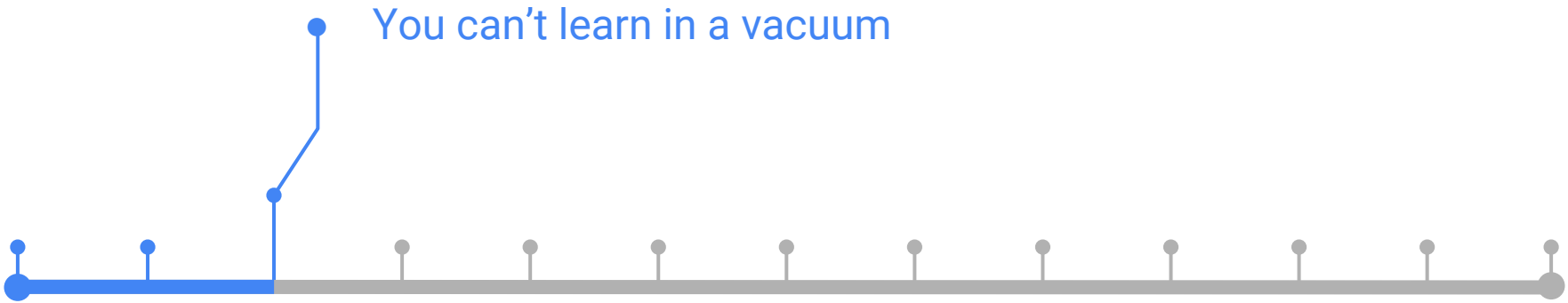
Deep Learning Walkthrough - 02

Cassie Kozyrkov
Chief Decision Scientist, Google Cloud
GitHub: kozyrkov; Twitter: @quaesita

Google Cloud

Step 2 | Get your data



Google Cloud

Confidential & Proprietary

Danger! Pitfall alert

Your effective dataset size is often much smaller than you think it is.*

Unless you've done a careful check, assume you're at small scale. It's safer.

Google Cloud

* A notorious example is classifying rare events. If you don't have a lot of "yes" labels, your dataset is small. :(

Easy mode

What's the most forgiving environment for ML?

One where you frequently have **new data** flowing in

Populate a BigQuery Table

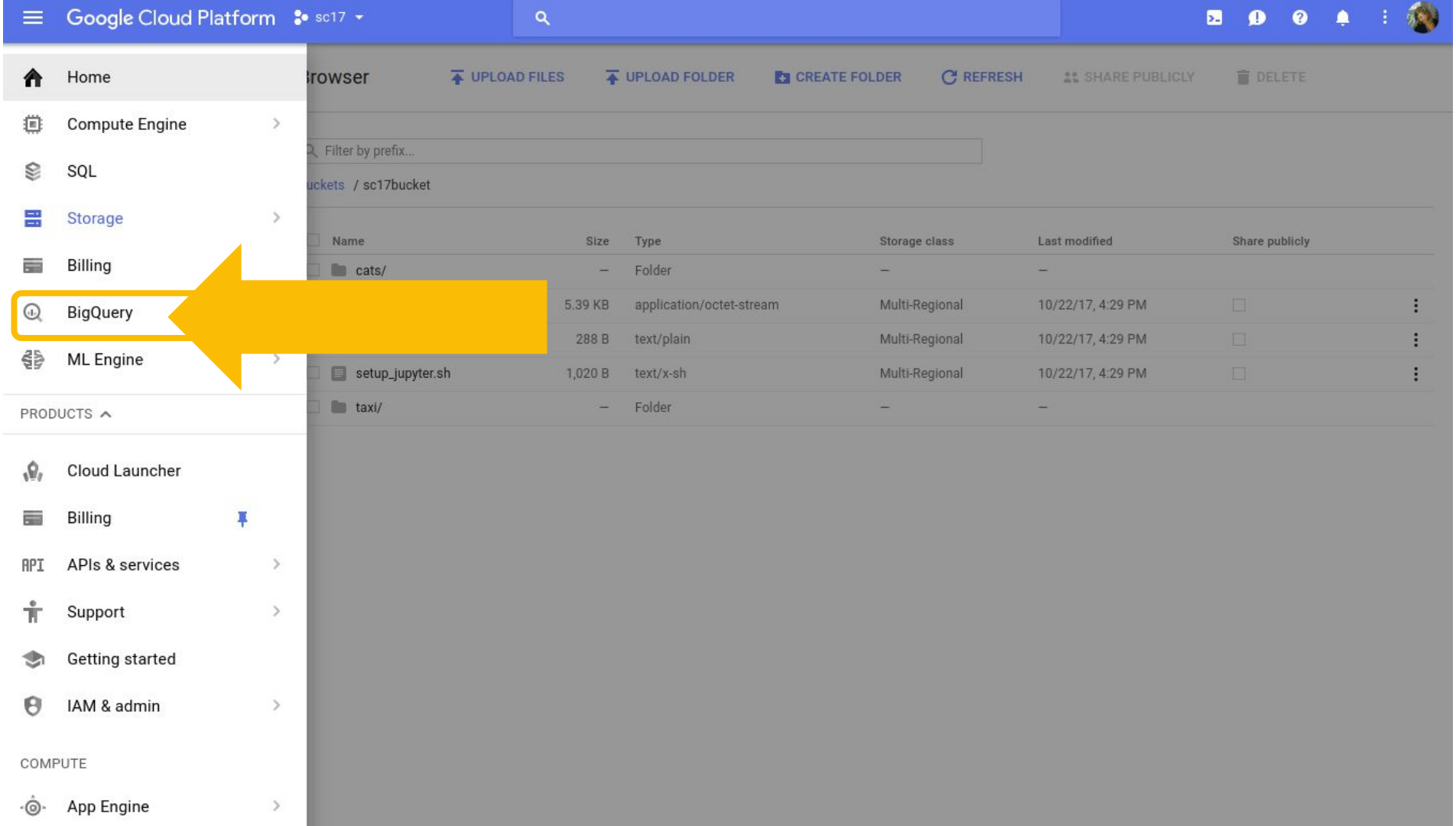
Secure | https://ssh.cloud.google.com/projects/super-188716/zones/us-east1-c/instances/super-188716-compute-instance?authus...

