SEP 720 – Cloud Computing : Assignment 3

Qwiklabs

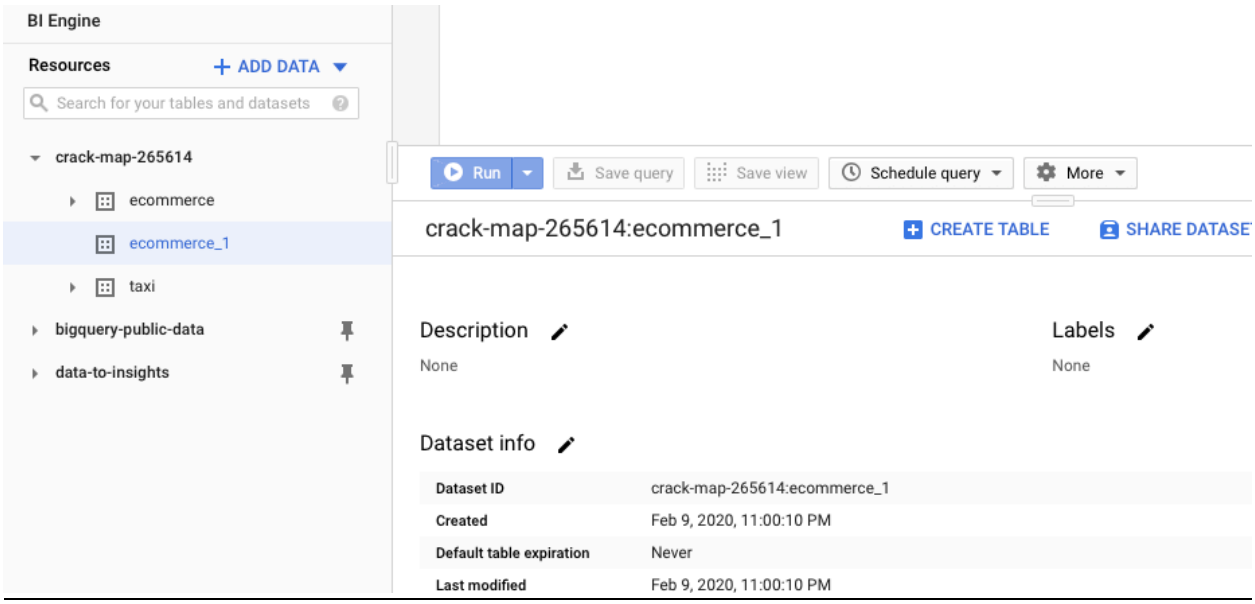
Submitted by,

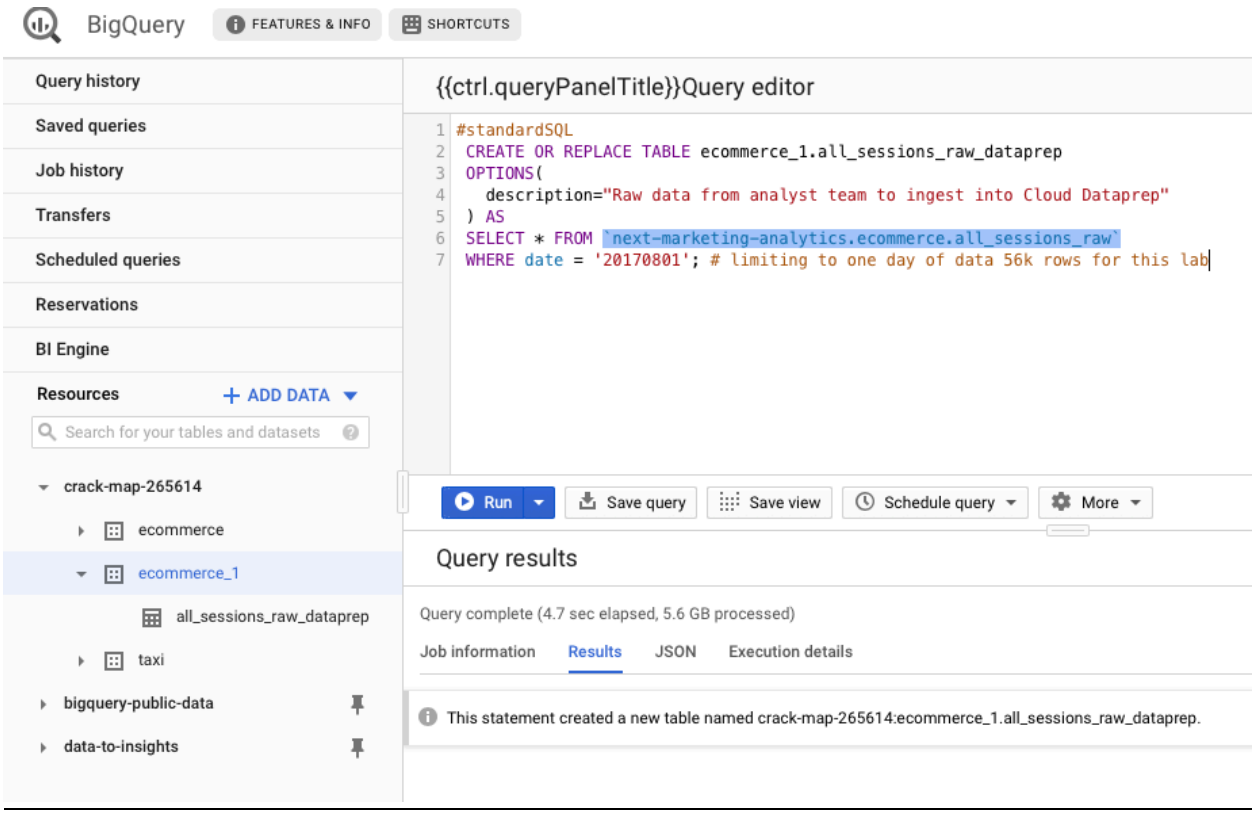
Greeshma Gopal(gopalg)

ID- 400245291

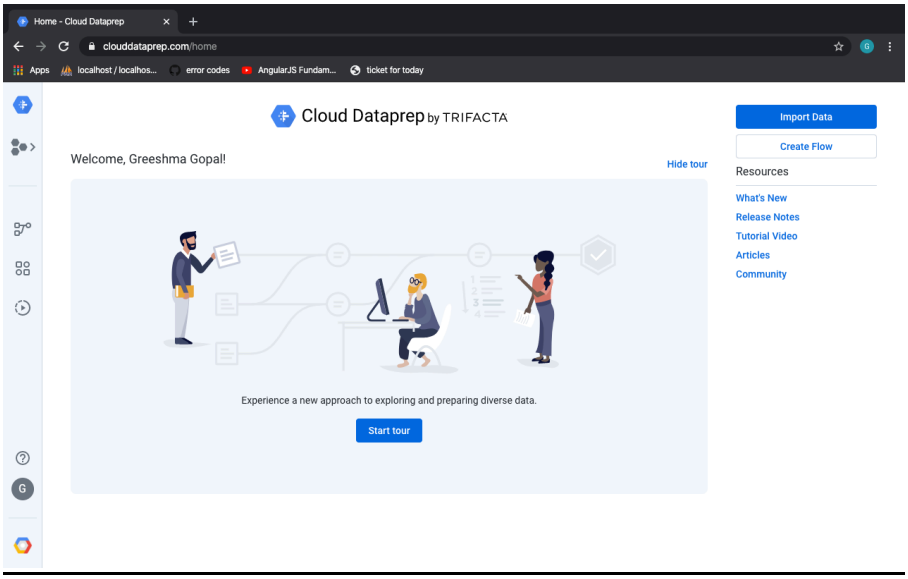
Lab1: Creating a Data Transformation Pipeline with Cloud Dataprep

