

Exploring and Creating an Ecommerce Analytics Pipeline with Cloud Dataprep v1.5

1 hour 30 minutesFree

Rate Lab

Overview

[Cloud Dataprep](#) by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. In this lab we

will explore the Cloud Dataprep UI to build an ecommerce transformation pipeline that will run at a scheduled interval and output results back into BigQuery. The dataset we will be using is an [ecommerce dataset](#) that has millions of Google Analytics records for the [Google Merchandise Store](#) loaded into BigQuery. We've made a copy of that dataset for this lab and will be exploring the available fields and rows for insights.

Objectives

In this lab, you learn how to perform these tasks:

- Connect BigQuery datasets to Cloud Dataprep
- Explore dataset quality with Cloud Dataprep
- Create a data transformation pipeline with Cloud Dataprep
- Schedule transformation jobs outputs to BigQuery

What you'll need

- A Google Cloud Platform project
- The [Google Chrome](#) browser. Cloud Dataprep supports only the Chrome browser.

Setup and requirements


For each lab, you get a new Google Cloud project and set of resources for a fixed time at no cost.

1. Sign in to Qwiklabs using an incognito window.
2. Note the lab's access time (for example, 1:15:00), and make sure you can finish within that time. There is no pause feature. You can restart if needed, but you have to start at the beginning.
3. When ready, click Start lab.
4. Note your lab credentials (Username and Password). You will use them to sign in to the Google Cloud Console.
5. Click Open Google Console.
6. Click Use another account and copy/paste credentials for this lab into the prompts. If you use other credentials, you'll receive errors or incur charges.
7. Accept the terms and skip the recovery resource page.

Note: Do not click End Lab unless you have finished the lab or want to restart it. This clears your work and removes the project.

Check project permissions

Before you begin your work on Google Cloud, you need to ensure that your project has the correct permissions within Identity and Access Management (IAM).

1. In the Google Cloud console, on the Navigation menu () , select IAM & Admin > IAM.
2. Confirm that the default compute Service Account `{project-number}-compute@developer.gserviceaccount.com` is present and has the `editor` role assigned. The account prefix is the project number, which you can find on Navigation menu > Home.

Google Cloud Platform | qwiklabs-gcp-03-e30ac90a32e4 | Search products and resources

IAM & Admin

IAM | ADD | REMOVE

PERMISSIONS | RECOMMENDATIONS HISTORY

Permissions for project "qwiklabs-gcp-03-e30ac90a32e4"

These permissions affect this project and all of its resources. [Learn more](#)

View By: **PRINCIPALS** | ROLES

Filter: Enter property name or value

Type	Principal ↑	Name	Role
<input type="checkbox"/>	407543585891-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor
<input type="checkbox"/>	407543585891@cloudbuild.gserviceaccount.com		Cloud Build Service Account
<input type="checkbox"/>	407543585891@cloudservices.gserviceaccount.com	Google APIs Service Agent	Editor
<input type="checkbox"/>	admiral@qwiklabs-services-prod.iam.gserviceaccount.com		Owner
<input type="checkbox"/>	qwiklabs-gcp-03-e30ac90a32e4@qwiklabs-gcp-03-e30ac90a32e4.iam.gserviceaccount.com	Qwiklabs User Service Account	App Engine Admin BigQuery Admin

Note: If the account is not present in IAM or does not have the `editor` role, follow the steps below to assign the required role.

1. In the Google Cloud console, on the Navigation menu, click Home.
2. Copy the project number (e.g. 729328892908).
3. On the Navigation menu, select IAM & Admin > IAM.
4. At the top of the IAM page, click Add.
5. For New principals, type:

`{project-number}-compute@developer.gserviceaccount.com`
Copied!