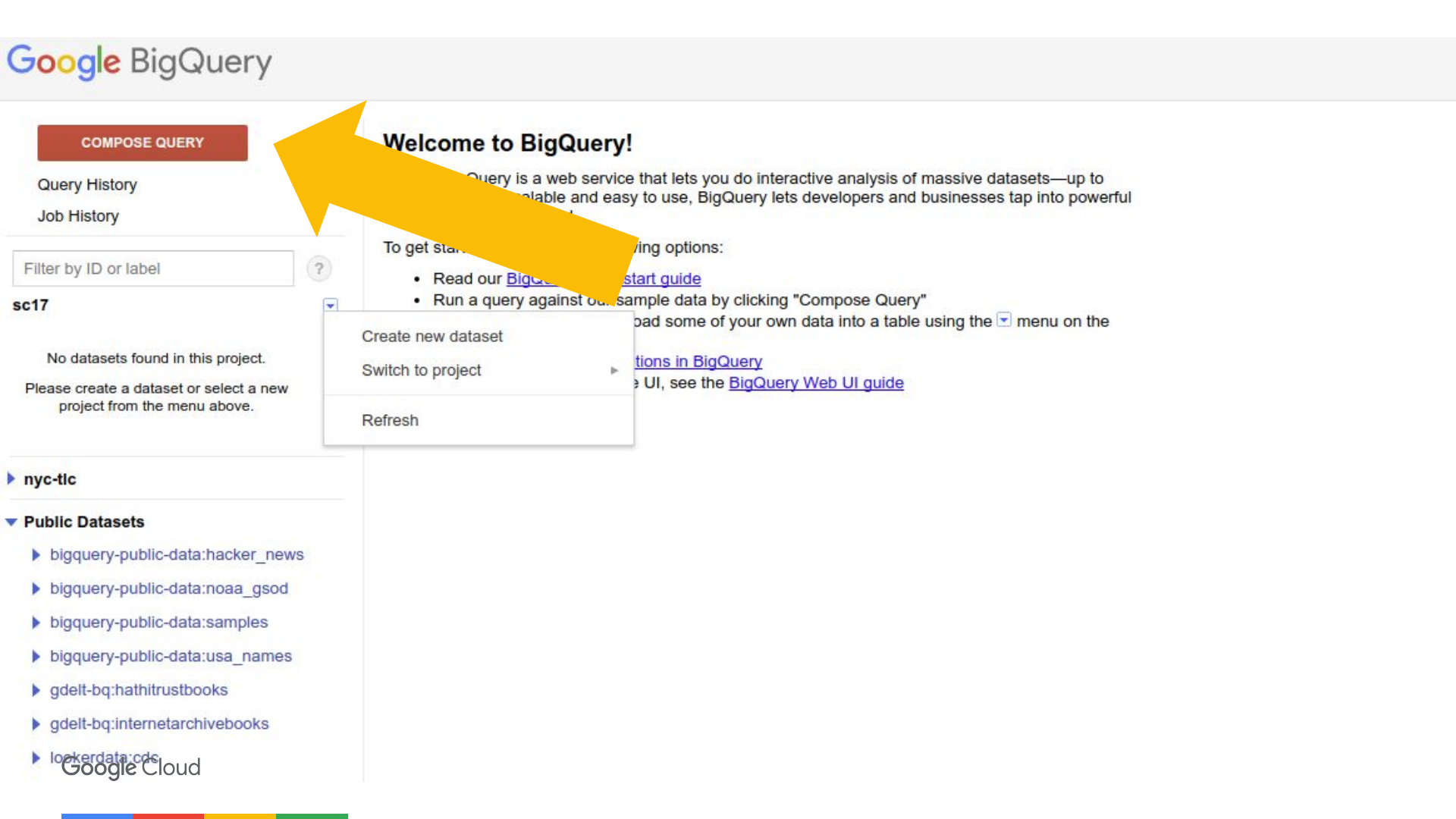
```
(env)  user  r@super-188716-compute-instance:~/sc17/cats$ sh run_step_2a_query.sh
Waiting on bqjob_r2b58c9818095dc06_0000016075203e0a_1 ... (14s) Current status: DONE
```