- Creating a BigQuery Dataset

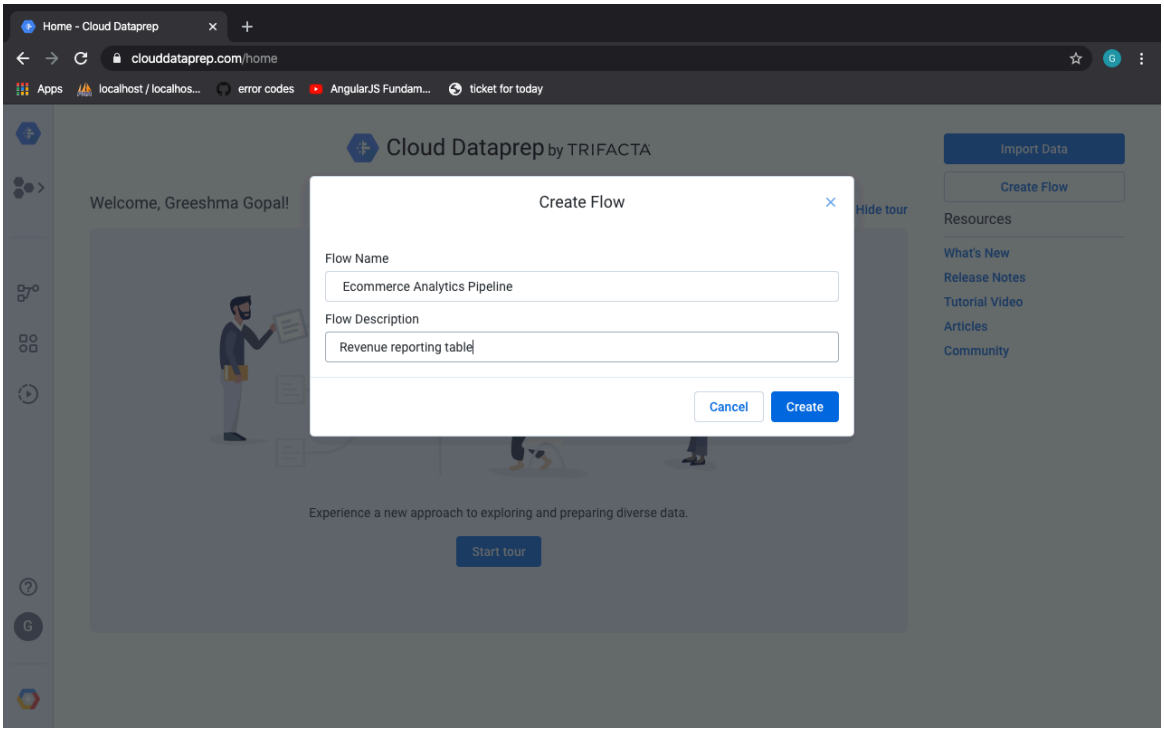


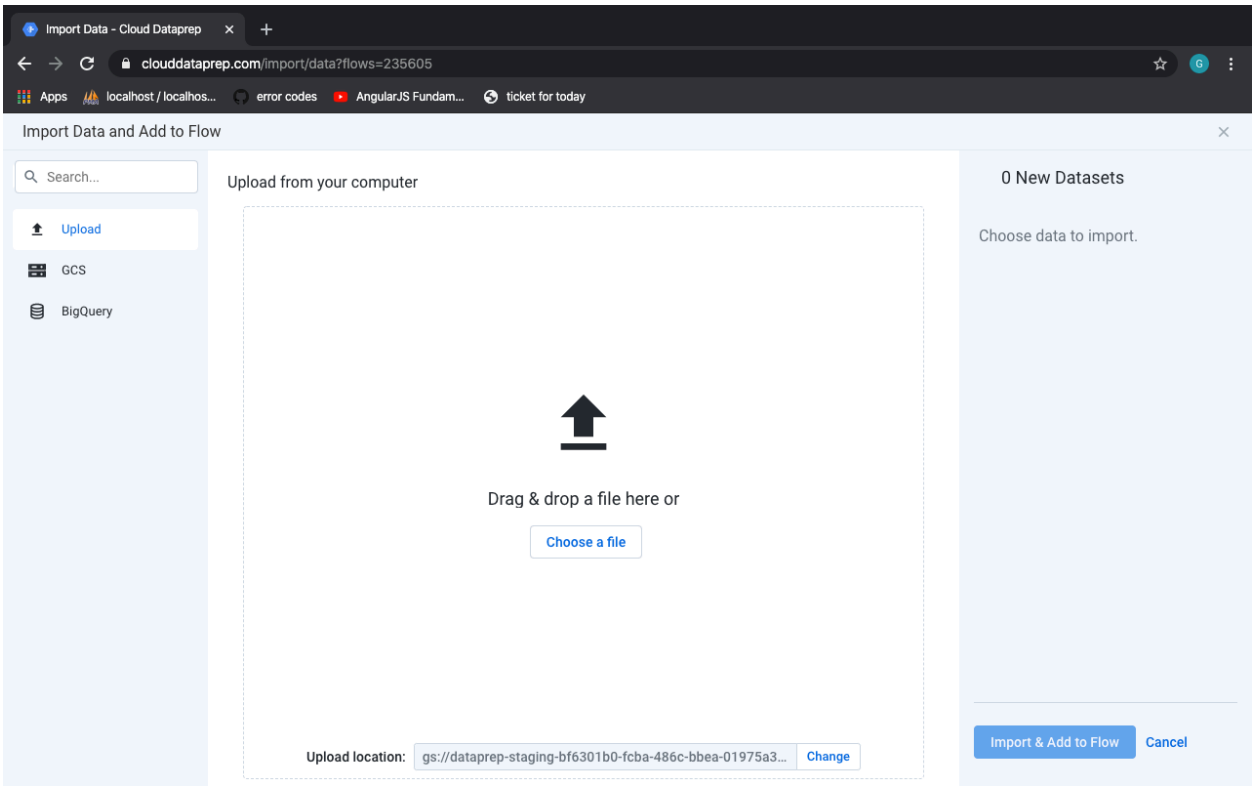


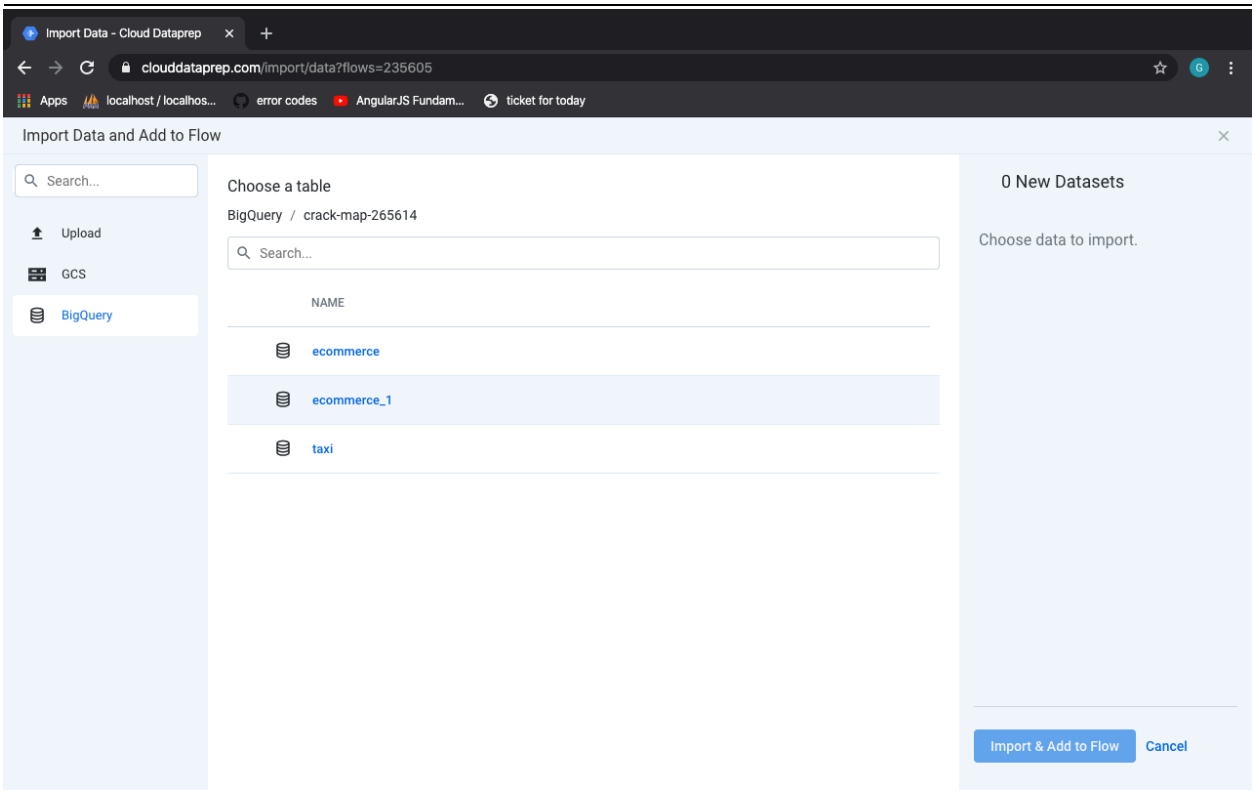
- Opening Cloud Dataprep



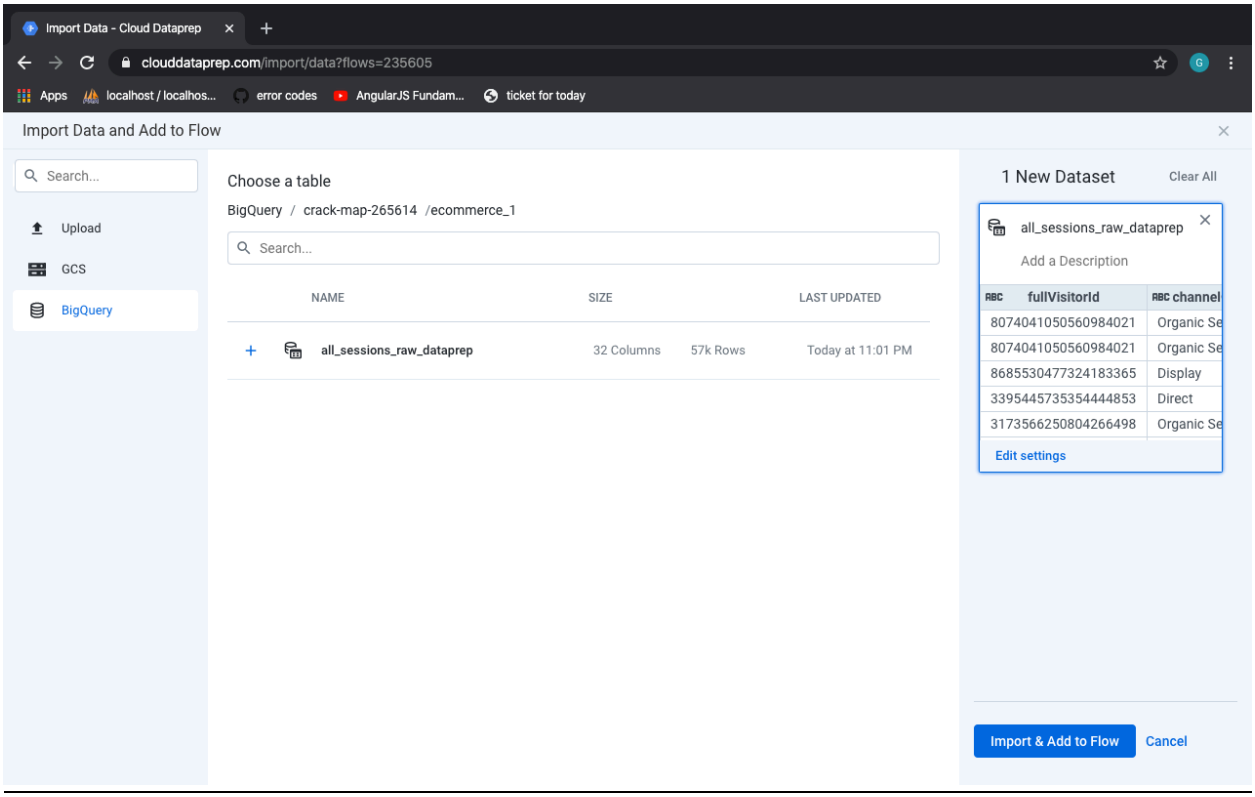
- Connecting BigQuery data to Cloud Dataprep

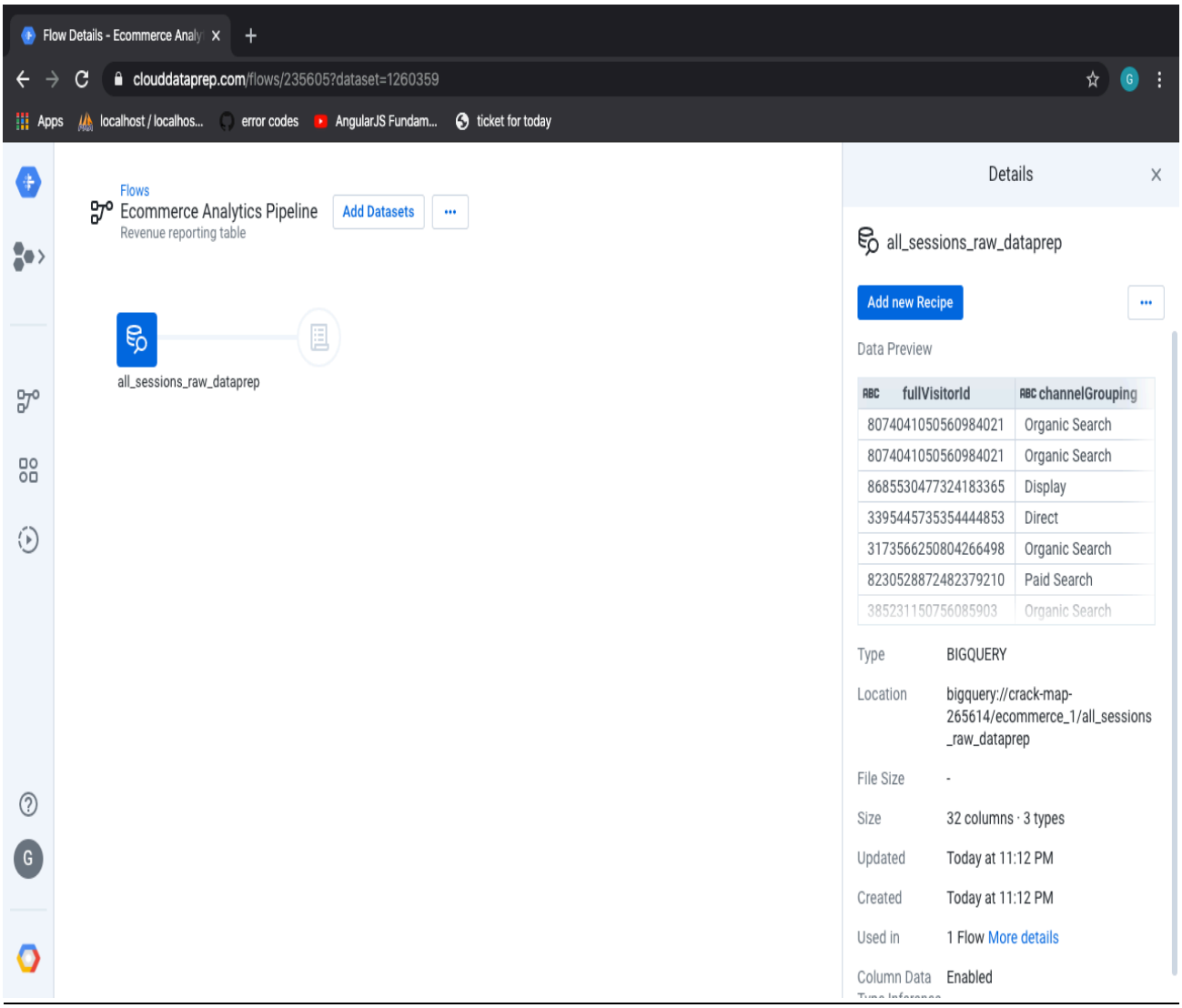


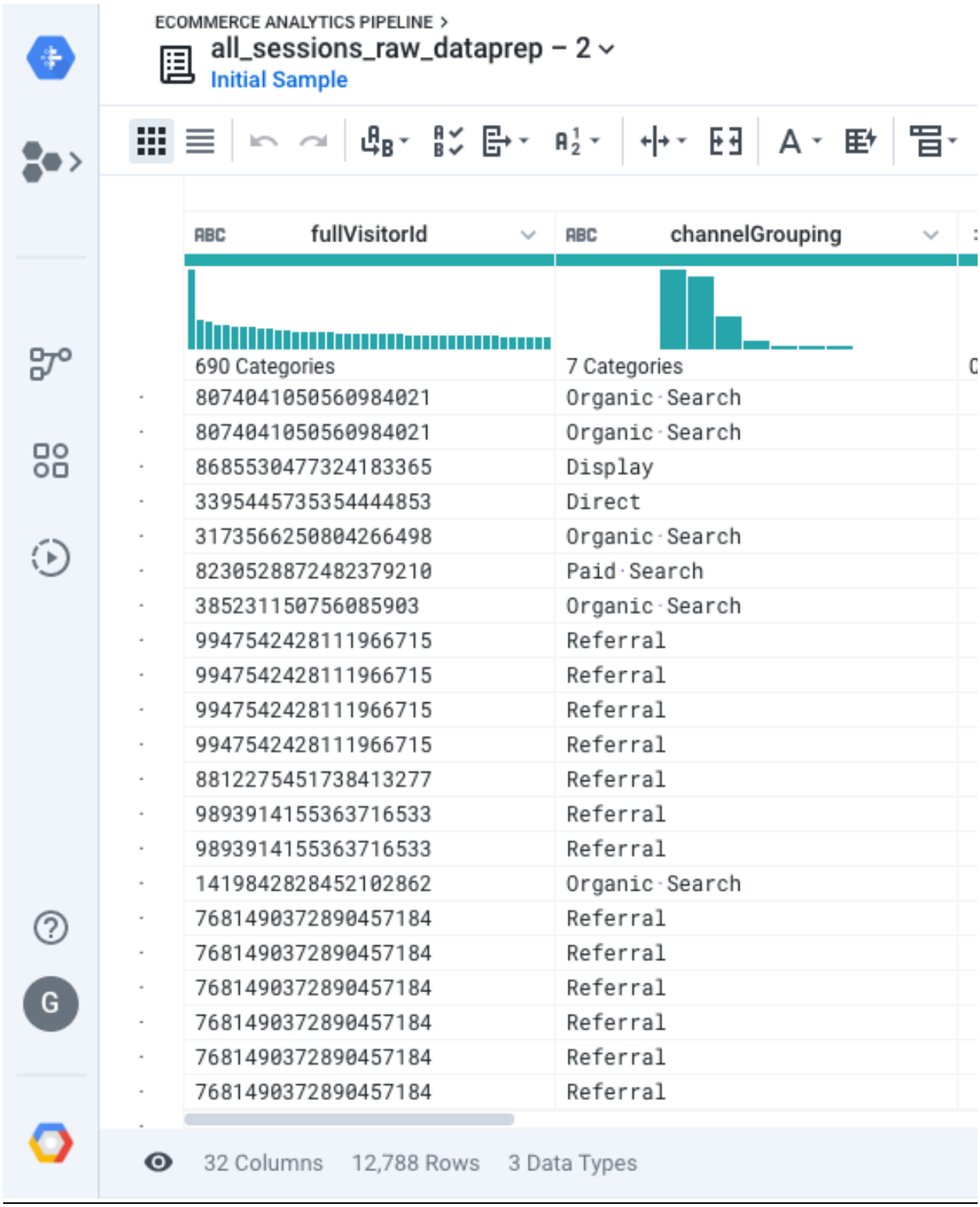

Import Data - Cloud Dataprep
clouddataprep.com/import/data?flows=235605
Apps
localhost / localhos...
error codes
AngularJS Fundam...
ticket for today
Import Data and Add to Flow
Search...
Upload
GCS
BigQuery
Upload from your computer
Drag & drop a file here or
Choose a file
Upload location:
gs://dataprep-staging-bf6301b0-fcba-486c-bbea-01975a3...
Change
0 New Datasets
Choose data to import.
Import & Add to Flow
Cancel

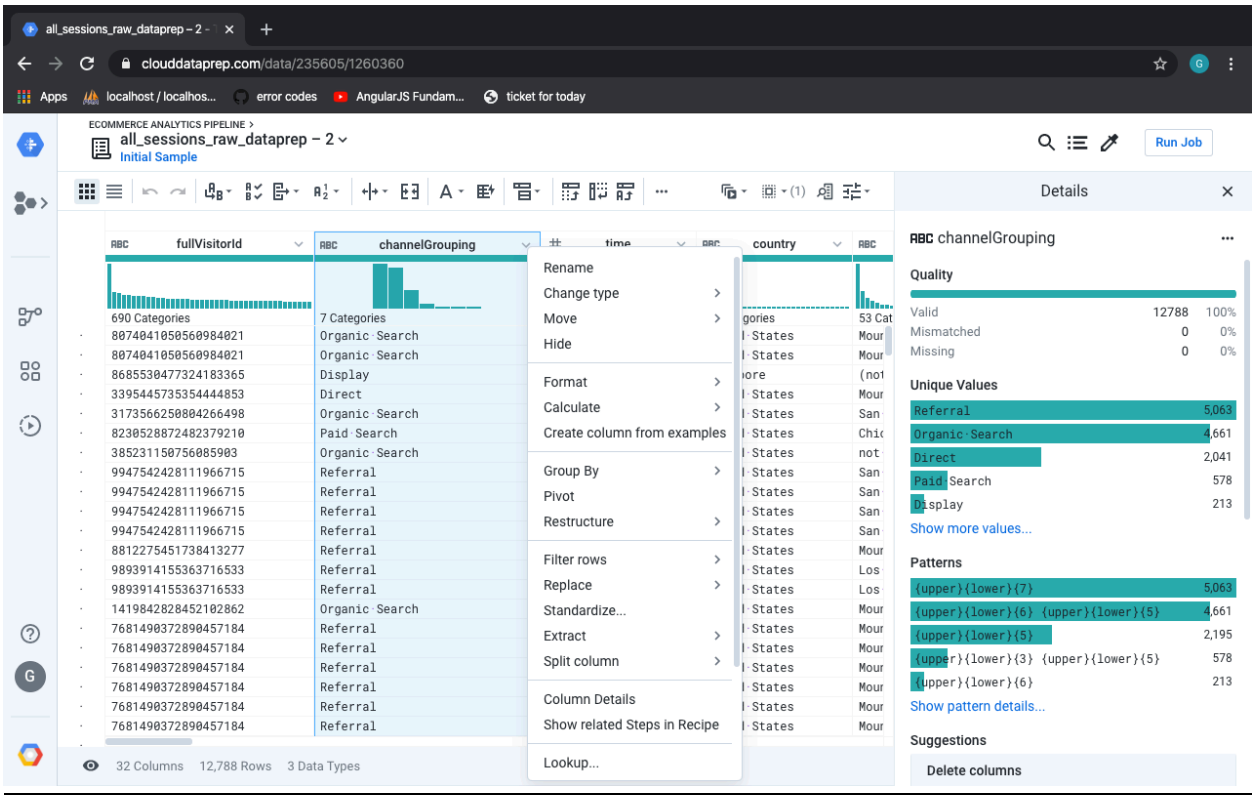

Import Data - Cloud Dataprep
clouddataprep.com/import/data?flows=235605
Apps
localhost / localhos...
error codes
AngularJS Fundam...
ticket for today
Import Data and Add to Flow
Search...
Upload
GCS
BigQuery
Choose a table
BigQuery / crack-map-265614
Search...
NAME
ecommerce
ecommerce_1
taxi
0 New Datasets
Choose data to import.
Import & Add to Flow
Cancel

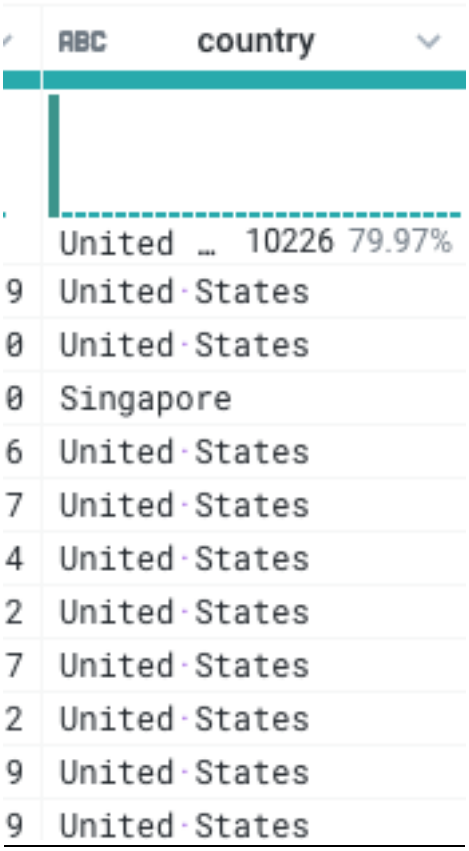
- Exploring ecommerce data fields with a UI