content_copy

6. Replace `{project-number}` with your project number.
7. For Role, select Project (or Basic) > Editor.
8. Click Save.

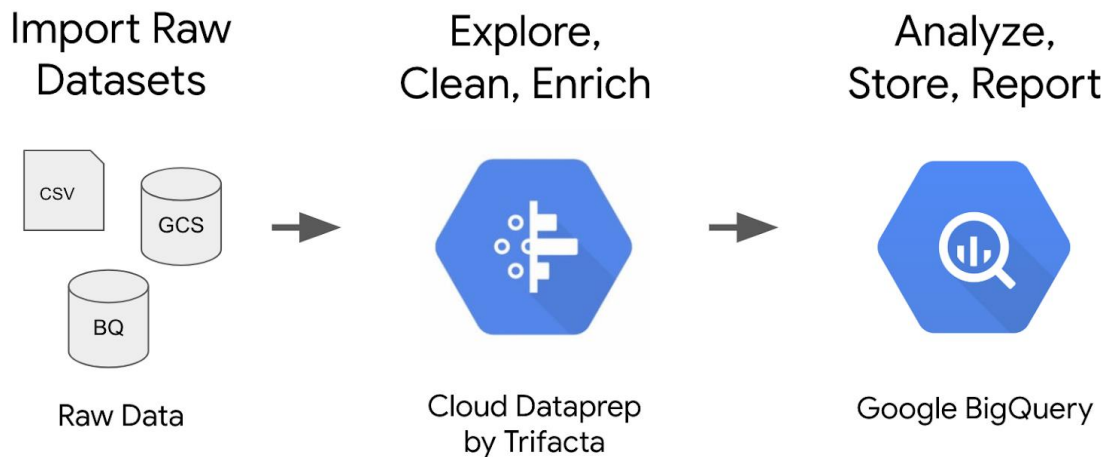
Open BigQuery Console

1. In the Google Cloud Console, select Navigation menu > BigQuery.

The Welcome to BigQuery in the Cloud Console message box opens. This message box provides a link to the quickstart guide and lists UI updates.

2. Click Done.

Although this lab is largely focused on Cloud Dataprep, you need BigQuery as an endpoint for dataset ingestion to the pipeline and as a destination for the output when the pipeline is completed.



Task 1. Create an empty BigQuery dataset

In this task, you create a new BigQuery dataset to receive the output table of your new pipeline.

1. In the left pane, click View actions (⋮) next to your project ID and select Create dataset.
2. In the Create dataset dialog:
 - For Dataset ID, type ecommerce.
 - Leave the other values at their defaults.
3. Click Create dataset.

4. Copy and paste this SQL query into the Query editor text field:

```
#standardSQL
CREATE OR REPLACE TABLE ecommerce.all_sessions_raw_dataprep
OPTIONS(
  description="Raw data from analyst team to ingest into Cloud
Dataprep"
) AS
SELECT * FROM `data-to-insights.ecommerce.all_sessions_raw`
WHERE date = '20170801'; # limiting to one day of data 56k rows for
this lab
Copied!
```

content_copy

5. Click Run.

This query copies over a subset of the public raw ecommerce dataset to your own project dataset for you to explore and clean in Cloud Dataprep.

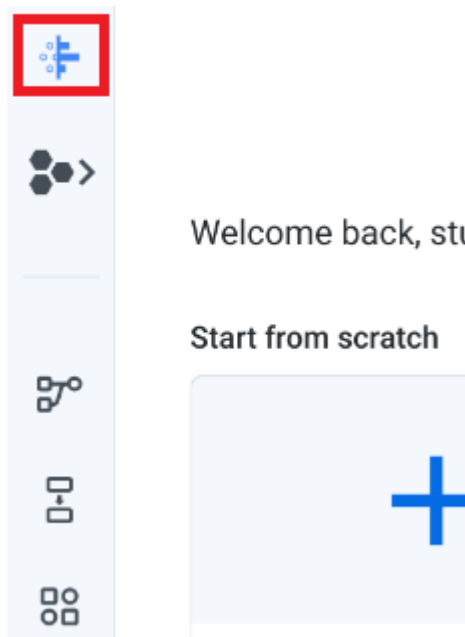
6. Confirm that the new raw data table exists in your project.

Task 2. Open Cloud Dataprep

In this task, you accept the terms of service for Google and Trifacta, and then you allow Trifacta to access your project data.

1. In the [GCP console](#), make sure that your lab's project is selected.
2. In the Navigation menu (≡), click Dataprep.
3. Select the Terms of Service for Google and Trifacta, and then click Accept.
4. In the Share account information with Trifacta dialog, select the checkbox, and then click Agree and Continue.
5. To allow Trifacta to access your project data, click Allow. This authorization process might take a few minutes.
6. In the Sign in with Google window appears, select your Qwiklab account and then click Allow. Click Accept if required after checking the checkbox.

7. To use the default location for the storage bucket, click Continue.



The homepage of Cloud Dataprep opens. If required, click Home.

Task 3. Connect BigQuery data to Cloud Dataprep

In this task, you connect Cloud Dataprep to your BigQuery data source.

On the Cloud Dataprep page:

1. Click Create a new flow.
2. Click Untitled Flow on the top of the page.
3. In the Rename dialog, specify these details:

- For Flow Name, type Ecommerce Analytics Pipeline
 - For Flow Description, type Revenue reporting table for Apparel
4. Click Ok.
 5. Click (+) icon to add a dataset.
 6. In the Add datasets to flow dialog, click Import datasets from bottom-left corner.
 7. In the left pane, click BigQuery.
 8. When your ecommerce dataset is loaded, click on it.
 9. To create a dataset, click Create dataset (+).
 10. Click Import & Add to Flow.

The data source automatically updates.

Task 4. Explore ecommerce data fields with a UI

In this task, you load and explore a sample of the dataset within Cloud Dataprep.