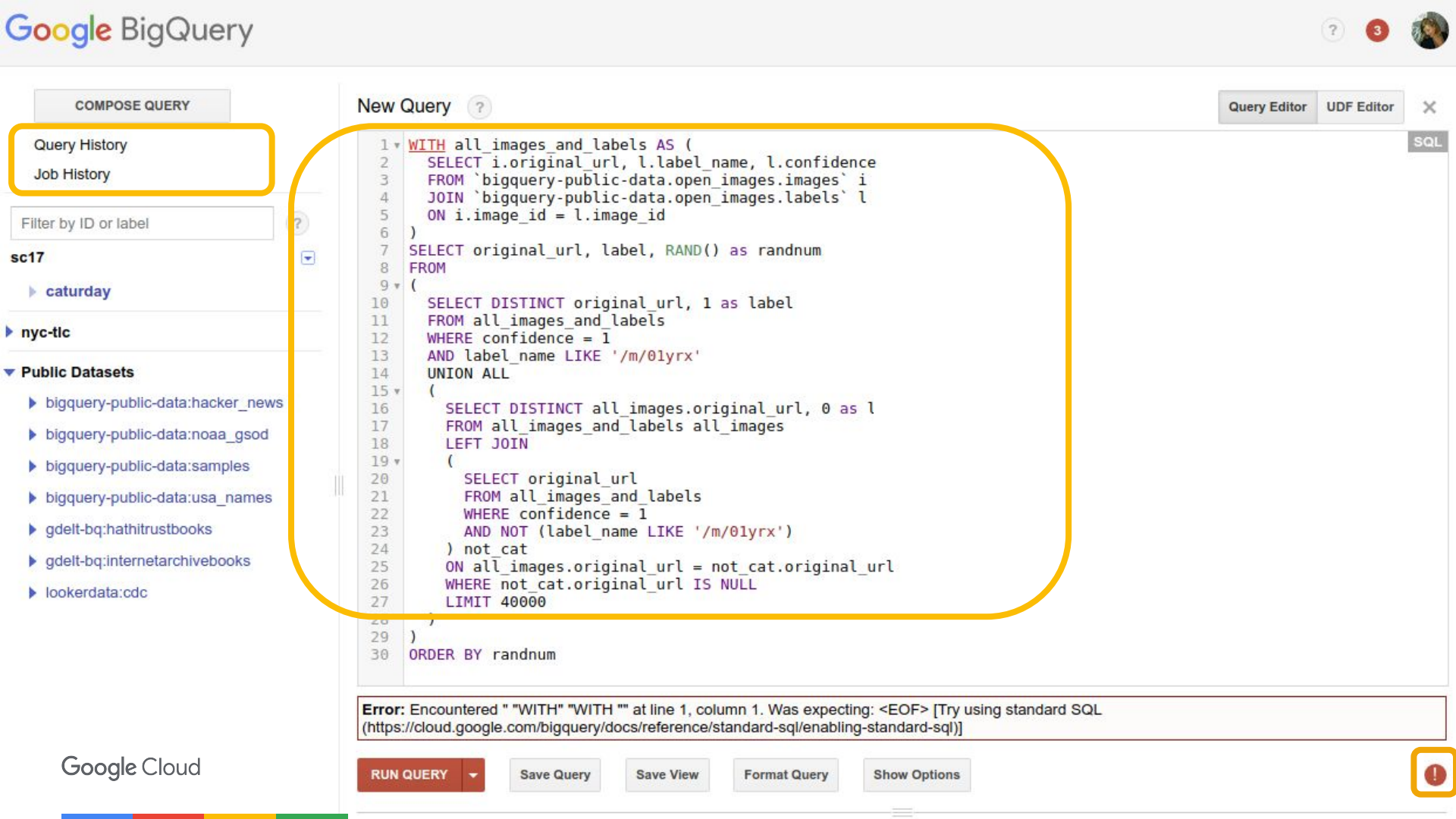
sh run_step_2a_query.sh

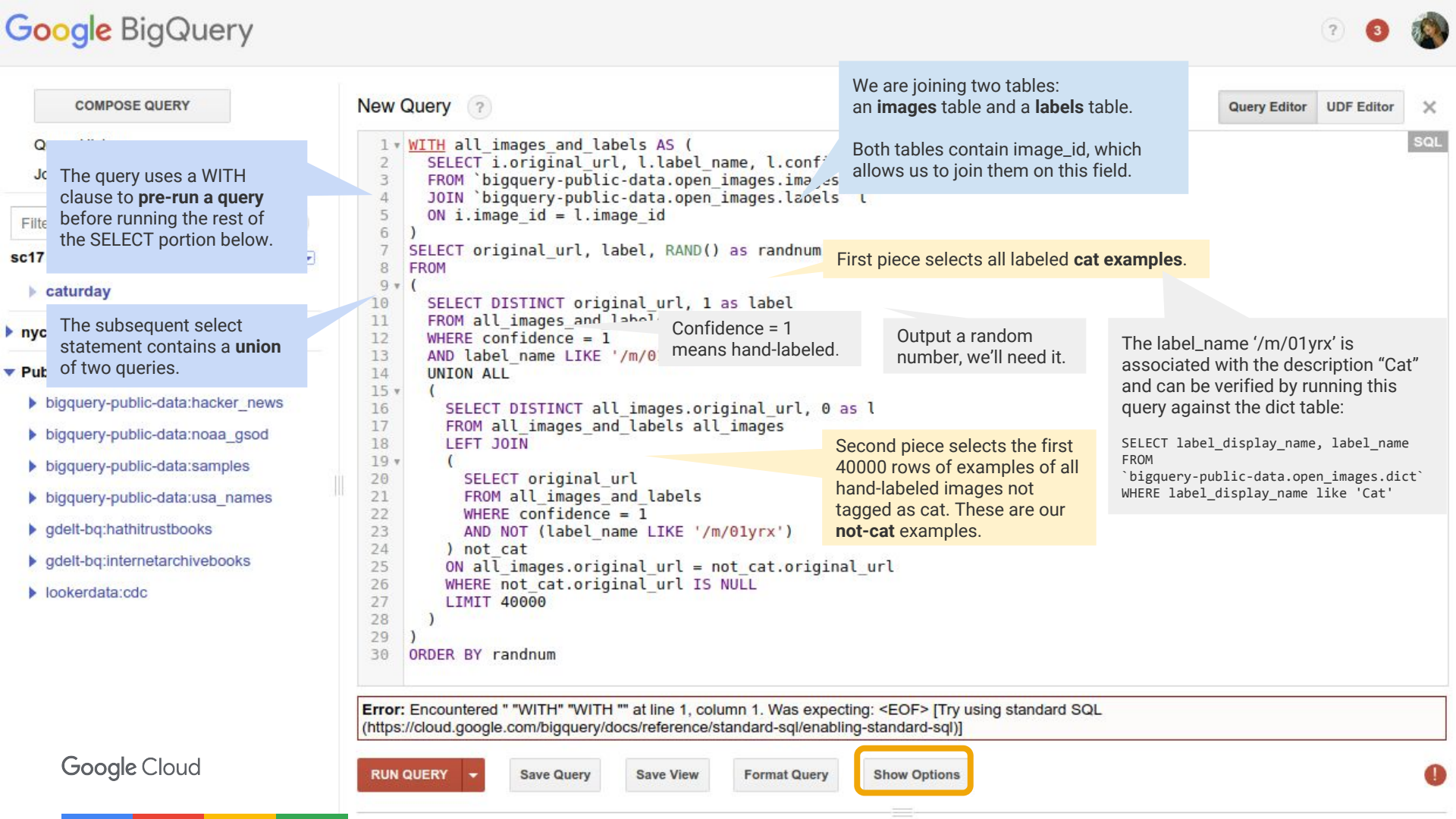
original_url	label	randnum
https://c7.staticflickr.com/9/8894/18277258689_5acf8b0a8a_o.jpg	0	8.526988257706591E-6
https://c2.staticflickr.com/3/2803/4020824567_0b7e36bfcd_o.jpg	0	1.3883631782896089E-5
https://c7.staticflickr.com/4/3318/3188958741_235b507552_o.jpg	1	2.702642935203947E-5
https://farm2.staticflickr.com/6171/6235175628_7f2b84a4fc_o.jpg	0	4.646176122080625E-5
https://c5.staticflickr.com/8/7153/6740118299_1aacb29588_o.jpg	1	6.0176315697990646E-5
https://farm7.staticflickr.com/5346/17204131514_007a763656_o.jpg	0	6.279622240379368E-5
https://c7.staticflickr.com/4/3751/9788632466_aa5204ba9f_o.jpg	1	6.490346933474278E-5
https://farm1.staticflickr.com/2593/4148830705_b01d9c83e8_o.jpg	1	6.60085155750334E-5
https://farm1.staticflickr.com/4045/4218663050_bd64e12b8a_o.jpg	1	8.965109192482501E-5
https://c1.staticflickr.com/4/3200/2350820461_a1f528269d_o.jpg	1	9.404920838229636E-5
https://c2.staticflickr.com/4/3224/2932224208_e799727148_o.jpg	1	1.0307303246426258E-4
https://c8.staticflickr.com/8/7105/7364627244_5ef122e36a_o.jpg	0	1.1777178603544769E-4
https://farm6.staticflickr.com/8109/8657170264_72a8146898_o.jpg	0	1.5044281923115577E-4
https://farm1.staticflickr.com/4034/4613672388_402a6de86c_o.jpg	0	1.5933507039811445E-4
https://farm4.staticflickr.com/195/499321718_b13b311e45_o.jpg	0	1.673010517302398E-4
https://farm3.staticflickr.com/7726/17339965475_e681255eb0_o.jpg	0	1.8154501185296823E-4
https://c1.staticflickr.com/1/67/203231915_df34a12f3d_o.jpg	1	1.880243997225417E-4
https://farm3.staticflickr.com/35/124377903_8d2463be2d_o.jpg	1	1.9661932271120008E-4
https://c4.staticflickr.com/4/3012/2947965305_d617988305_o.jpg	0	2.3681323217839E-4
https://c1.staticflickr.com/4/3644/3559941164_de8840e814_o.jpg	0	2.5992094483573425E-4
https://c2.staticflickr.com/3/2746/4479280283_0937d4f1e7_o.jpg	1	2.6391619407029054E-4
https://farm6.staticflickr.com/6160/6145126622_dbc68ec602_o.jpg	0	2.68379707281996E-4
https://c2.staticflickr.com/9/8264/8666029623_caa21476ea_o.jpg	1	2.7360821774990995E-4
https://c7.staticflickr.com/1/171/470614612_79aae80715_o.jpg	0	2.8884169378939895E-4
https://farm6.staticflickr.com/3856/14394042358_04e8040e77_o.jpg	0	2.974360479515159E-4
https://farm5.staticflickr.com/8610/16181928586_bc12b486f4_o.jpg	0	2.9849741811061037E-4
https://farm4.staticflickr.com/5612/15559754015_981c5b64b0_o.jpg	1	2.993901183343528E-4
https://c2.staticflickr.com/6/5084/5236877415_316d4398ed_o.jpg	1	3.080053757211765E-4
https://c3.staticflickr.com/1/153/379515012_9a99bb1584_o.jpg	1	3.1225777428489643E-4