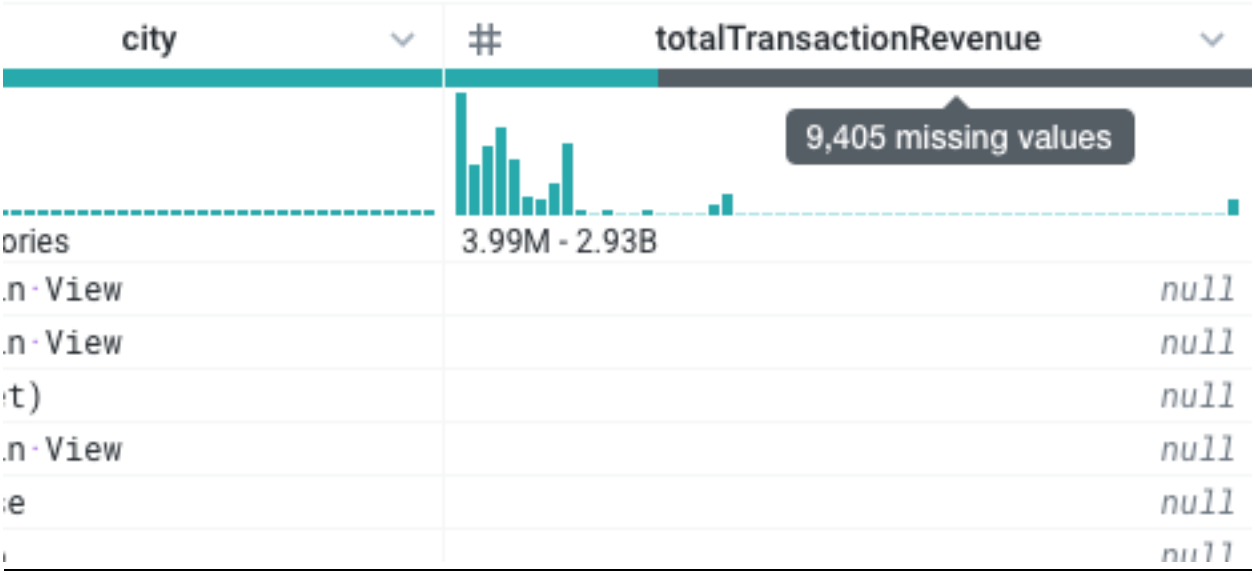


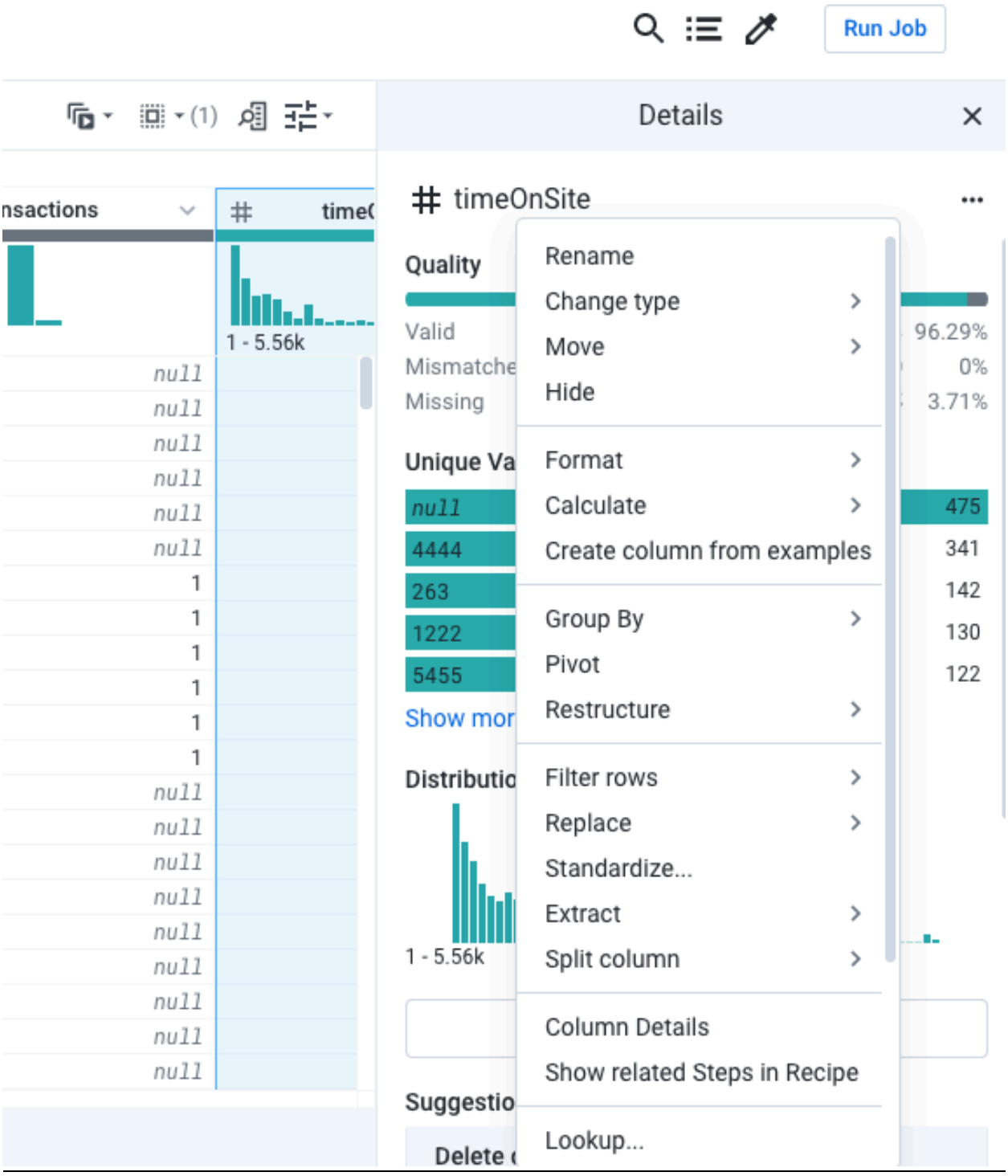

Flow Details - Ecommerce Analy
clouddataprep.com/flows/235605?dataset=1260359
Apps
localhost / localhos...
error codes
AngularJS Fundam...
ticket for today
Flows
Ecommerce Analytics Pipeline
Revenue reporting table
Add Datasets
all_sessions_raw_dataprep
Details
all_sessions_raw_dataprep
Add new Recipe
Data Preview
fullVisitorId
channelGrouping
8074041050560984021
Organic Search
8074041050560984021
Organic Search
8685530477324183365
Display
3395445735354444853
Direct
3173566250804266498
Organic Search
8230528872482379210
Paid Search
38523115075608590 3
Organic Search
Type
BIGQUERY
Location
bigquery://crack-map-265614/ecommerce_1/all_sessions_raw_dataprep
File Size
-
Size
32 columns · 3 types
Updated
Today at 11:12 PM
Created
Today at 11:12 PM
Used in
1 Flow More details
Column Data
Enabled

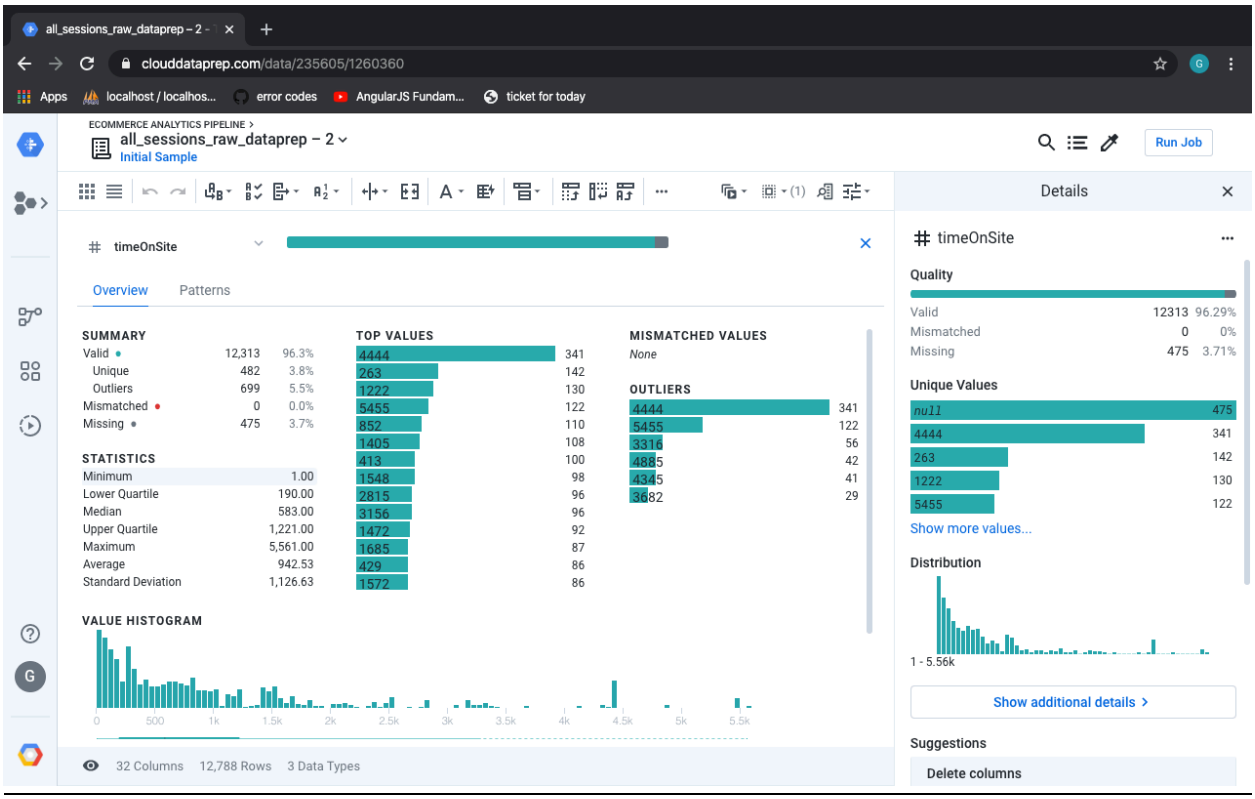

ECOMMERCE ANALYTICS PIPELINE >
all_sessions_raw_dataprep – 2
Initial Sample
ABC fullVisitorId
ABC channelGrouping
690 Categories
7 Categories
8074041050560984021 Organic Search
8074041050560984021 Organic Search
8685530477324183365 Display
3395445735354444853 Direct
3173566250804266498 Organic Search
8230528872482379210 Paid Search
385231150756085903 Organic Search
9947542428111966715 Referral
9947542428111966715 Referral
9947542428111966715 Referral
9947542428111966715 Referral
8812275451738413277 Referral
9893914155363716533 Referral
9893914155363716533 Referral
1419842828452102862 Organic Search
7681490372890457184 Referral
7681490372890457184 Referral
7681490372890457184 Referral
7681490372890457184 Referral
7681490372890457184 Referral
7681490372890457184 Referral
32 Columns
12,788 Rows
3 Data Types

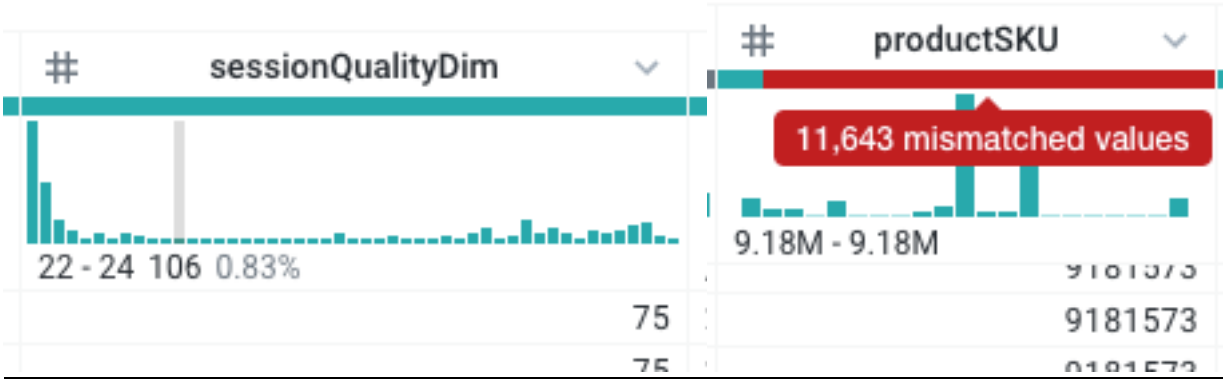

all_sessions_raw_dataprep – 2
clouddataprep.com/data/235605/1260360
Apps
localhost / localhos...
error codes
AngularJS Fundam...
ticket for today
ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep – 2
Initial Sample
Run Job
Details
fullVisitorId
channelGrouping
time
country
690 Categories
7 Categories
Rename
Change type
Move
Hide
Format
Calculate
Create column from examples
Group By
Pivot
Restructure
Filter rows
Replace
Standardize...
Extract
Split column
Column Details
Show related Steps in Recipe
Lookup...
channelGrouping
Quality
Valid 12788 100%
Mismatched 0 0%
Missing 0 0%
Unique Values
Referral 5,063
Organic Search 4,661
Direct 2,041
Paid Search 578
Display 213
Show more values...
Patterns
{upper}{lower}{7} 5,063
{upper}{lower}{6} {upper}{lower}{5} 4,661
{upper}{lower}{5} 2,195
{upper}{lower}{3} {upper}{lower}{5} 578
{upper}{lower}{6} 213
Show pattern details...
Suggestions
Delete columns
32 Columns
12,788 Rows
3 Data Types

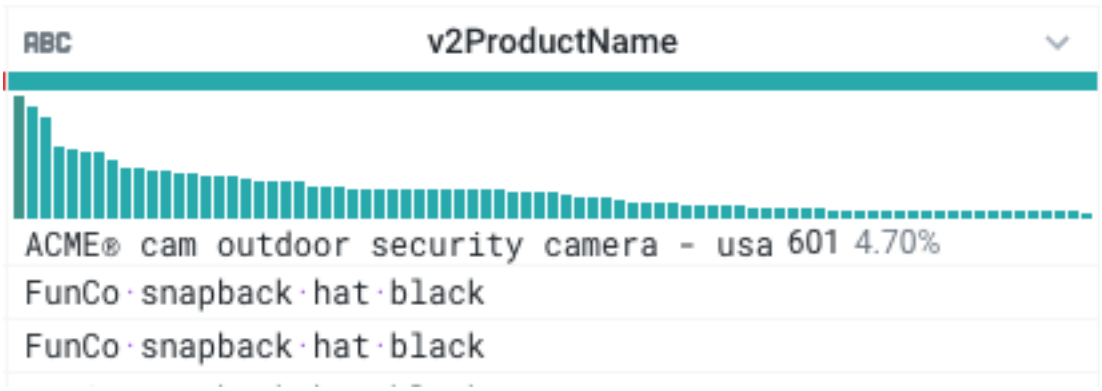

country
United … 10226 79.97%
9 United States
0 United States
0 Singapore
6 United States
7 United States
4 United States
2 United States
7 United States
2 United States
9 United States
9 United States

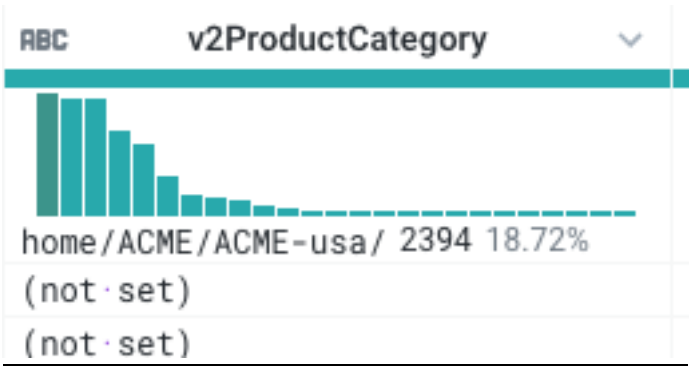

city
totalTransactionRevenue
9,405 missing values
ories
3.99M - 2.93B
.n View
.n View
t)
.n View
e
null
null
null
null
null