1. Click Edit Recipe in the right panel.
2. Click Don't show me any helpers in The Transformer dialog if required.

Cloud Dataprep loads a sample of your dataset into the Transformer view. This process might take a few minutes.



all_sessions_raw_dataprep – 2 ▾

Ecommerce Analytics Pipeline • Initial Sample

Grid

Columns

ABC	fullVisitorId	ABC	channelGrouping	#	time
	689 Categories		7 Categories		0 - 5.39M
·	8074041050560984021	·	Organic · Search	·	572599
·	8074041050560984021	·	Organic · Search	·	374400
·	8685530477324183365	·	Display	·	772010
·	3395445735354444853	·	Direct	·	1110096
·	3173566250804266498	·	Organic · Search	·	840497
·	8230528872482379210	·	Paid · Search	·	1270584
·	385231150756085903	·	Organic · Search	·	88302
·	9947542428111966715	·	Referral	·	22232
·	9947542428111966715	·	Referral	·	341867
·	9947542428111966715	·	Referral	·	405999
·	9947542428111966715	·	Referral	·	409719
·	8812275451738403277	·	Referral	·	29767



32 Columns

12,783 Rows

3 Data Types

Answer the questions:

How many columns are in the dataset?



64 columns



3 columns



32 columns

Submit

- Cloud Dataprep will load a sample of the source dataset for speed of exploration.

Note: When your pipeline is run, it will operate over the entire source dataset. How many rows does the sample contain?



all_sessions_raw_dataprep – 2 ▾

Ecommerce Analytics Pipeline • Initial Sample

Grid

Columns

ABC	fullVisitorId	ABC	channelGrouping	#	time
	689 Categories		7 Categories		0 - 5.39M
8074041050560984021		Organic · Search		572599	
8074041050560984021		Organic · Search		374400	
8685530477324183365		Display		772010	
3395445735354444853		Direct		1110096	
3173566250804266498		Organic · Search		840497	
8230528872482379210		Paid · Search		1270584	
385231150756085903		Organic · Search		88302	
9947542428111966715		Referral		22232	
9947542428111966715		Referral		341867	
9947542428111966715		Referral		405999	
9947542428111966715		Referral		409719	
8812275451738413277		Referral		29767	



32 Columns

12,783 Rows

3 Data Types

Answer : **About 12 thousands rows**

- What is the most common value in the channelGrouping column?



all_sessions_raw_dataprep – 2 ▾

Ecommerce Analytics Pipeline • Initial Sample

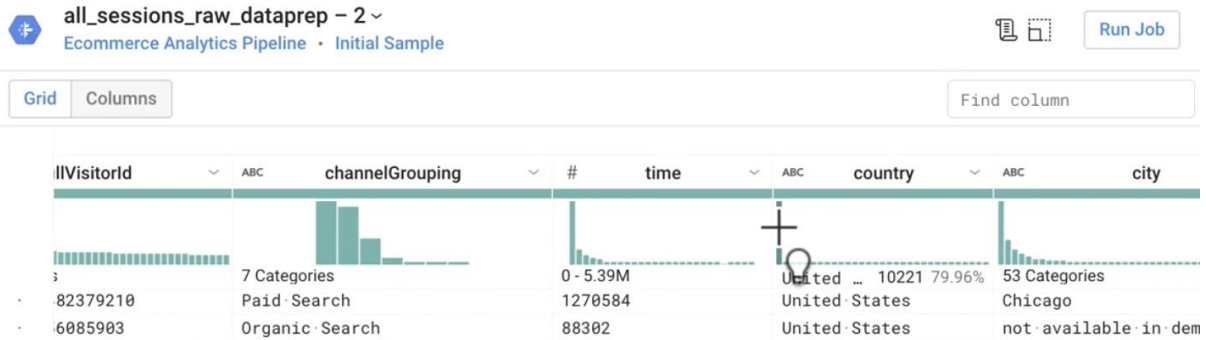
Grid

Columns

ABC	fullVisitorId	ABC	channelGrouping	#	time
	689 Categories		7 Categories		0 - 5.39M
8230528872482379210		Referral 5068 39.61%		1270584	
385231150756085903		Paid · Search		88302	
9947542428111966715		Organic · Search		22232	
		Referral			