Populate a BigQuery table [Behind the scenes]


Google Cloud Platform
sc17
Home
Compute Engine
SQL
Storage
Billing
BigQuery
ML Engine
PRODUCTS
Cloud Launcher
Billing
APIs & services
Support
Getting started
IAM & admin
COMPUTE
App Engine
Browser
UPLOAD FILES
UPLOAD FOLDER
CREATE FOLDER
REFRESH
SHARE PUBLICLY
DELETE
Filter by prefix...
uckets / sc17bucket
Name
Size
Type
Storage class
Last modified
Share publicly
cats/
5.39 KB
application/octet-stream
Multi-Regional
10/22/17, 4:29 PM
288 B
text/plain
Multi-Regional
10/22/17, 4:29 PM
setup_jupyter.sh
1,020 B
text/x-sh
Multi-Regional
10/22/17, 4:29 PM
taxi/
Folder


Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
sc17
No datasets found in this project.
Please create a dataset or select a new project from the menu above.
Create new dataset
Switch to project
Refresh
nyc-tlc
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
Welcome to BigQuery!
Google Cloud


Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
sc17
caturday
nyc-tlc
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
New Query
Query Editor
UDF Editor
SQL
WITH all_images_and_labels AS (
SELECT i.original_url, l.label_name, l.confidence
FROM `bigquery-public-data.open_images.images` i
JOIN `bigquery-public-data.open_images.labels` l
ON i.image_id = l.image_id
)
SELECT original_url, label, RAND() as randnum
FROM
(
SELECT DISTINCT original_url, 1 as label
FROM all_images_and_labels
WHERE confidence = 1
AND label_name LIKE '/m/01yrx'
UNION ALL
(
SELECT DISTINCT all_images.original_url, 0 as l
FROM all_images_and_labels all_images
LEFT JOIN
(
SELECT original_url
FROM all_images_and_labels
WHERE confidence = 1
AND NOT (label_name LIKE '/m/01yrx')
) not_cat
ON all_images.original_url = not_cat.original_url
WHERE not_cat.original_url IS NULL
LIMIT 40000
)
)
ORDER BY randnum
Error: Encountered " "WITH" "WITH "" at line 1, column 1. Was expecting: <EOF> [Try using standard SQL (https://cloud.google.com/bigquery/docs/reference/standard-sql/enabling-standard-sql)]
RUN QUERY
Save Query
Save View
Format Query
Show Options
Google Cloud



We are joining two tables:
an **images** table and a **labels** table.

Both tables contain image_id, which allows us to join them on this field.

The query uses a WITH clause to **pre-run a query** before running the rest of the SELECT portion below.

The subsequent select statement contains a **union** of two queries.

First piece selects all labeled **cat examples**.