Run Job
Details
timeOnSite
Quality
Valid
Mismatche
Missing
96.29%
0%
3.71%
Unique Va
Show mor
Distributio
1 - 5.56k
Rename
Change type
Move
Hide
Format
Calculate
Create column from examples
Group By
Pivot
Restructure
Filter rows
Replace
Standardize...
Extract
Split column
Column Details
Show related Steps in Recipe
Suggestio
Lookup...
Delete
nsactions
1 - 5.56k


all_sessions_raw_dataprep – 2
clouddataprep.com/data/235605/1260360
Apps
localhost / localhos...
error codes
AngularJS Fundam...
ticket for today
ECOMMERCE ANALYTICS PIPELINE
all_sessions_raw_dataprep – 2
Initial Sample
Run Job
timeOnSite
Overview
Patterns
SUMMARY
Valid 12,313 96.3%
Unique 482 3.8%
Outliers 699 5.5%
Mismatched 0 0.0%
Missing 475 3.7%
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
TOP VALUES
4444 341
263 142
1222 130
5455 122
852 110
1405 108
413 100
1548 98
2815 96
3156 96
1472 92
1685 87
429 86
1572 86
MISMATCHED VALUES
None
OUTLIERS
4444 341
5455 122
3316 56
4885 42
4345 41
3682 29
VALUE HISTOGRAM
0 500 1k 1.5k 2k 2.5k 3k 3.5k 4k 4.5k 5k 5.5k
32 Columns 12,788 Rows 3 Data Types
Details
timeOnSite
Quality
Valid 12313 96.29%
Mismatched 0 0%
Missing 475 3.71%
Unique Values
null 475
4444 341
263 142
1222 130
5455 122
Show more values...
Distribution
1 - 5.56k
Show additional details
Suggestions
Delete columns


sessionQualityDim
22 - 24 106 0.83%
75
productSKU
11,643 mismatched values
9.18M - 9.18M
9181573


ABC
v2ProductName
ACME® cam outdoor security camera - usa 601 4.70%
FunCo snapback hat black
FunCo snapback hat black



ABC transactionId

Overview Patterns

SUMMARY

Valid	582	4.6%
Unique	97	0.8%
Outliers	0	0.0%
Mismatched	0	0.0%
Missing	12,206	95.4%

STRING LENGTH STATISTICS

Minimum	15.00
Lower Quartile	15.00
Median	15.00
Upper Quartile	15.00
Maximum	15.00
Average	15.00
Standard Deviation	0.00

TOP VALUES

ORD201708012517	48
ORD201708012584	44
ORD201708012558	22
ORD201708012449	20
ORD201708012572	20
ORD201708012481	18
ORD201708012533	16
ORD201708012550	16
ORD201708012562	14
ORD201708012360	12
ORD201708012415	12
ORD201708012493	12
ORD201708012543	12
ORD201708012473	10

