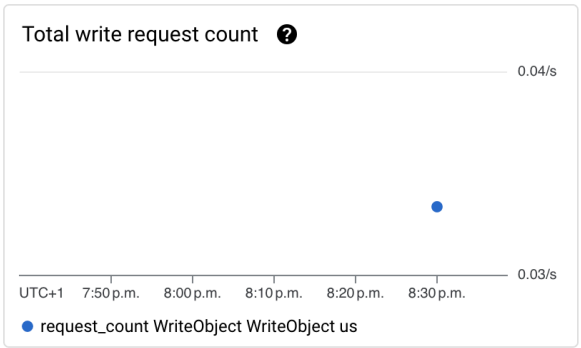

Server error rate
1
0
UTC+1 7:50 p.m. 8:00 p.m. 8:10 p.m. 8:20 p.m. 8:30 p.m.
request_count)us

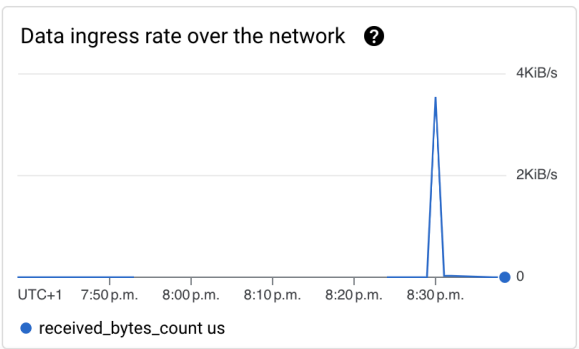

Client error rate
40%
20%
0
UTC+1 7:50 p.m. 8:00 p.m. 8:10 p.m. 8:20 p.m. 8:30 p.m.
request_count)us

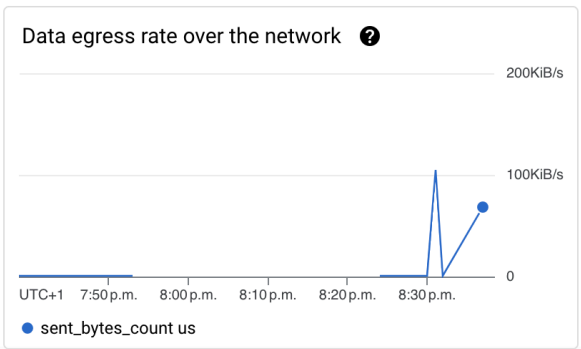

Top 5 buckets by read errors
No hay datos disponibles para el período seleccionado.
UTC+1 7:50 p.m. 8:00 p.m. 8:10 p.m. 8:20 p.m. 8:30 p.m.
0 time series


Top 5 buckets by write errors
No hay datos disponibles para el período seleccionado.
UTC+1 7:50 p.m. 8:00 p.m. 8:10 p.m. 8:20 p.m. 8:30 p.m.
0 time series


Total read/list/get request count
0.5/s
0
UTC+1 7:50 p.m. 8:00 p.m. 8:10 p.m. 8:20 p.m. 8:30 p.m.
GetBucketMetadata us
ListObjects us
ReadObject us


Total write request count
0.04/s
0.03/s
UTC+1 7:50 p.m. 8:00 p.m. 8:10 p.m. 8:20 p.m. 8:30 p.m.
request_count WriteObject WriteObject us


Data ingress rate over the network
4KiB/s
2KiB/s
0
UTC+1 7:50 p.m. 8:00 p.m. 8:10 p.m. 8:20 p.m. 8:30 p.m.
received_bytes_count us


Data egress rate over the network
200KiB/s
100KiB/s
0
UTC+1 7:50 p.m. 8:00 p.m. 8:10 p.m. 8:20 p.m. 8:30 p.m.
sent_bytes_count us