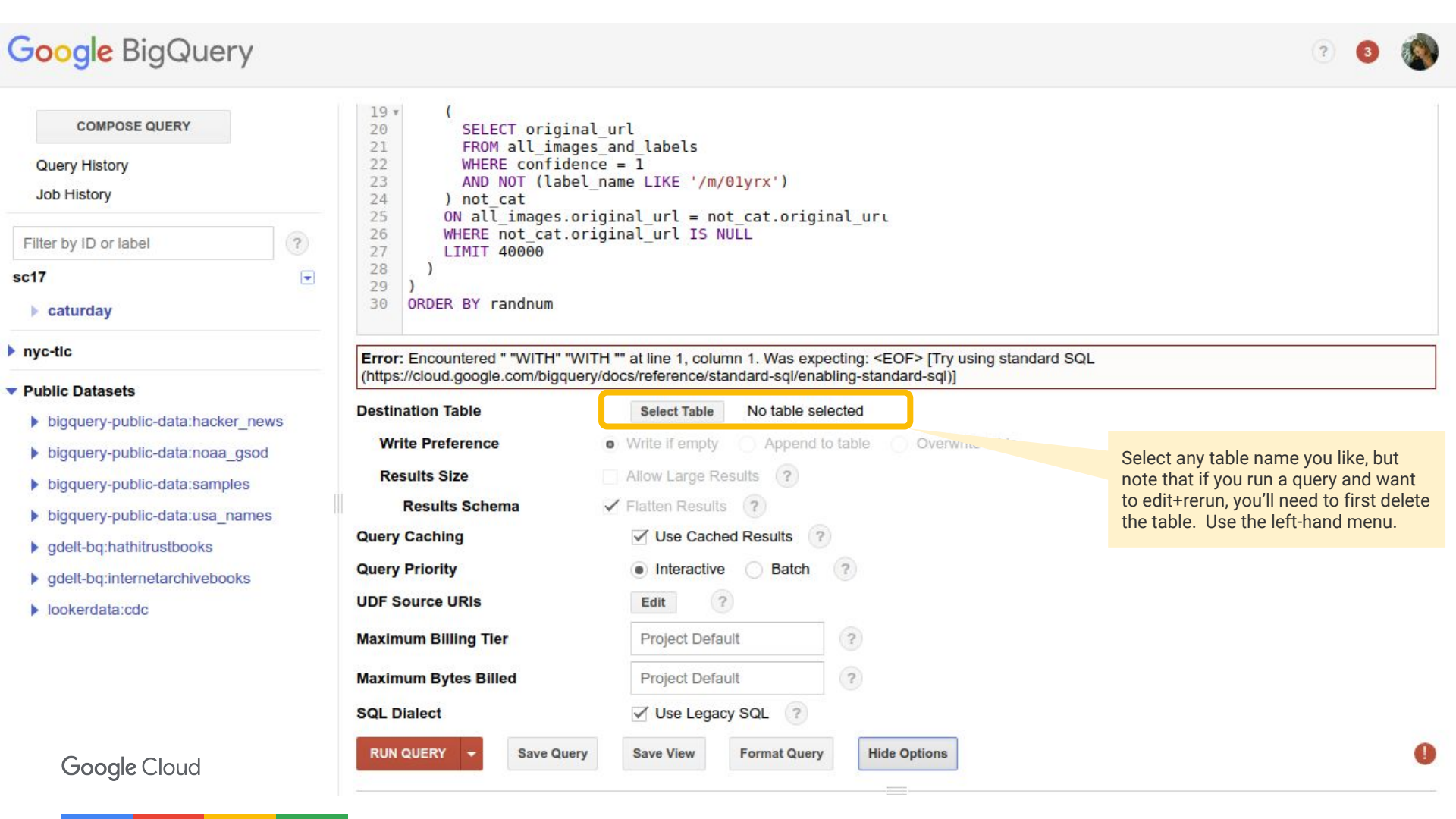
Confidence = 1 means hand-labeled.

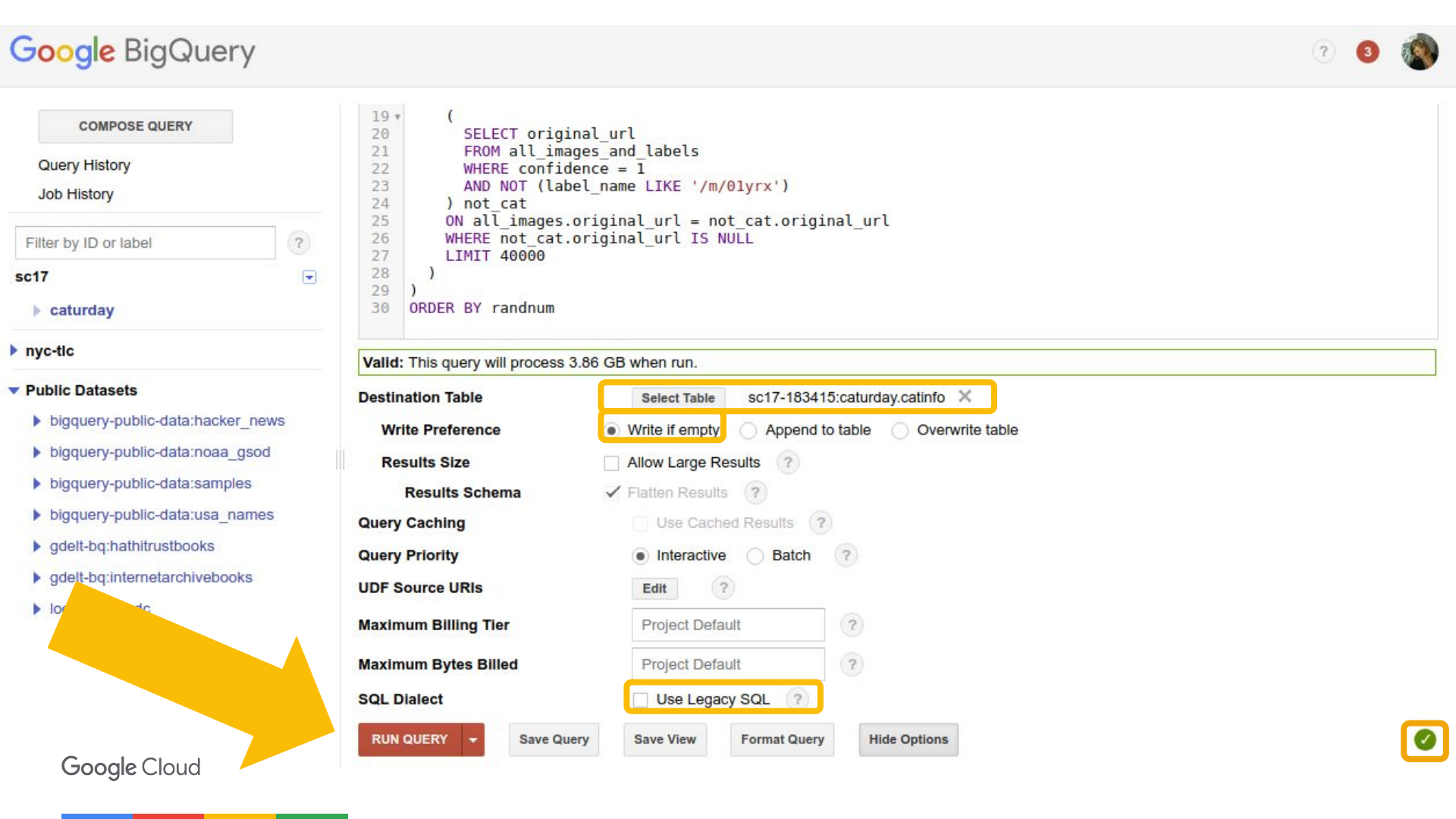
Output a random number, we'll need it.