MISMATCHED VALUES

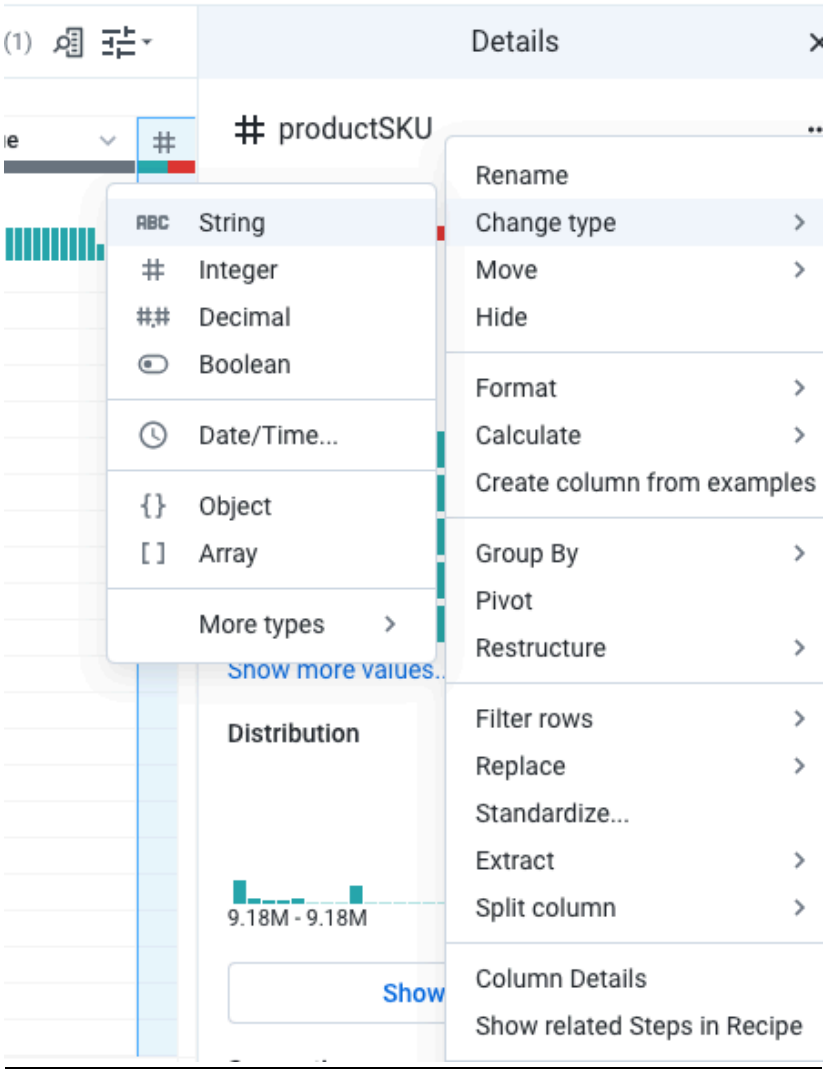
None

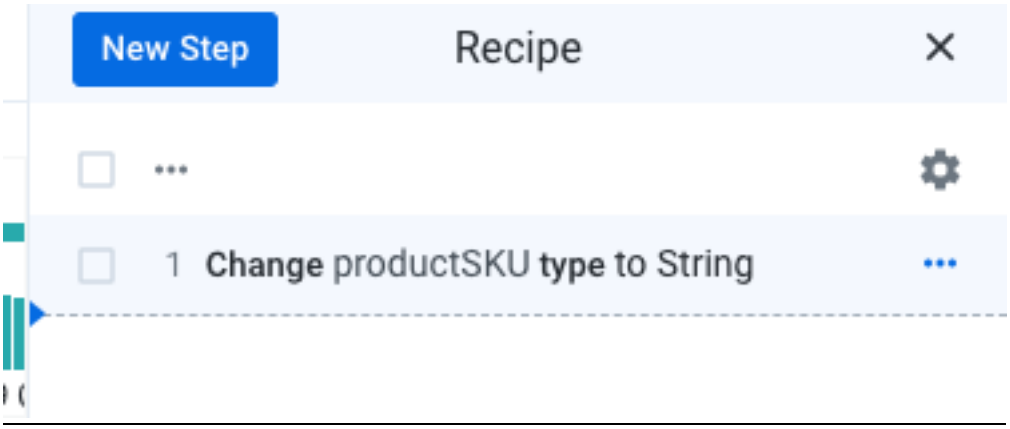
STRING LENGTH OUTLIERS

None

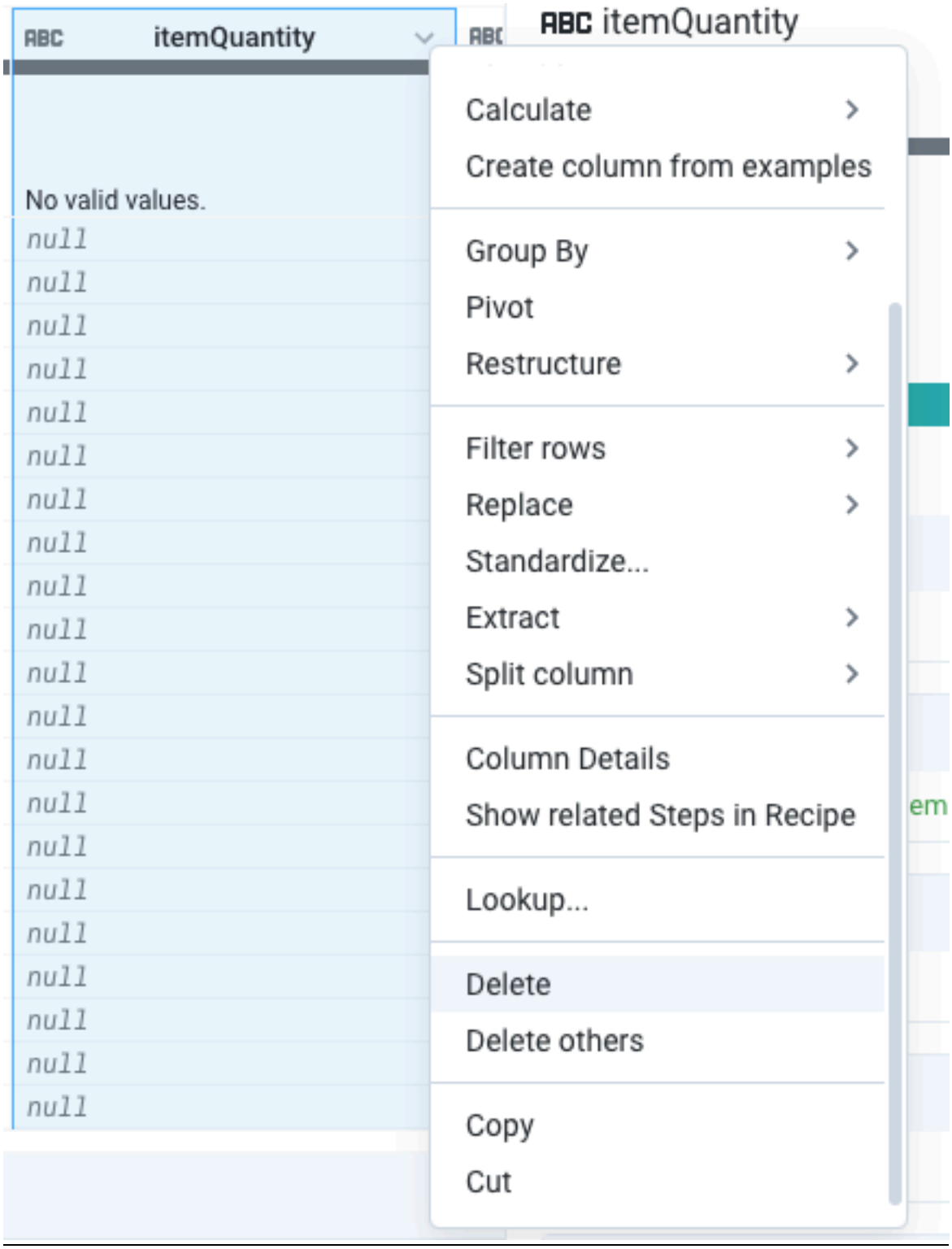
STRING LENGTH

- Cleaning the data

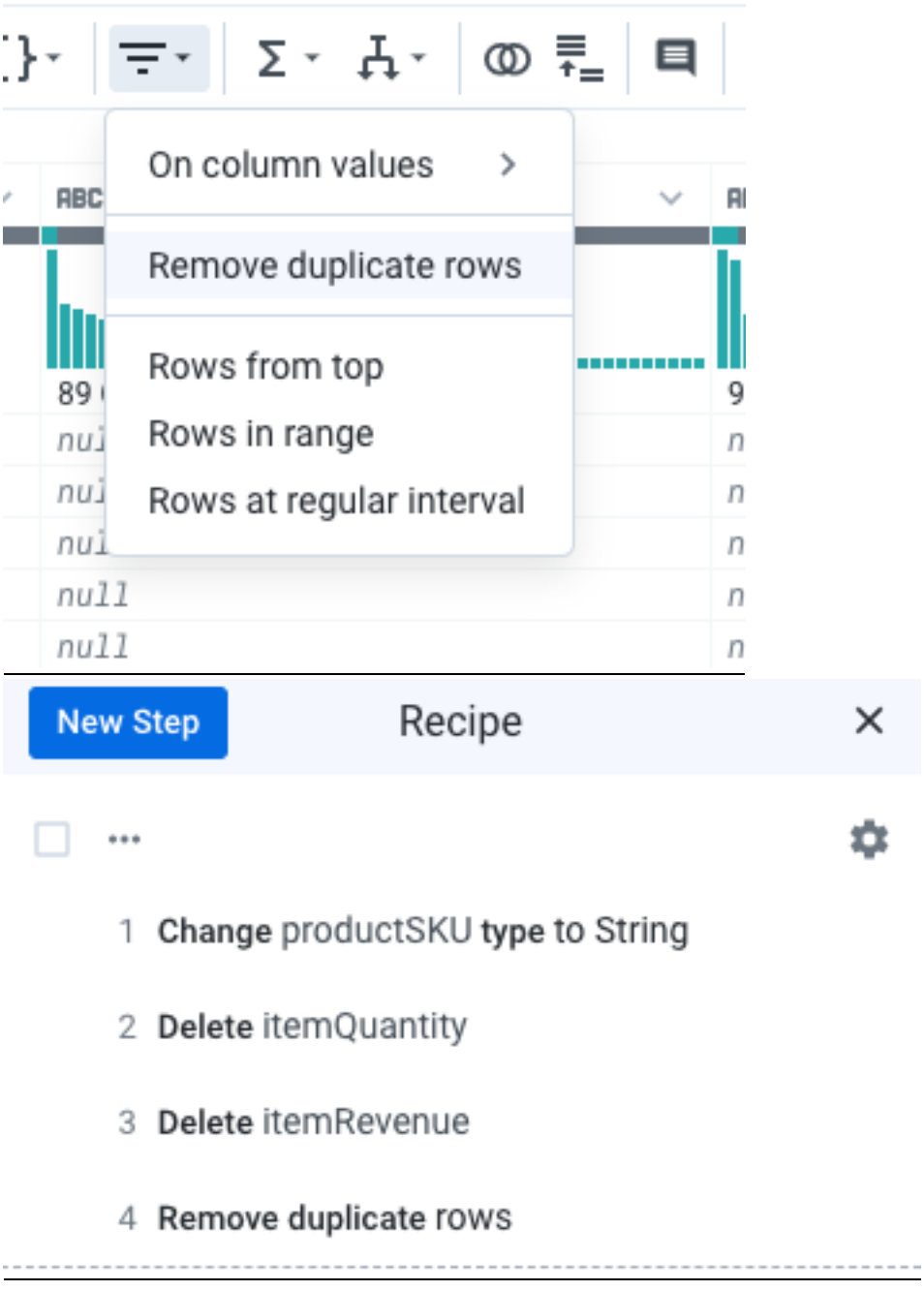




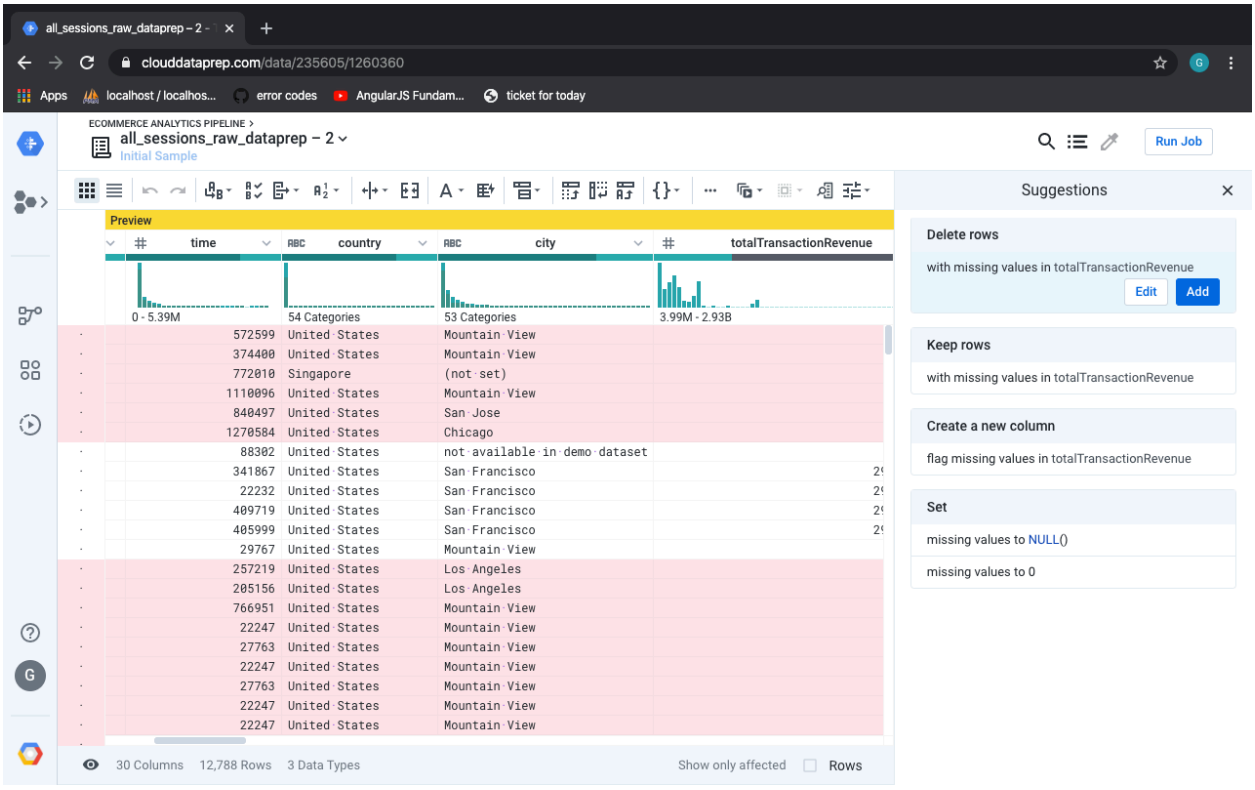
- Deleting unused columns

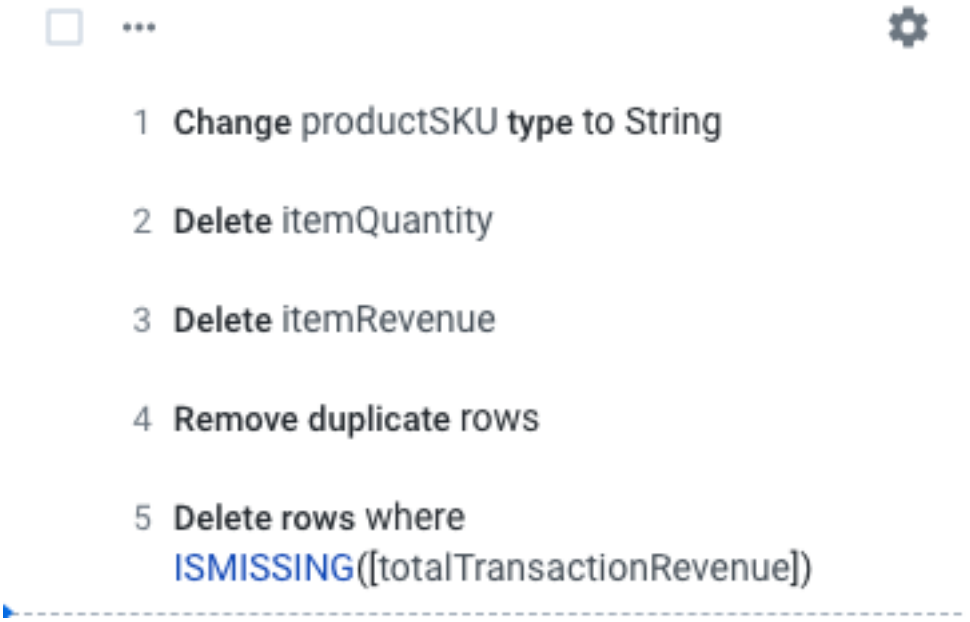


- Deduplicating rows

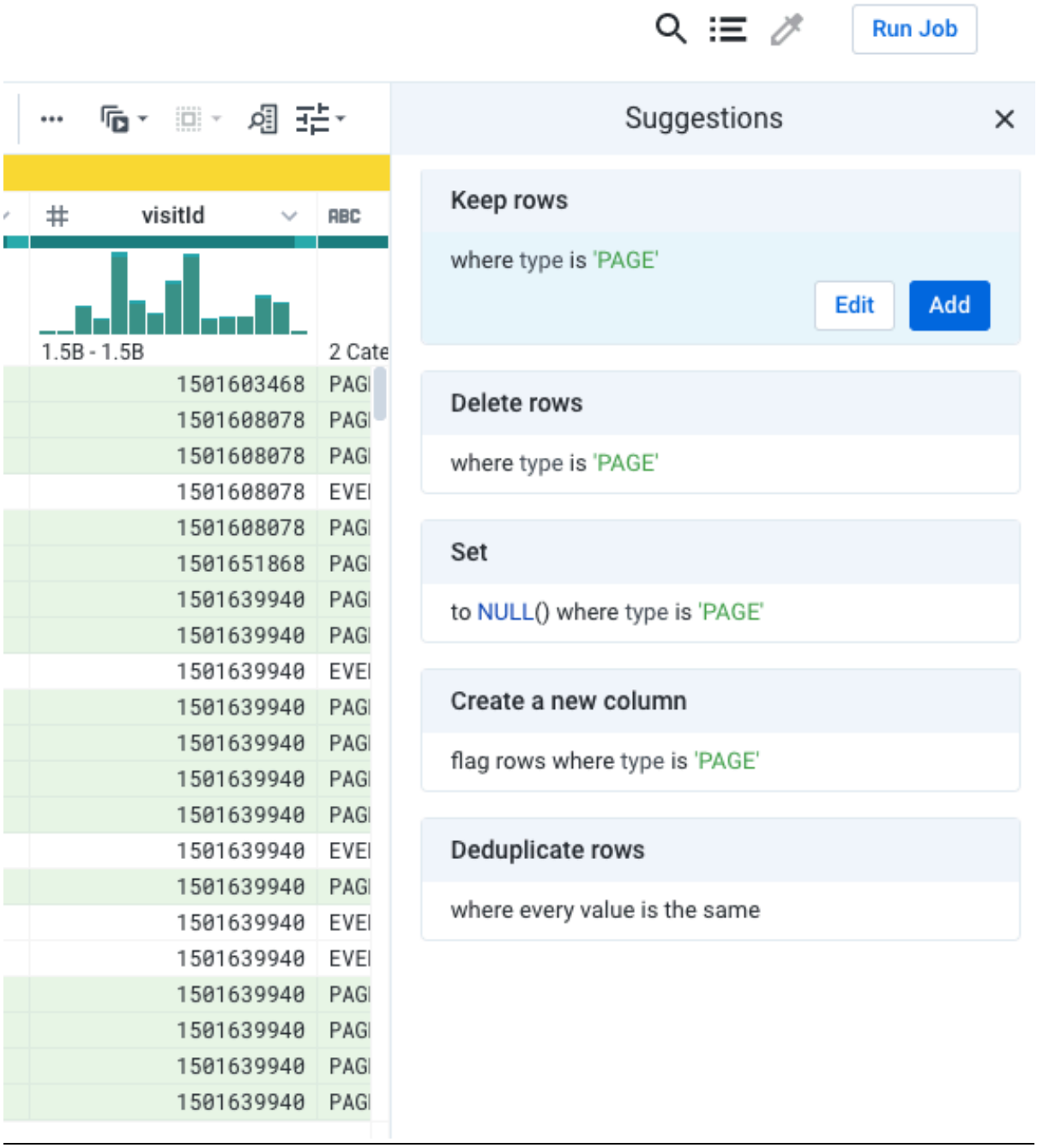


- Filtering out sessions without revenue





- Filtering sessions for PAGE views



1 **Change** productSKU **type** to String

2 **Delete** itemQuantity

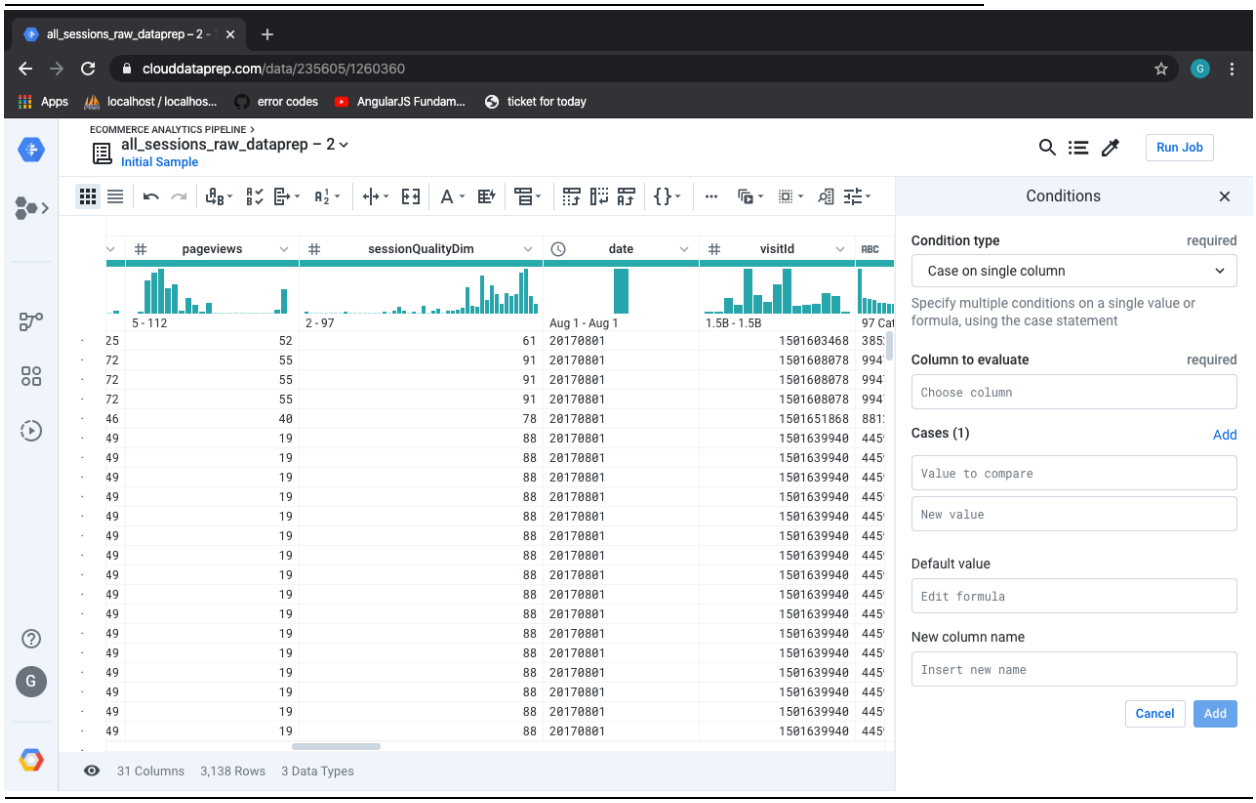
3 **Delete** itemRevenue

4 **Remove duplicate rows**

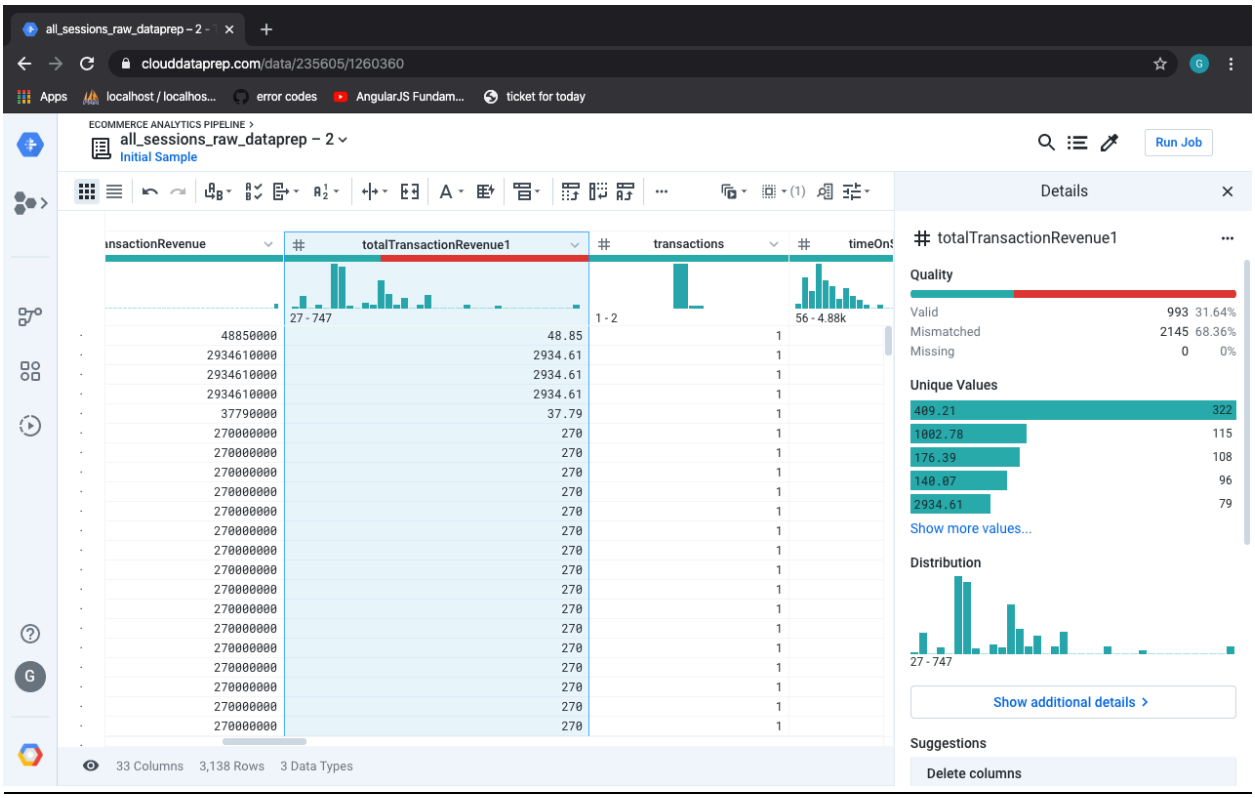
5 **Delete rows where**
ISMISSING([totalTransactionRevenue])

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

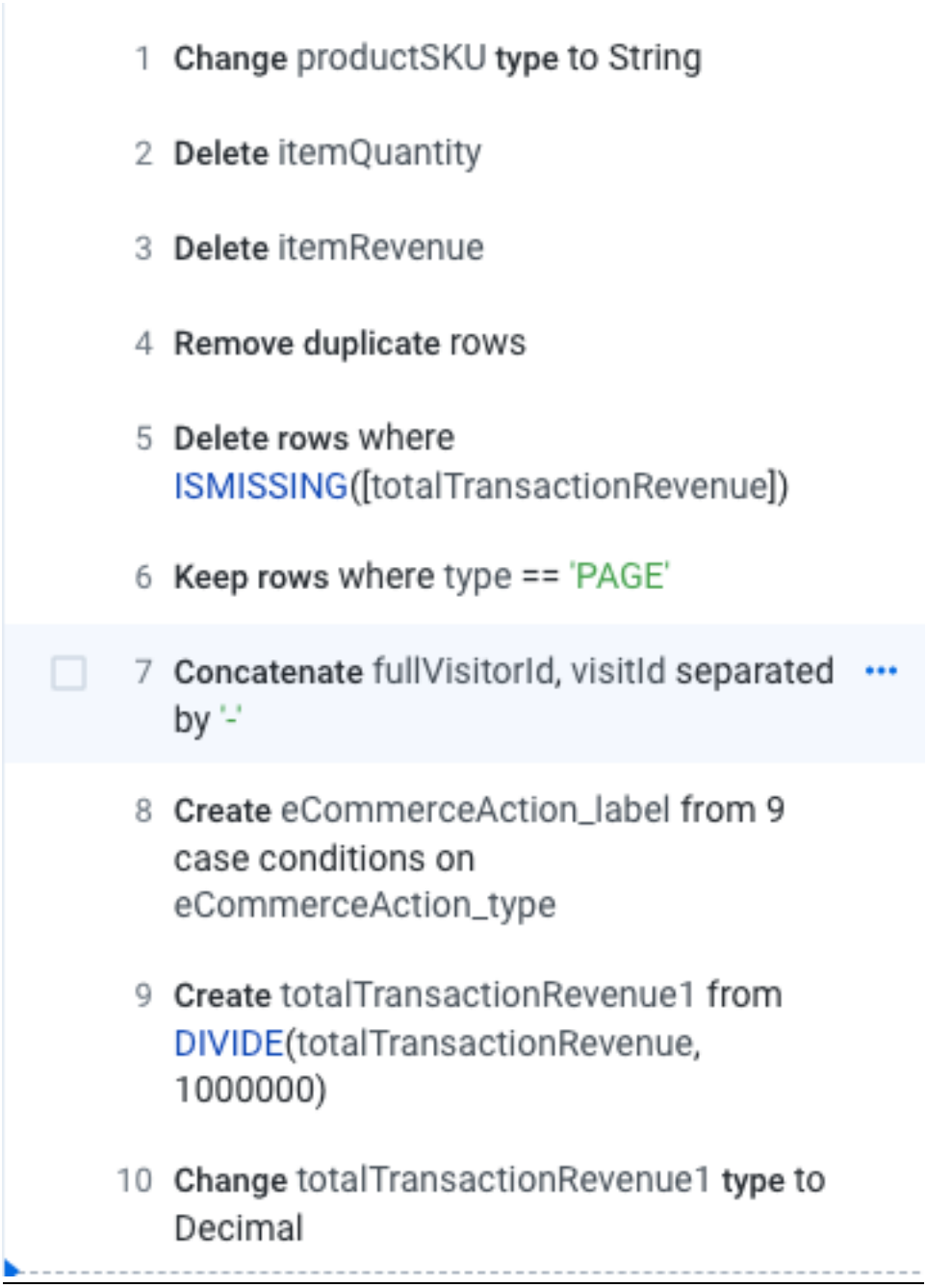
7 **Concatenate** fullVisitorId, visitId **separated by** '-'