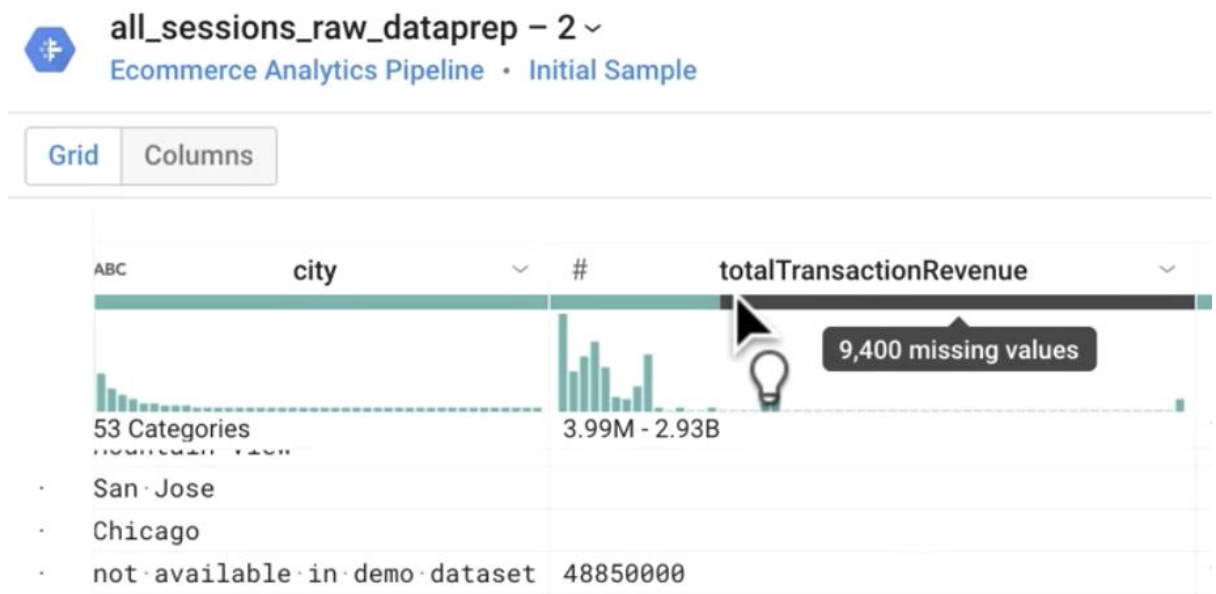
Answer : **Referral**

- What are the top three countries that sessions originate from?



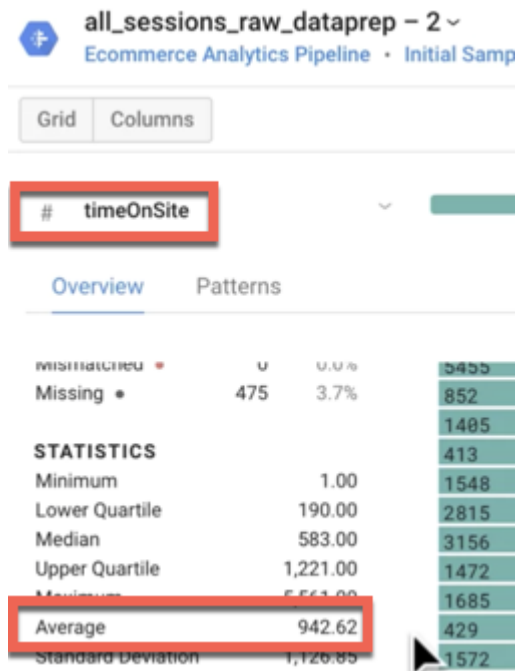
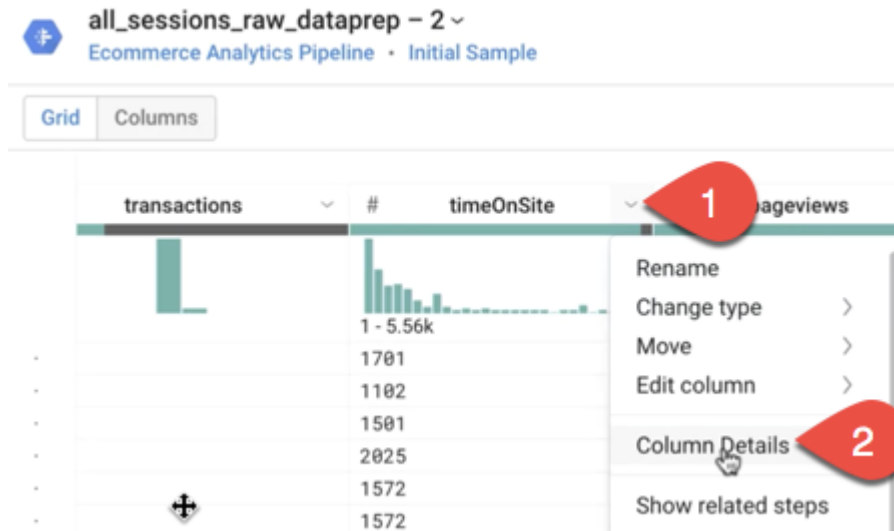
Answer: **US, India, United Kingdom**

- What does the gray bar under totalTransactionRevenue represent?



Answer: **Missing values**

- What is the average timeOnSite in seconds, average pageviews, and average sessionQualityDim for the data sample? (Hint: Use Column Details.)