The label_name '/m/01yrx' is associated with the description "Cat" and can be verified by running this query against the dict table:

```
SELECT label_display_name, label_name
FROM
`bigquery-public-data.open_images.dict`
WHERE label_display_name like 'Cat'
```

Second piece selects the first 40000 rows of examples of all hand-labeled images not tagged as cat. These are our **not-cat** examples.


Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
sc17
caturday
nyc-tlc
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
lookerdata:cdc
19 (
20 SELECT original_url
21 FROM all_images_and_labels
22 WHERE confidence = 1
23 AND NOT (label_name LIKE '/m/01yrx')
24) not_cat
25 ON all_images.original_url = not_cat.original_url
26 WHERE not_cat.original_url IS NULL
27 LIMIT 40000
28)
29)
30 ORDER BY randnum
Error: Encountered " "WITH" "WITH "" at line 1, column 1. Was expecting: <EOF> [Try using standard SQL (https://cloud.google.com/bigquery/docs/reference/standard-sql/enabling-standard-sql)]
Destination Table
Select Table
No table selected
Write Preference
Write if empty
Append to table
Results Size
Allow Large Results
Results Schema
Flatten Results
Query Caching
Use Cached Results
Query Priority
Interactive
Batch
UDF Source URIs
Edit
Maximum Billing Tier
Project Default
Maximum Bytes Billed
Project Default
SQL Dialect
Use Legacy SQL
RUN QUERY
Save Query
Save View
Format Query
Hide Options
Select any table name you like, but note that if you run a query and want to edit+rerun, you'll need to first delete the table. Use the left-hand menu.
Google Cloud


Google BigQuery
COMPOSE QUERY
Query History
Job History
Filter by ID or label
sc17
caturday
nyc-tlc
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
gdelt-bq:hathitrustbooks
gdelt-bq:internetarchivebooks
SELECT original_url
FROM all_images_and_labels
WHERE confidence = 1
AND NOT (label_name LIKE '/m/01yrx')
) not_cat
ON all_images.original_url = not_cat.original_url
WHERE not_cat.original_url IS NULL
LIMIT 40000
)
)
ORDER BY randnum
Valid: This query will process 3.86 GB when run.
Destination Table
Select Table
sc17-183415:caturday.catinfo
Write Preference
Write if empty
Append to table
Overwrite table
Results Size
Allow Large Results
Results Schema
Flatten Results
Query Caching
Use Cached Results
Query Priority
Interactive
Batch
UDF Source URIs
Edit
Maximum Billing Tier
Project Default
Maximum Bytes Billed
Project Default
SQL Dialect
Use Legacy SQL
RUN QUERY
Save Query
Save View
Format Query
Hide Options
Google Cloud

- nyc-tlc
- **Public Datasets**
  - bigquery-public-data:hacker_news
  - bigquery-public-data:noaa_gsod
  - bigquery-public-data:samples
  - bigquery-public-data:usa_names
  - gdelt-bq:hathitrustbooks
  - gdelt-bq:internetarchivebooks
  - lookerdata:cdc

```
12    WHERE confidence = 1
13    AND label_name LIKE '/m/01yrx'
14    UNION ALL
15    (
16      SELECT DISTINCT all_images.original_url, 0 as label
```

Valid: This query will process 3.86 GB when run.

Destination Table Select Table sc17-183415:caturday.catinfo

Write Preference Write if empty / Append to table / Overwrite table

Results Size Allow Large Results

Results Schema Flatten Results

Query Caching Use Cached Results

Query Priority Interactive / Batch

UDF Source URIs Edit

Maximum Billing Tier Project Default

Maximum Bytes Billed Project Default

SQL Dialect Use Legacy SQL

RUN QUERY | Save Query | Save View | Format Query | Hide Options | Query complete (14.6s elapsed, 3.86 GB processed)

Results | Explanation | Job Information | Download as CSV | Download as JSON | Save as Table | Save to Google Sheets

Row	original_url	label	randnum
1	https://c8.staticflickr.com/4/3359/4615280607_966caab870_o.jpg	1	1.749116054572397E-5
2	https://c1.staticflickr.com/3/2794/4086459730_0fb1054cb2_o.jpg	0	1.7994502079510255E-5
3	https://c1.staticflickr.com/9/8087/8566786907_67c59c434b_o.jpg	0	2.4605481685258113E-5
4	https://farm5.staticflickr.com/8323/8109857300_9ec1e9fbf7_o.jpg	1	4.6472283419696225E-5
5	https://c5.staticflickr.com/4/3457/3309470896_cb8fceddba_o.jpg	1	7.413447964088604E-5

Table | JSON

First < Prev Rows 1 - 5 of 80231 Next > Last

Google Cloud

- nyc-tlc
- Public Datasets
  - bigquery-public-data:hacker_news
  - bigquery-public-data:noaa_gsod
  - bigquery-public-data:samples
  - bigquery-public-data:usa_names
  - gdelt-bq:hathitrustbooks
  - gdelt-bq:internetarchivebooks
  - lookerdata:cdc

```
13
14
15
16                                    inal_url, 0 as label
```