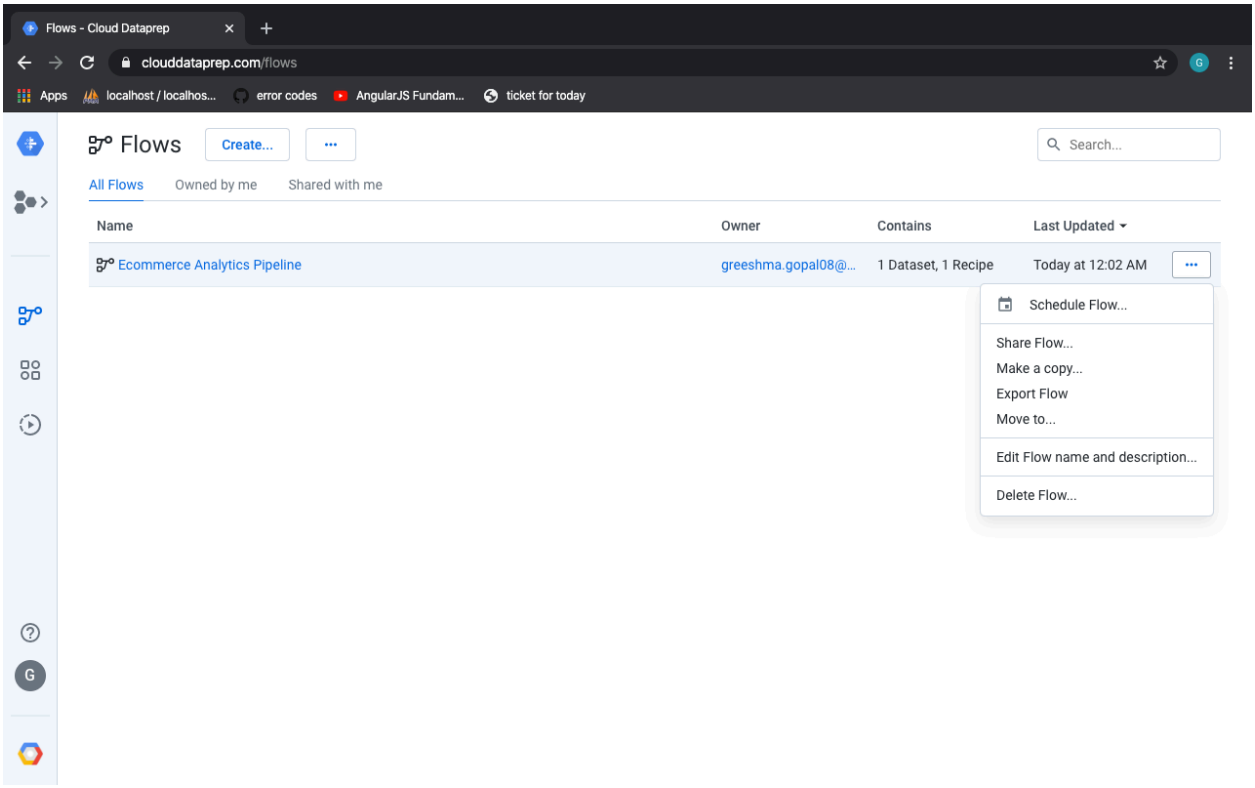
- Creating a new column for a unique session ID

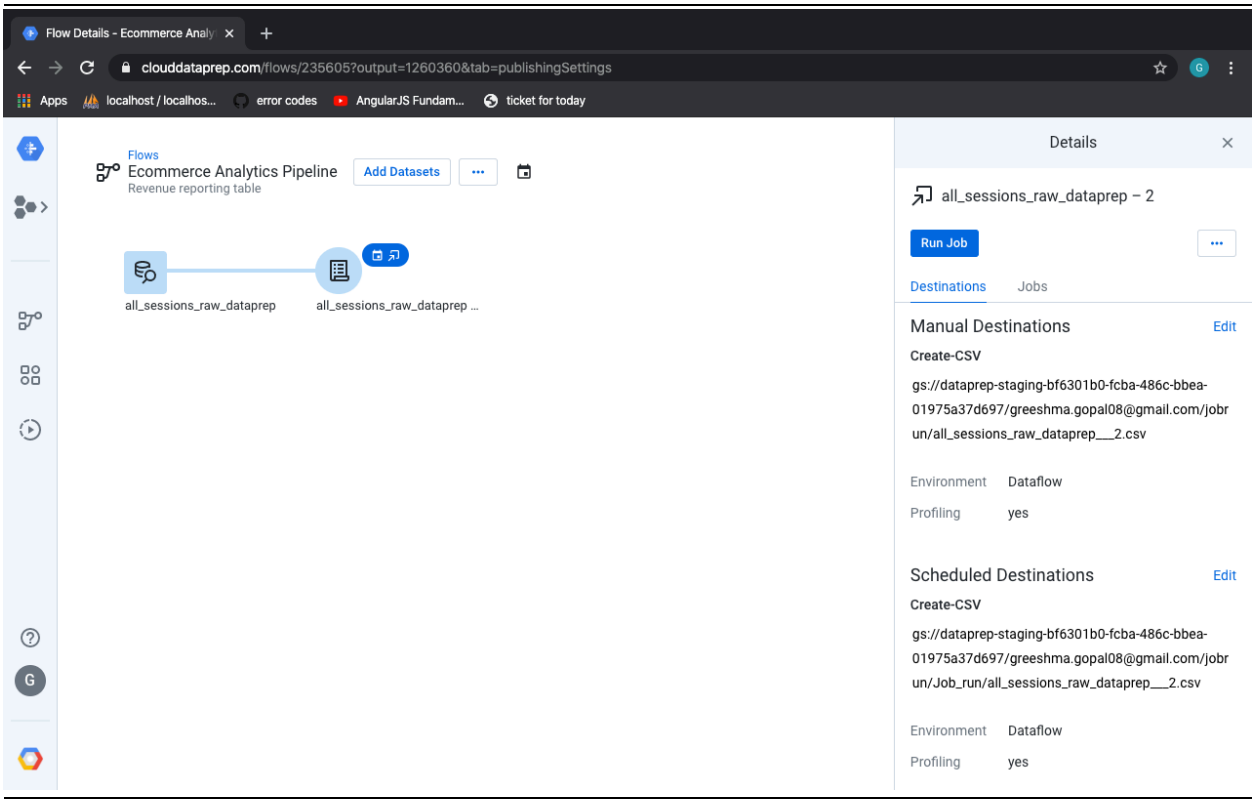


- Creating a case statement for the ecommerce action type along with all other manipulations which were done can be viewed

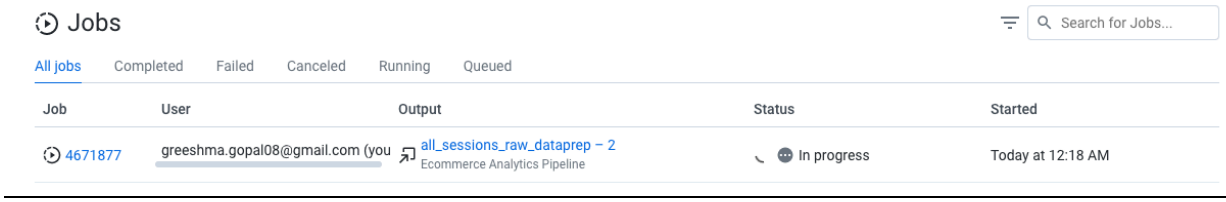


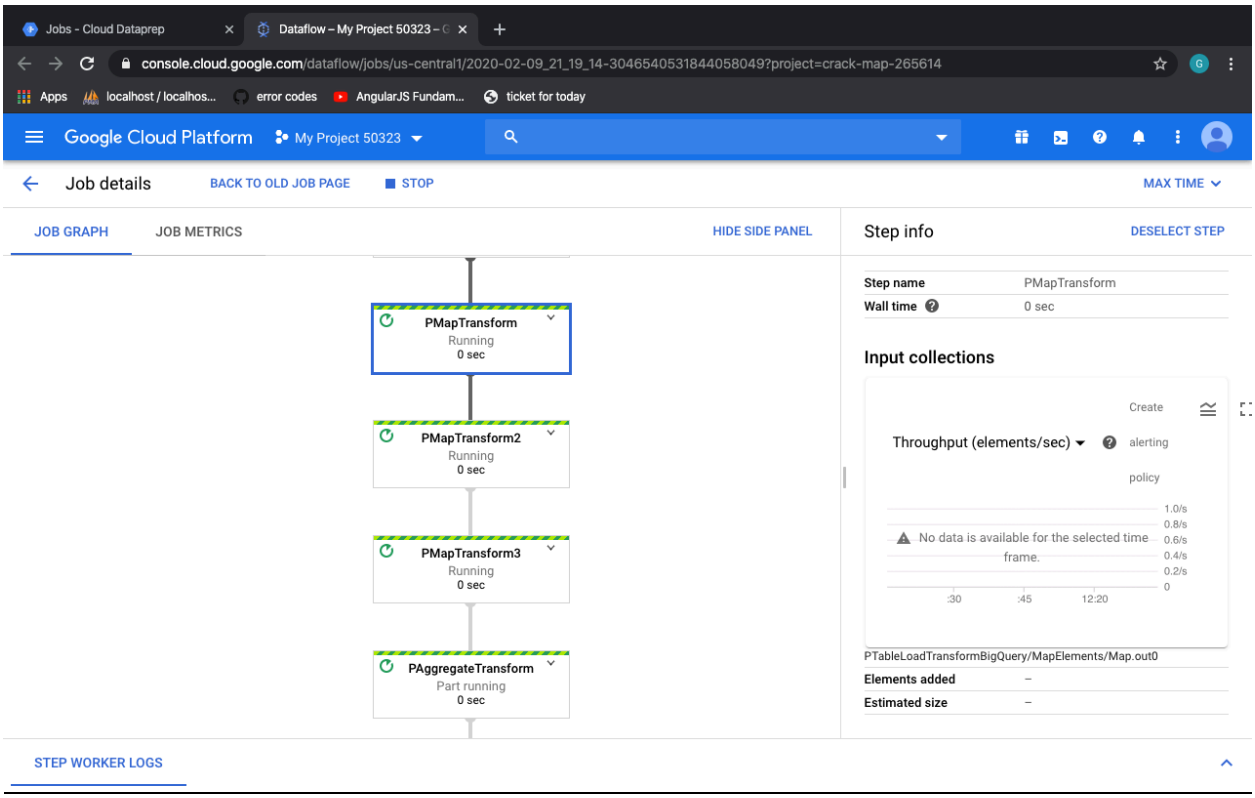
- Running and scheduling Cloud Dataprep jobs to BigQuery



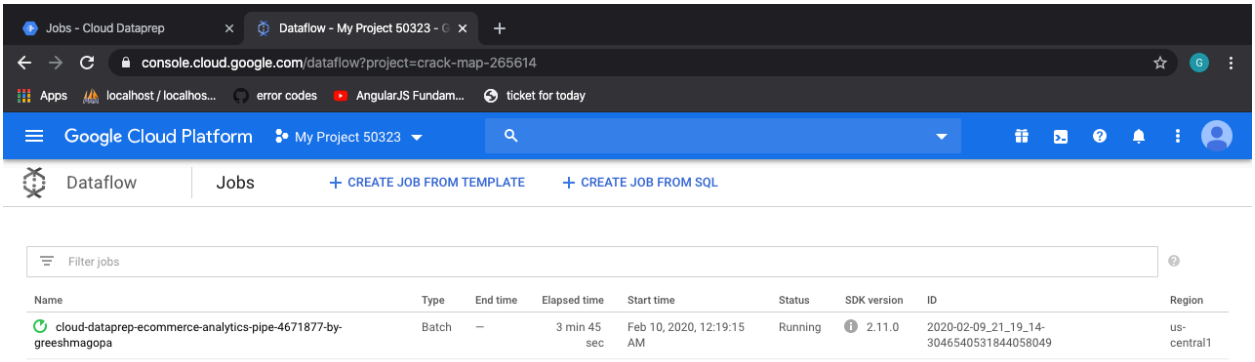


- Job scheduled is running. Can be verified this in the job status.

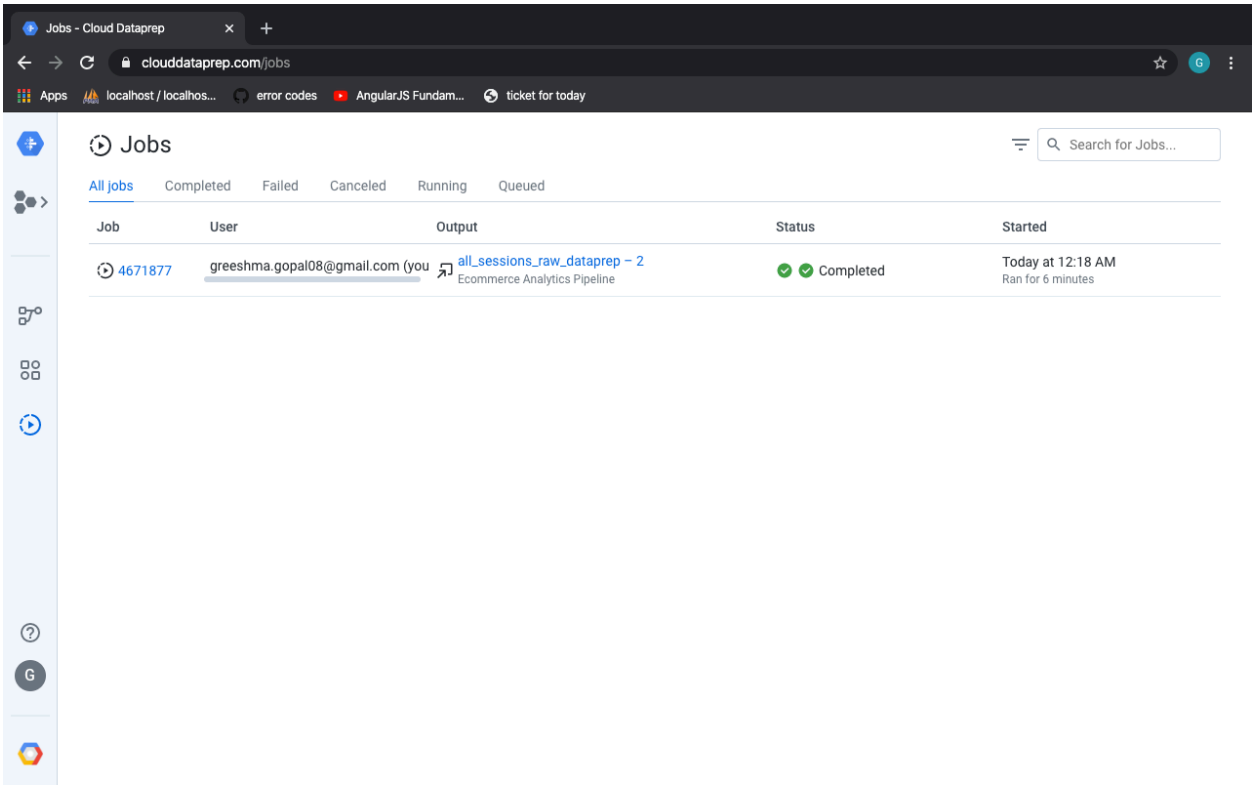




- The status of the job is running

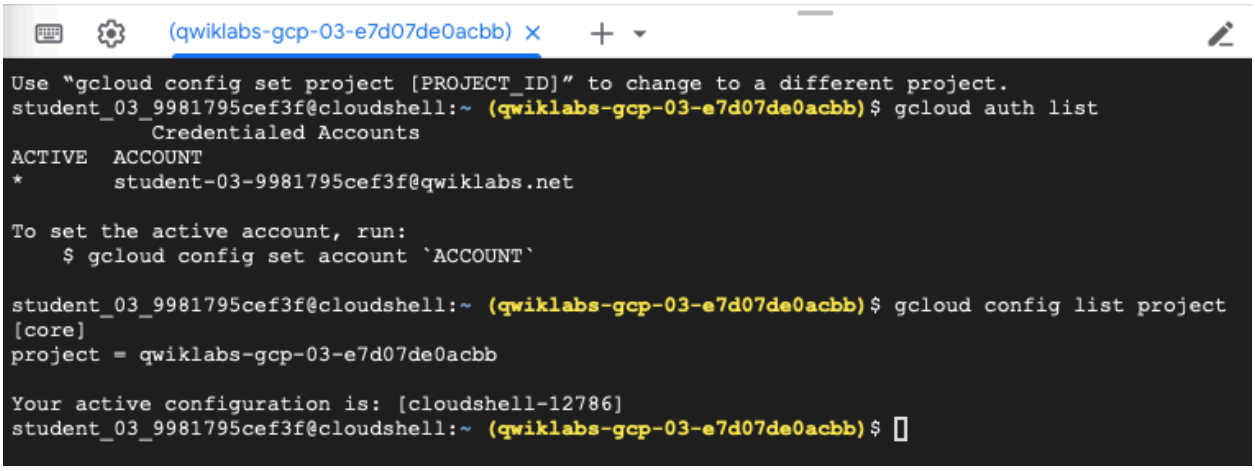


- The job has been successfully run and the first run is completed.