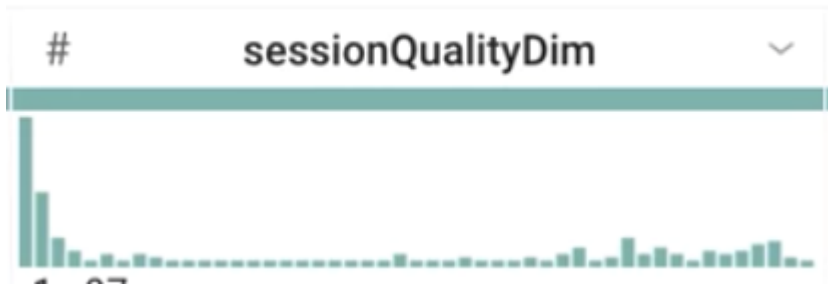
Answers : **Average Time On Site: 942 seconds (or 15.7 minutes)**

Average Pageviews: 20.44 pages

Average Session Quality Dimension: 38.36

Note: Your answers may vary slightly due to the data sample used by Cloud Dataprep.

- Looking at the histogram for sessionQualityDim, are the data values evenly distributed?



Answer: No, they are skewed to lower values (low quality sessions), which is expected.


- What is the date range for the dataset sample?

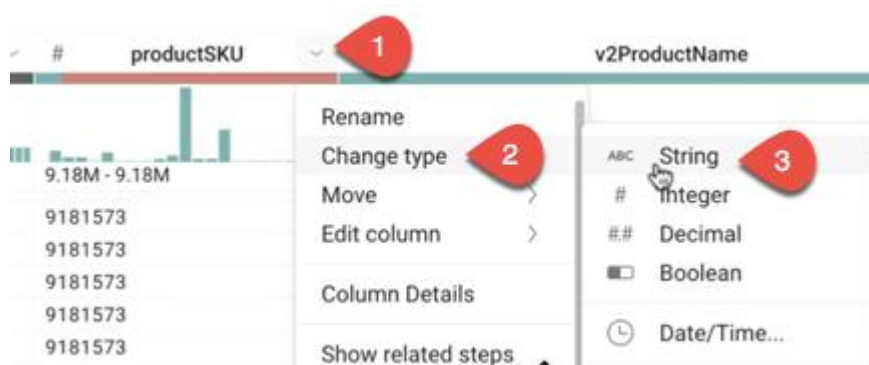
Answer: 8/1/2017 (one day of data)

- Why is there a red bar under the productSKU column?

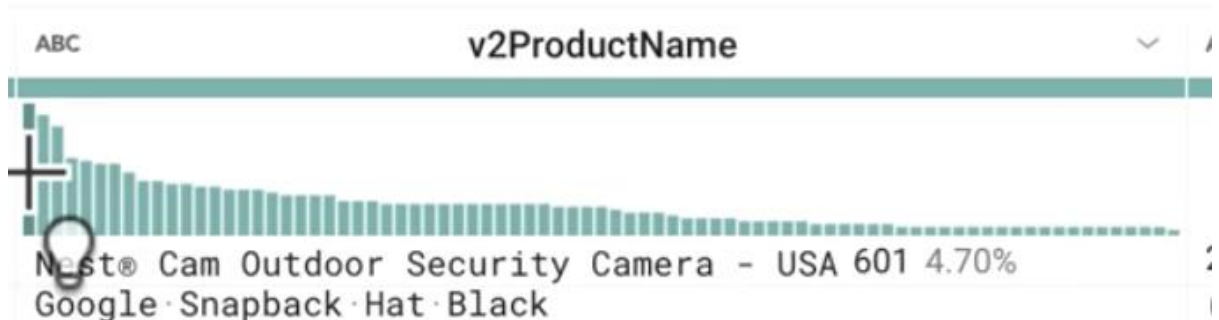
Answer: The red bar indicates mismatched values. Cloud Dataprep automatically identified the productSKU column type as an integer. Cloud Dataprep also detected some non-integer values and therefore flagged those as mismatched. In fact, the productSKU is not always an integer (for example, a correct value might be "GGOEGOCD078399"). So in this case, Cloud Dataprep incorrectly identified the column type: it should be a string, not an integer. You fix that in the next step.

Note: If the productSKU column already has a type `String` then you can't see the red bar.

- To convert the productSKU column type to a string data type, open the menu to the right of the productSKU column by clicking , then click Change type > String.

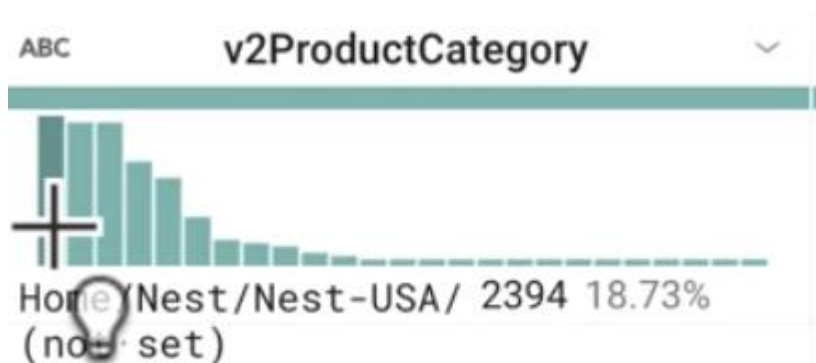


- Looking at v2ProductName, what are the most popular products?