Vali

Dest sc17-183415:caturday.catinfo

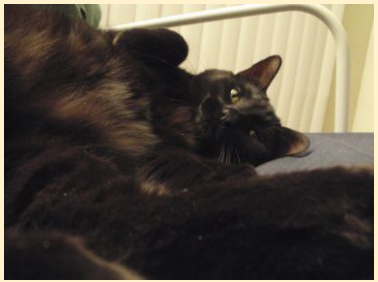
Append to table Overwrite table

sults

Query Caching

Use Legacy

Save View



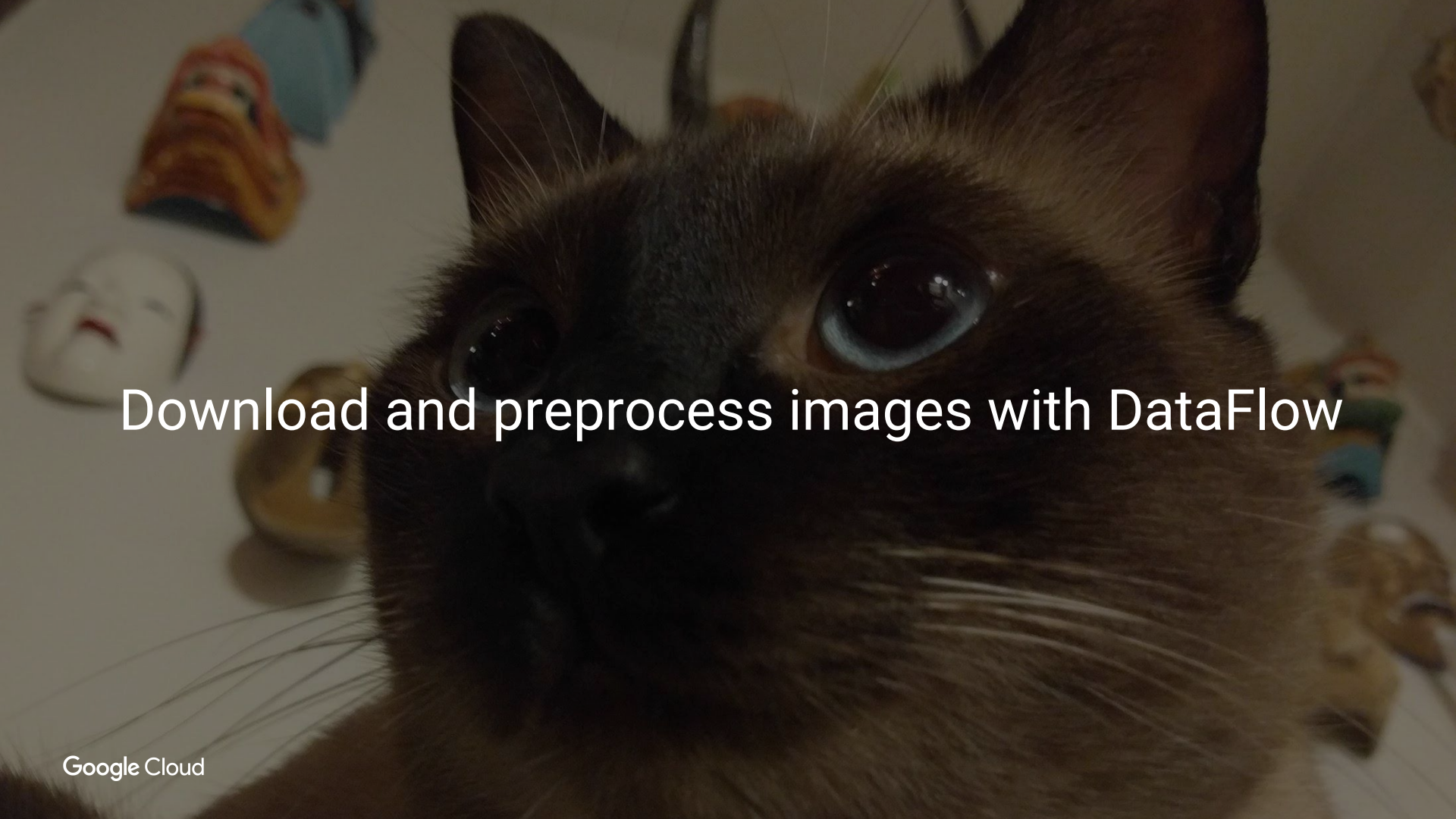
Results | Explanation | Job Information | Download | Download as J | Save as Table | Save to Google Sheets

Row	original_url	abel	randnum
1	https://c8.staticflickr.com/4/3359/4615280607_966caab870_o.jpg	1	/491160545 7E-5
2	https://c1.staticflickr.com/3/2794/4086459730_0fb1054cb2_o.jpg	0	.799450 9510255E-5
3	https://c1.staticflickr.com/9/8087/8566786907_67c59c434b_o.jpg	0	2.4 5481685258113E-5
4	https://farm5.staticflickr.com/8323/8109857300_9ec1e9fbf7_o.jpg	1	4.6472283419696225E-5
5	https://c5.staticflickr.com/4/3457/3309470896_cb8fceddba_o.jpg	1	