Lab 2: Visualizing Data with Google Data Studio

- Activating the Google Cloud Shell



- Connecting to the Lab Cloud SQL instance

- Connect to the mysql command line interface

```
student-03-9981795cef3f@vm-qldm-11645385-79e98852656c6b99:~$ MYSQLIP=$(gcloud sql instances describe \
> flights --format="value(ipAddresses.ipAddress)")
student-03-9981795cef3f@vm-qldm-11645385-79e98852656c6b99:~$ mysql --host=$MYSQLIP --user=root  --password
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 100
Server version: 5.7.25-google (Google)

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]>
```

- Count the number of records by running the SELECT query

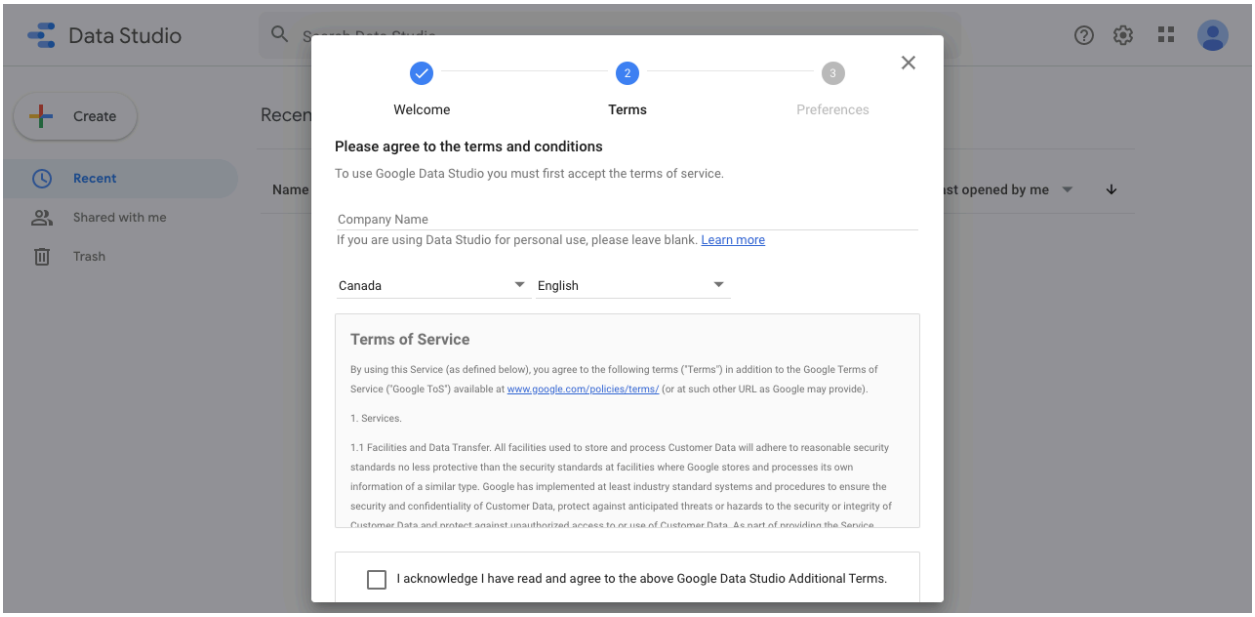
```
Database changed
MySQL [bts]> describe flights;select count(*) from flights;
+-----------------------+-------------+------+-----+---------+-------+
| Field                 | Type        | Null | Key | Default | Extra |
+-----------------------+-------------+------+-----+---------+-------+
| FL_DATE               | date        | YES  | MUL | NULL    |       |
| UNIQUE_CARRIER        | varchar(16) | YES  |     | NULL    |       |
| AIRLINE_ID            | varchar(16) | YES  |     | NULL    |       |
| CARRIER               | varchar(16) | YES  |     | NULL    |       |
| FL_NUM                | int(11)     | YES  |     | NULL    |       |
| ORIGIN_AIRPORT_ID     | varchar(16) | YES  | MUL | NULL    |       |
| ORIGIN_SEQ_ID         | varchar(16) | YES  |     | NULL    |       |
| ORIGIN_CITY_MARKET_ID | varchar(16) | YES  |     | NULL    |       |
| ORIGIN                | varchar(16) | YES  |     | NULL    |       |
| DEST_AIRPORT_ID       | varchar(16) | YES  |     | NULL    |       |
| DEST_AIRPORT_SEQ_ID   | varchar(16) | YES  |     | NULL    |       |
| DEST_CITY_MARKET_ID   | varchar(16) | YES  |     | NULL    |       |
| DEST                  | varchar(16) | YES  |     | NULL    |       |
| CRS_DEP_TIME          | int(11)     | YES  |     | NULL    |       |
| DEP_TIME              | int(11)     | YES  | MUL | NULL    |       |
| DEP_DELAY             | float       | YES  | MUL | NULL    |       |
| TAXI_OUT              | float       | YES  |     | NULL    |       |
| WHEELS_OFF            | int(11)     | YES  |     | NULL    |       |
| WHEELS_ON             | int(11)     | YES  |     | NULL    |       |
| TAXI_IN               | float       | YES  |     | NULL    |       |
| CRS_ARR_TIME          | int(11)     | YES  |     | NULL    |       |
| ARR_TIME              | int(11)     | YES  |     | NULL    |       |
| ARR_DELAY             | float       | YES  | MUL | NULL    |       |
| CANCELLED             | float       | YES  |     | NULL    |       |
| CANCELLATION_CODE     | varchar(16) | YES  |     | NULL    |       |
| DIVERTED              | float       | YES  |     | NULL    |       |
| DISTANCE              | float       | YES  |     | NULL    |       |
+-----------------------+-------------+------+-----+---------+-------+
27 rows in set (0.00 sec)

+----------+
| count(*) |
+----------+
|   899159 |
+----------+
1 row in set (0.23 sec)

MySQL [bts]>
```

- Creating some table views to fetch sets of flights that are delayed by 10, 15 and 20 minutes respectively

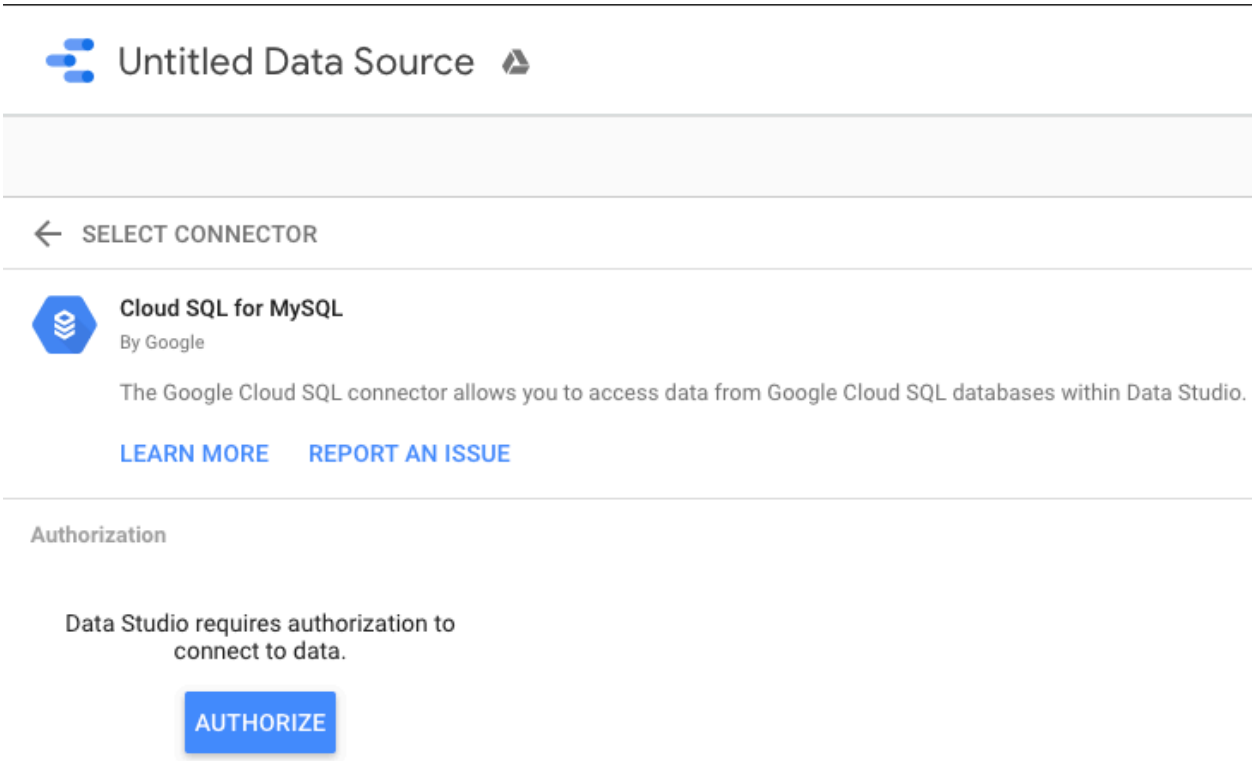
```
MySQL [bts]> CREATE VIEW delayed_10 AS SELECT * FROM flights WHERE dep_delay > 10;
Query OK, 0 rows affected (0.00 sec)

MySQL [bts]> CREATE VIEW delayed_15 AS SELECT * FROM flights WHERE dep_delay > 15;
Query OK, 0 rows affected (0.01 sec)

MySQL [bts]> CREATE VIEW delayed_20 AS SELECT * FROM flights WHERE dep_delay > 20;
Query OK, 0 rows affected (0.00 sec)
```

Connect Data Studio to visually analyze the dataset

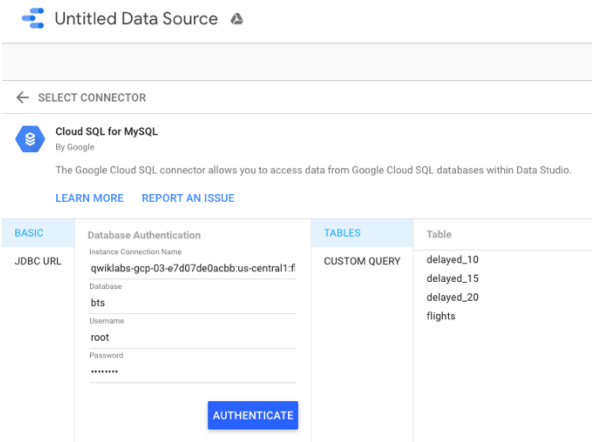
- Creating datasource



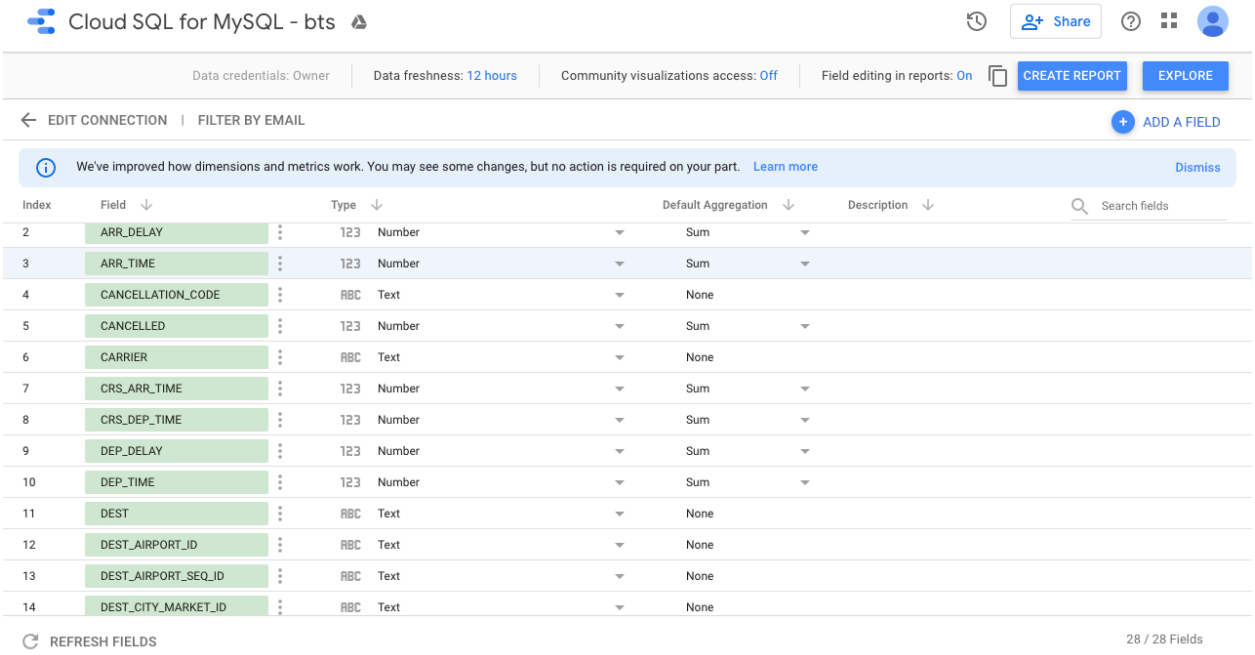
- Authorizing cloud SQL for mySQL for the datasource



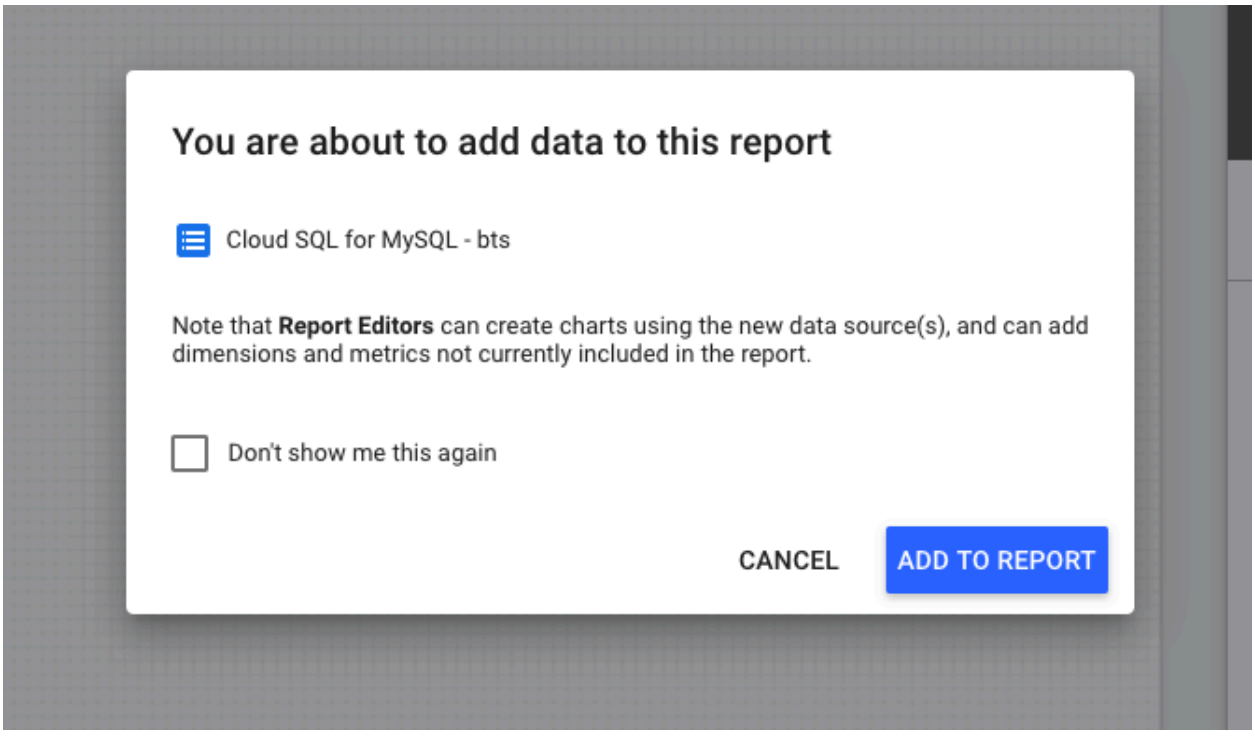
- Choosing the database to make the chart