Answer: **Nest products**

- Looking at v2ProductCategory, what are some of the most popular products? How many categories were sampled?



Answer: **Nest, (not set), and Apparel are the most popular out of approximately 25 categories.**

- True or False: The most common productVariant is COLOR.

Answer: **False. It's (not set) because most products do not have variants (80%+)**

- What are the two categories of type?

Answer: **PAGE and EVENT**

- What is the average productQuantity?

Answer: **3.45 (your answer may vary)**

- How many distinct SKUs are in the dataset?

Answer: **Over 600+**

- What are some of the most popular product names by row count? The most popular categories?

Answer:

Cam Outdoor Security Camera - USA

Cam Indoor Security Camera - USA

Learning Thermostat 3rd Gen-USA - Stainless Steel

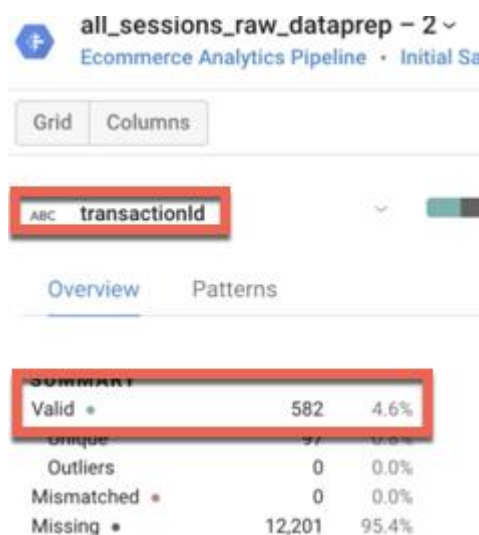
- What is the dominant currency code for transactions?

Answer: USD (United States Dollar)

- Are there valid values for itemQuantity or itemRevenue?

Answer: No, they are all NULL values.

- What percentage of transaction IDs have a valid value? What does this represent for our ecommerce dataset?



The screenshot shows a data pipeline interface for 'all_sessions_raw_dataprep'. The 'transactionId' column is selected. The 'Overview' tab is active, displaying a summary table. The table has three columns: 'Valid', 'Count', and 'Percentage'. The 'Valid' row shows 582 valid values, which is 4.6% of the total. The 'Missing' row shows 12,201 missing values, which is 95.4% of the total. The 'Outliers' and 'Mismatched' rows both show 0 values, representing 0.0%.

Valid *	Count	Percentage
Valid *	582	4.6%
Unique	97	0.8%
Outliers	0	0.0%
Mismatched *	0	0.0%
Missing *	12,201	95.4%

Answer: About 4.6% of transaction IDs have a valid value, which represents the average conversion rate of the website (4.6% of visitors transact).

- How many eCommerceAction_type are there, and what is the most popular eCommerceAction_step?

Answers:

Six types have data in our sample.

0 or NULL is the most popular.

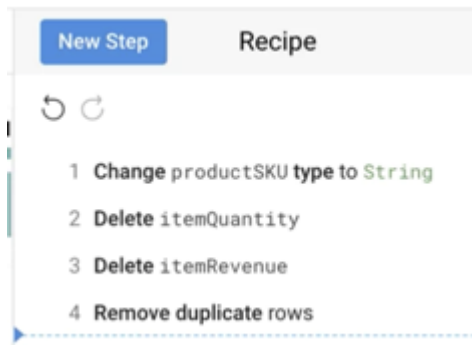
Task 5. Clean the data

In this task, you clean the data by deleting unused columns, eliminating duplicates, creating calculated fields, and filtering the rows. Deleting columns is common for when fields are depreciated in the schema or have all NULL values.

Delete unused columns

- Select the unwanted column, and then click Delete. Do this for the following columns which have all NULL values:

1. Click Recipe (☰) in the top right and select New Step.
2. In the Transformation search box, type deduplicate and select Remove duplicate rows.
3. Click Add.
4. Review the recipe created so far:



Filter out sessions without revenue

Your team has asked you to create a table of all user sessions that bought at least one item from the website. Filter out user sessions with NULL revenue.

1. Under the totalTransactionRevenue column, click the missing values bar.
2. In the Suggestions panel, click Delete rows with missing values, and then click Add (as shown).

The screenshot displays a data table with columns: country, city, totalTransactionRevenue, transactions, and timeOnSite. The 'totalTransactionRevenue' column has a missing values bar. The 'Suggestions' panel on the right provides three options: 'Delete rows' (where ISMISSING([totalTransactionRevenue])), 'Keep rows' (where ISMISSING([totalTransactionRevenue])), and 'Create a new column' (ISMISSING([totalTransactionRevenue])). The 'Delete rows' option is selected, and the 'Add' button is highlighted.