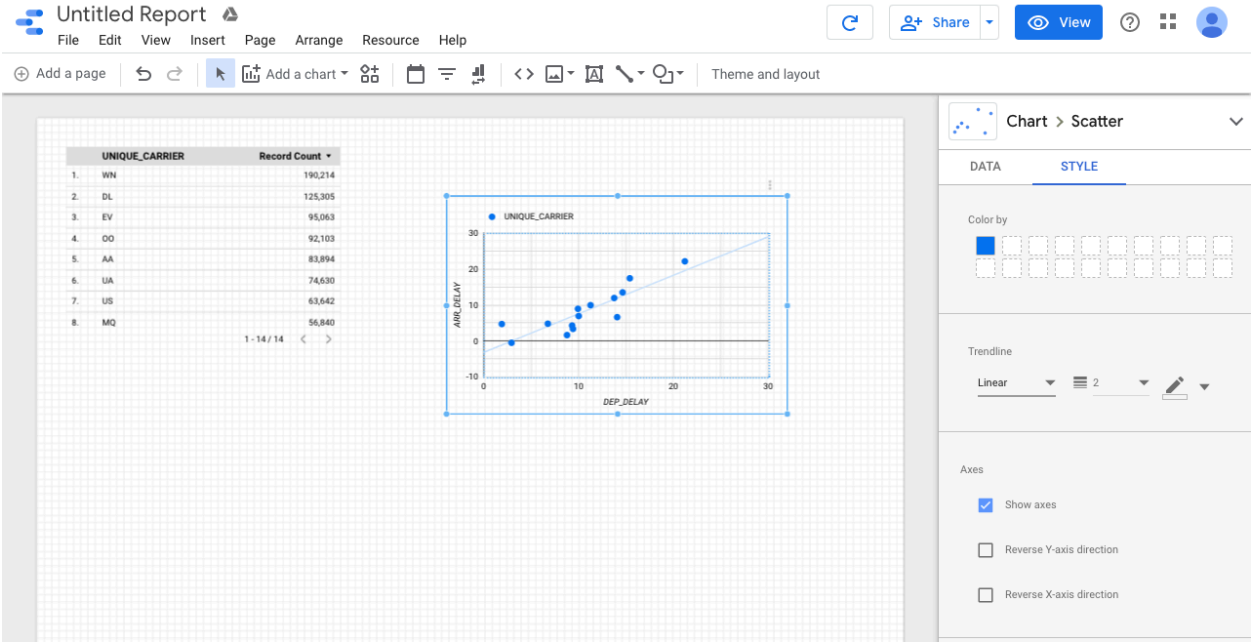
- We can see 28 fields when connected to the flights database



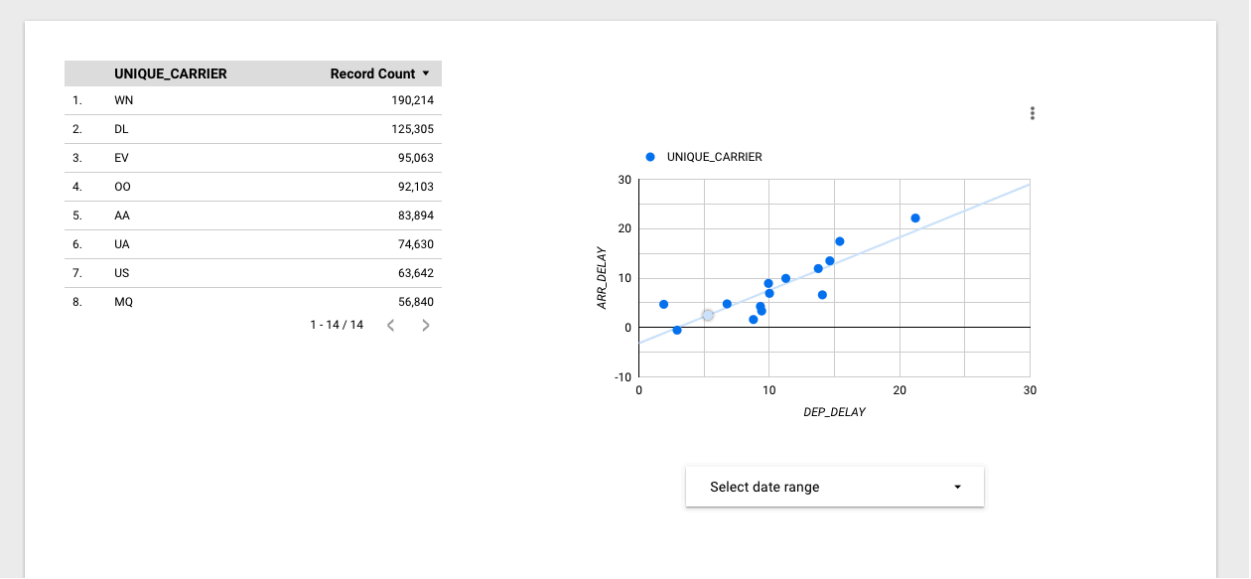
Create a Scatter chart using Data Studio



- Selecting the parameters to plot the scatter chart

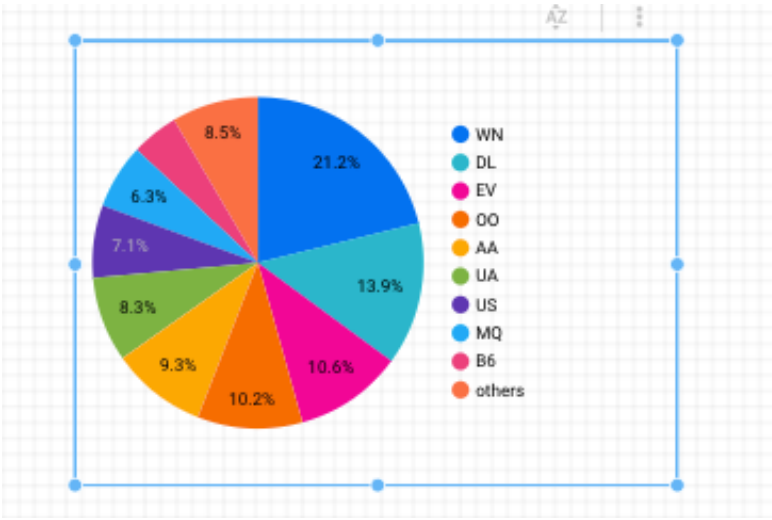


- Scattered chart along with date range selection has been added

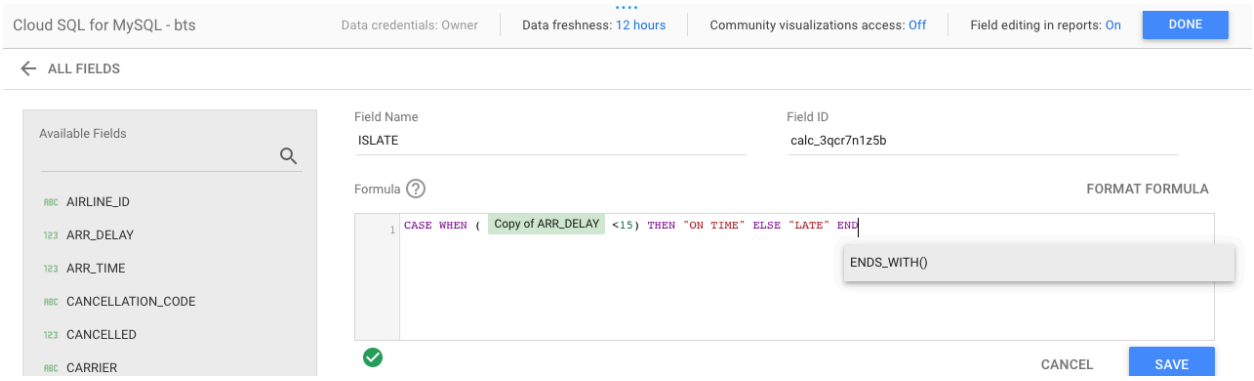


Add additional chart types to your report

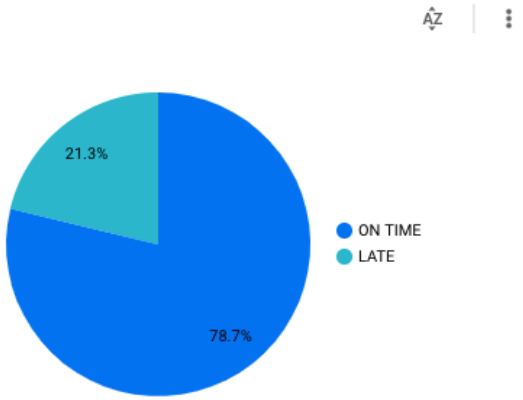
- Default pie chart has been created



- Creating new field for the chart

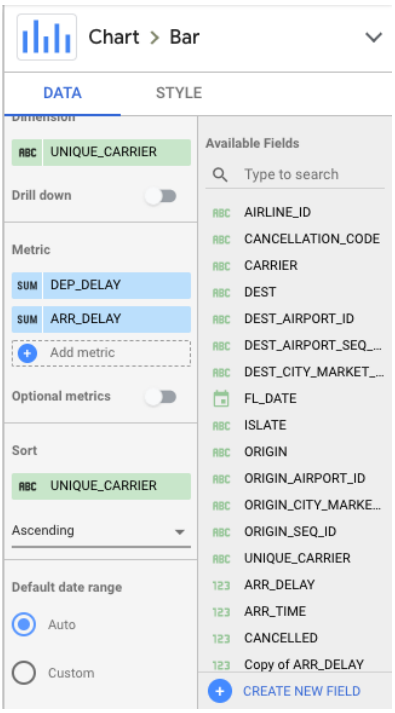


- Pie chart displaying the on time and late flights

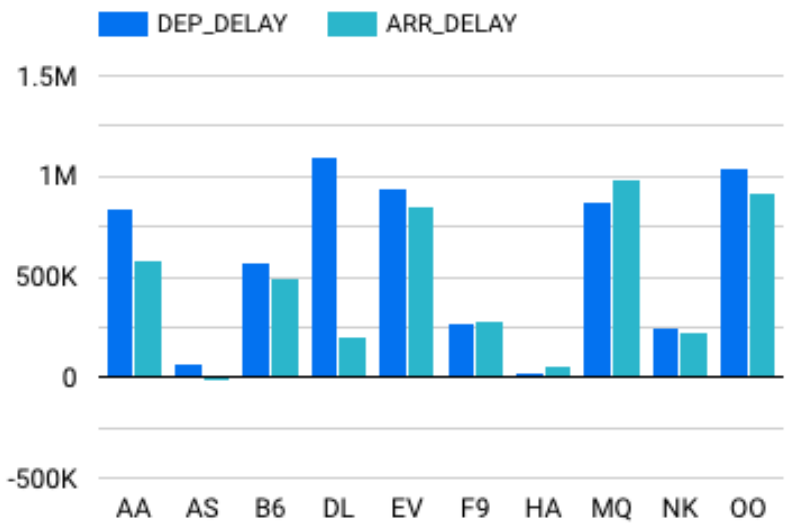


Add a bar graph

- Adding the settings of the bar chart



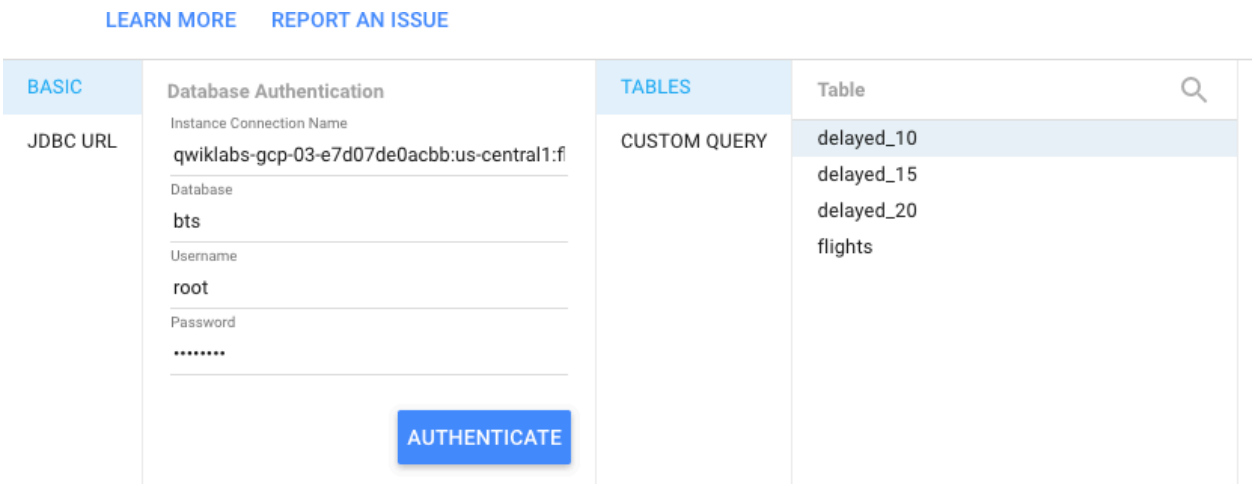
- Final bar chart of the flights

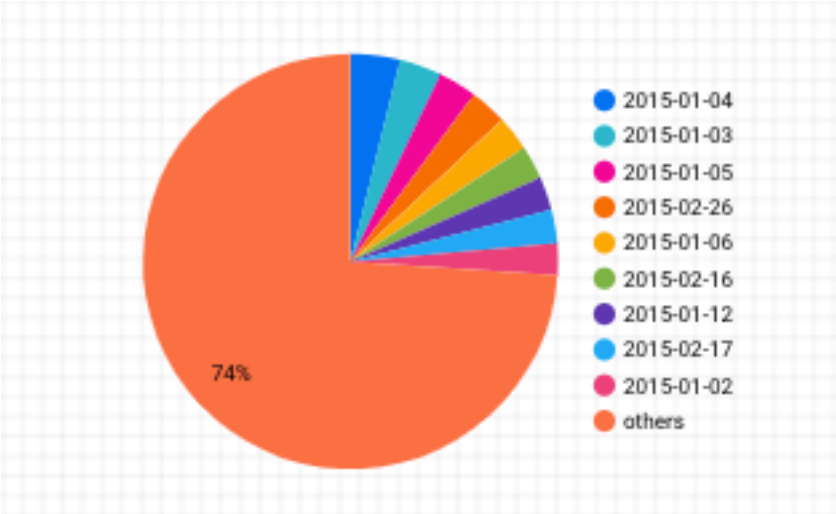


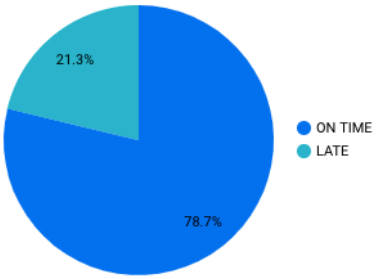
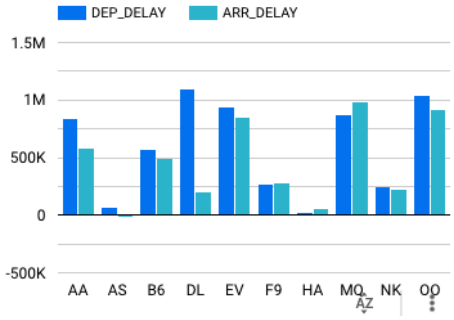
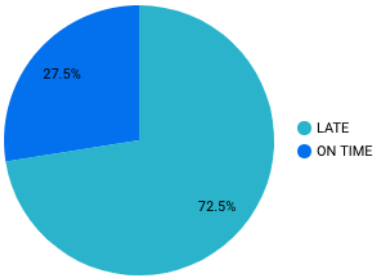
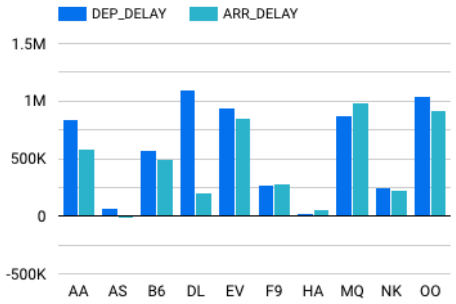
Create additional dashboard items for different departure delay thresholds

Added an additional Data Source for the Delayed_10 database table view

- Creating new datasource for delayed_10 view





Recreate the copy of the Arr_Delay field and the ISLATE calculated field.

- New field for islate and The second pie chart now displays the percentage of ontime and late flights for the Dealyed_10 view. Dashboards were created for four views delayed_10, delayed_15 and delayed_20