This step filters your dataset to only include transactions with revenue (where totalTransactionRevenue is NULL).

Filter out sessions for just Type = 'PAGE'

The dataset contains both views of website Pages and triggered Events like “viewed product categories” or “added to cart”. To avoid double counting session pageviews, add a filter to only include pageview-related events.

1. In the type column, click the bar for PAGE.
2. In the Suggestions panel, click Keep rows where type is PAGE, and then click Add.

Filter for apparel products

Your team has now asked you to further filter your output to only include transactions in the Apparel category (apparel includes items like T-Shirts and other clothing items)

1. Next to the v2ProductCategory column, click the drop down icon.
2. Select Filter rows > On column values.
3. Select Contains.
4. In Pattern to match type 'Apparel' (case sensitive) and then click Add.

Note: Products in the catalog can belong to more than one category ('Apparel' and 'Home/Apparel/') which is why we are matching any rows that have Apparel anywhere in the category name.

The screenshot displays a data analytics pipeline interface. At the top, it shows 'ECOMMERCE ANALYTICS PIPELINE' and a dataset named 'all_sessions_raw_dataprep'. Below this is a toolbar with various icons for data manipulation. The main area features a table with columns: v2ProductName, v2ProductCategory, productVariant, and currency. The table is filtered to show 177 categories. To the right of the table, a 'Filter rows' panel is open, showing a configuration for filtering on the 'v2ProductCategory' column. The condition is set to 'Contains' and the pattern to match is 'Apparel'. The action is set to 'Keep matching rows'.

v2ProductName	v2ProductCategory	productVariant	currency
Google Snapback Hat Black	(not set)	(not set)	USD
Google Snapback Hat Black	(not set)	(not set)	USD
Google Snapback Hat Black	(not set)	(not set)	USD
Nestle Learning Thermostat 3rd Gen-USA - Copper	Home/Nest/Nest-USA/	(not set)	USD
Nestle Cam Outdoor Security Camera - USA	Home/Nest/Nest-USA/	(not set)	USD
Nestle Learning Thermostat 3rd Gen-USA - Stainless Steel	Home/Nest/Nest-USA/	(not set)	USD
Nestle Learning Thermostat 3rd Gen-USA - Stainless Steel	Home/Nest/Nest-USA/	(not set)	USD
Nestle Cam Indoor Security Camera - USA	Home/Nest/Nest-USA/	(not set)	USD
Nestle Learning Thermostat 3rd Gen-USA - White	Home/Nest/Nest-USA/	(not set)	USD
Nestle Cam Indoor Security Camera - USA	Home/Nest/Nest-USA/	(not set)	USD
Nestle Cam Indoor Security Camera - USA	Home/Nest/Nest-USA/	(not set)	USD
Nestle Protect Smoke + CO White Wired Alarm-USA	Home/Nest/Nest-USA/	(not set)	USD
Nestle Protect Smoke + CO White Battery Alarm-USA	Home/Nest/Nest-USA/	(not set)	USD
Nestle Learning Thermostat 3rd Gen-USA - Stainless Steel	Home/Nest/Nest-USA/	(not set)	USD

Filter rows

Condition: Contains

Filter rows that contain a specified value or pattern

Column: v2ProductCategory

Pattern to match: 'Apparel'

Action: Keep matching rows

Task 6. Enrich the data

To learn about the schema used in this lab, refer to [\[UA\] BigQuery Export schema](#). Search this article for visitId and read the description to determine if it is unique across all user sessions or just the user.

VisitId = An identifier for this session. This is part of the value usually stored as the _utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.

visitId is not unique across all users.

In this task, you add a new concatenated column to create a unique session ID field. Then you will enrich your ecommerce label data with a case statement.

Create a new column for a unique session ID

As you discovered, the dataset has no single column for a unique visitor session. Create a unique ID for each session by concatenating the fullVisitorId and visitId fields.

1. Click on New Step.
2. For Search transformation, type concat, and then select Merge columns.
3. For Columns, select fullVisitorId and visitId.
4. For the New column name, type unique_session_id, and leave the other inputs as their default values and click Add.

Create a case statement for the ecommerce action type

The eCommerceAction_type field is an integer that maps to actual ecommerce actions performed in that session like 3 = "Add to Cart" or 5 = "Check out." Create a calculated column that maps to the integer value.

1. Click on New Step.
2. In the Transformation panel, type case, and then select Conditional column.

3. Select Case on single column from the drop-down.
4. For Column to evaluate, specify eCommerceAction_type.
5. Next to Cases (X), click Add 8 times for a total of 9 cases.
6. For each Case, specify the following mapping values (including the quotes):

Value to compare	New value
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'
0	'Unknown'

Leave the other fields at their default values.

7. For New column name, type eCommerceAction_label, and then click Add.

8. Review the Recipe and compare it to this example:

New Step

Recipe

×

↶ ↷

ooo

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

Keep rows

8

Concatenate fullVisitorId, visitId

9

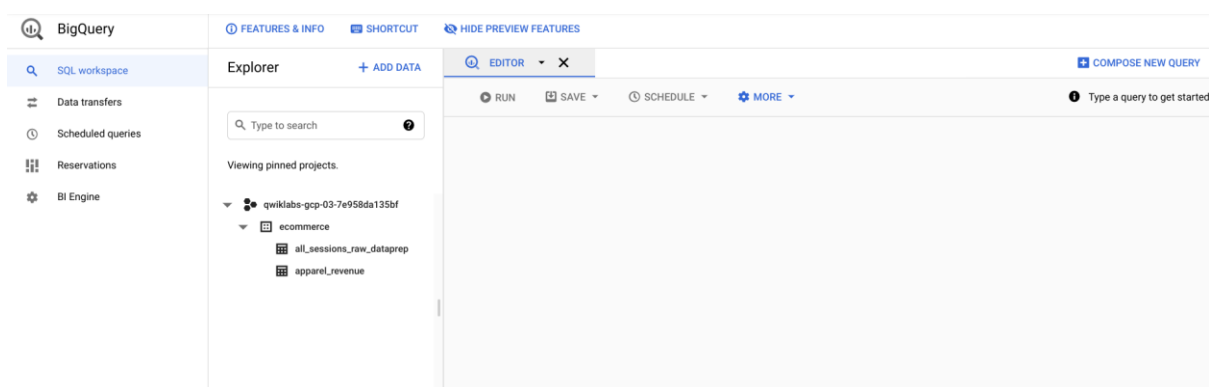
Create eCommerceAction_label from 9
case conditions on eCommerceAction_type

Task 7. Run Cloud Dataprep job to load BigQuery

When you are satisfied with the flow, it's time to execute the transformation recipe against your source dataset. To do that, you execute and monitor a Cloud Dataprep job (which starts and runs a Cloud Dataflow job).

1. From the Transformer page, in the upper right, click Run.
2. In the Publishing Actions section, hover over on `Create-CSV` then click Edit.
3. Select BigQuery in the left panel and go into your ecommerce dataset, and then click Create a new table.
4. Name the output table `apparel_revenue` and select Drop the table every run in the right panel.
5. Click Update.
6. Click Run.
7. Click Job history in the left panel to monitor your Cloud Dataprep job.
8. Wait 1 - 2 minutes for your job to run

After your Cloud Dataprep job finishes, refresh your BigQuery page and confirm that the output table `apparel_revenue` exists.



Select `apparel_revenue` > Preview and ensure you have revenue transaction data for Apparel products.

BigQuery

SQL workspace

Data transfers

Scheduled queries

Reservations

BI Engine

Release Notes

FEATURES & INFO

SHORTCUT

HIDE PREVIEW FEATURES

Explorer

+ ADD DATA

Type to search

Viewing pinned projects.

qwilibs-gcp-03-7e958da135bf

e-commerce

all_sessions_raw_dataprep

apparel_revenue

EDITOR

APPARE...

COMPOSE NEW QUERY

apparel_revenue

QUERY TABLE

SHARE TABLE

COPY TABLE

DELETE TABLE

EXPORT

Schema

Details

Preview

Row	fullVisitorId	channelGrouping	time	country	city	totalTransactionRevenue	transactions	timeOnSite	page
1	4293484339755189100	Display	55993	United States	Mountain View	37290000	1	56	
2	4293484339755189100	Display	55993	United States	Mountain View	37290000	1	56	
3	4293484339755189100	Display	55994	United States	Mountain View	37290000	1	56	
4	4293484339755189100	Display	55994	United States	Mountain View	37290000	1	56	
5	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
6	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
7	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
8	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
9	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
10	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
11	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
12	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
13	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
14	5408509515083537446	Referral	2889976	United States	not available in demo dataset	14190000	1	3682	
15	00122544416887869	Direct	734490	United States	not available in demo dataset	87540000	1	996	

Rows per page: 100 1 - 100 of 2503

First page

Last page

JOB HISTORY

QUERY HISTORY

SAVED QUERIES

Congratulations!

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

Already have a Google Analytics account and want to query your own datasets in BigQuery? Follow this [Set up BigQuery Export guide](#).

End your lab

When you have completed your lab, click End Lab. Google Cloud Skills Boost removes the resources you've used and cleans the account for you.

You will be given an opportunity to rate the lab experience. Select the applicable number of stars, type a comment, and then click Submit.

The number of stars indicates the following:

- **1 star = Very dissatisfied**
- **2 stars = Dissatisfied**
- **3 stars = Neutral**
- **4 stars = Satisfied**
- **5 stars = Very satisfied**

You can close the dialog box if you don't want to provide feedback.

For feedback, suggestions, or corrections, please use the Support tab.

Copyright 2022 Google LLC All rights reserved. Google and the Google logo are trademarks of Google LLC. All other company and product names may be trademarks of the respective companies with which they are associated.