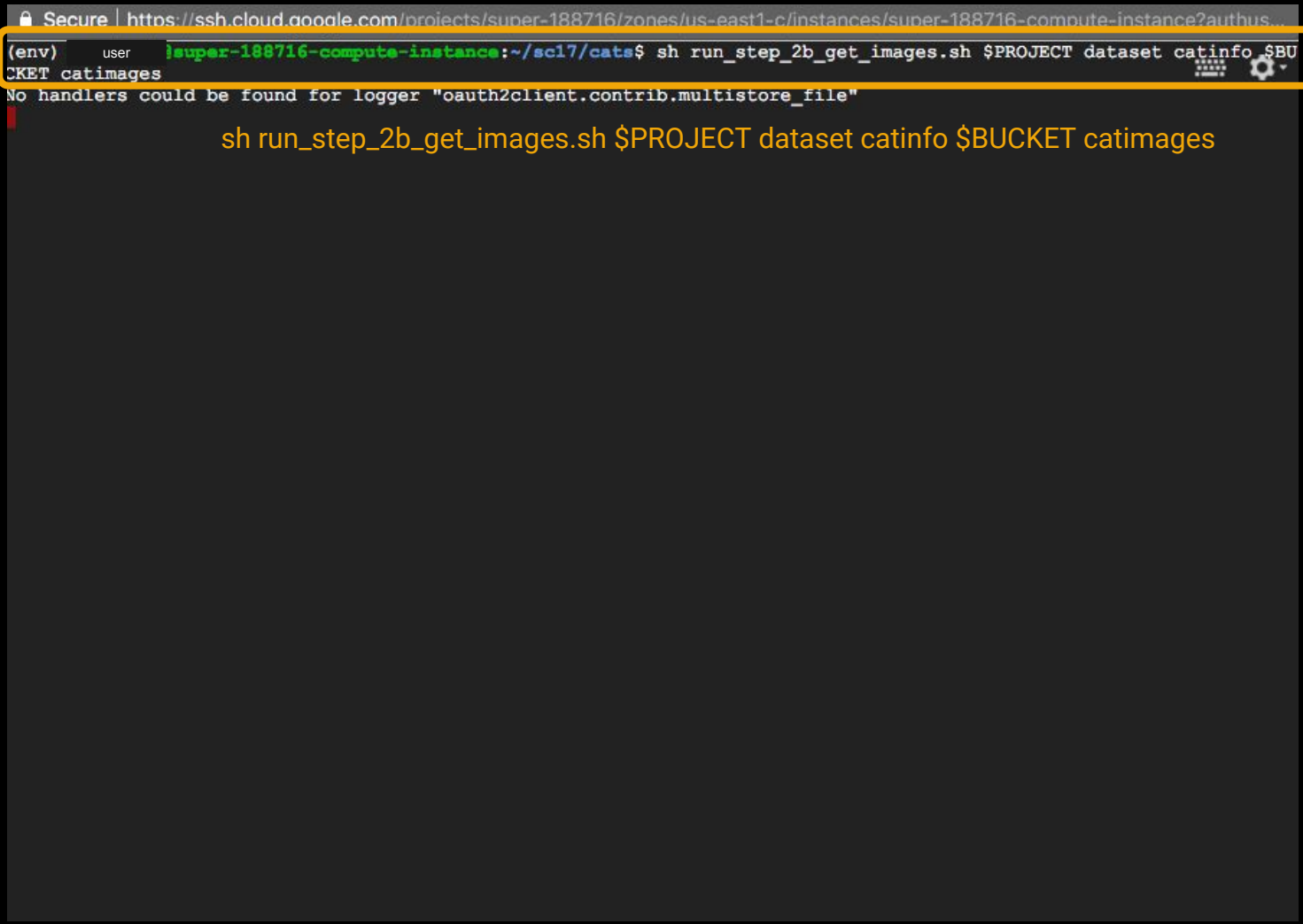
Table | JSON

First < Prev Rows 1 - 5 of 80231 Next > Last

Google Cloud

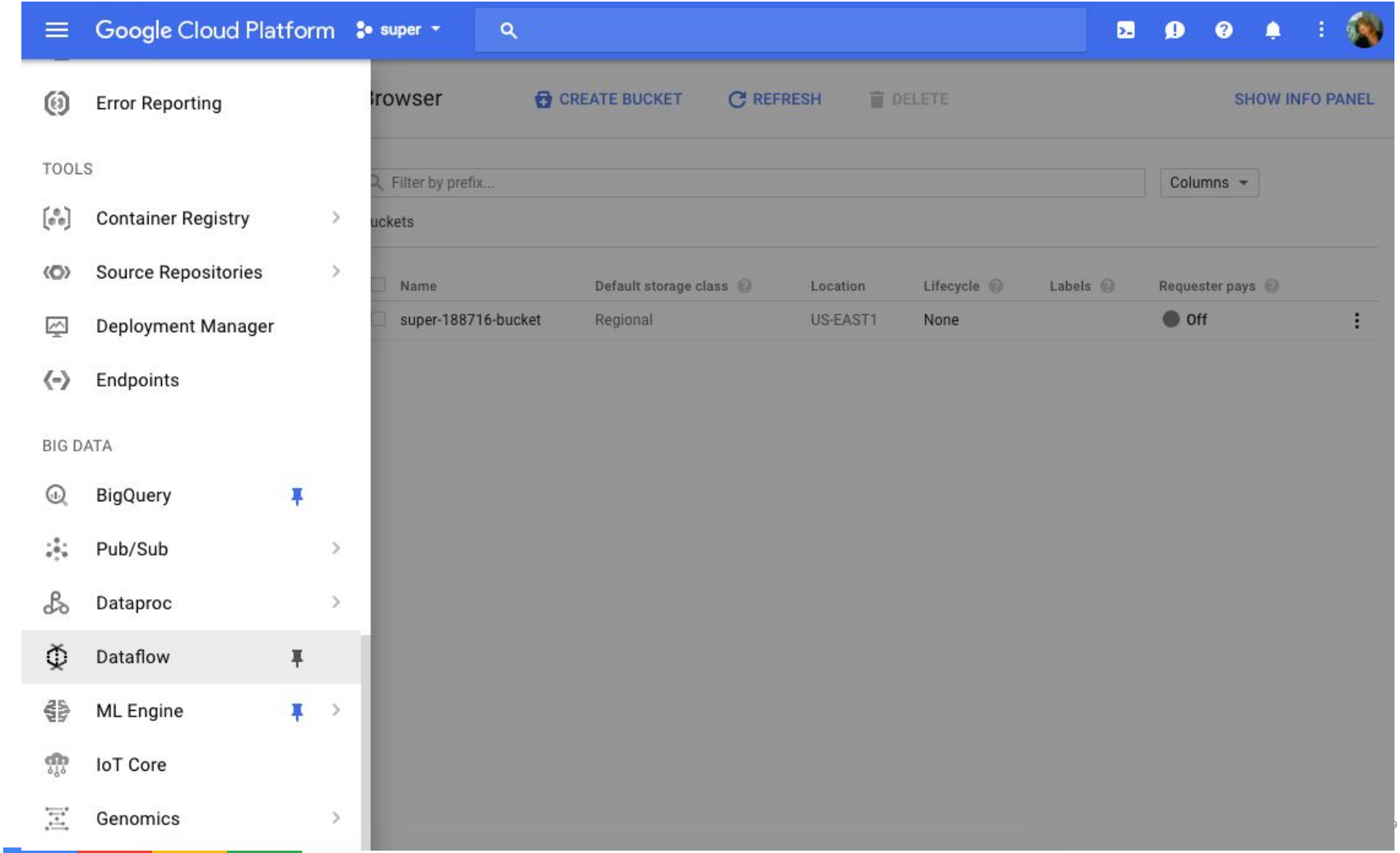


Google Cloud


Secure | https://ssh.cloud.google.com/projects/super-188716/zones/us-east1-c/instances/super-188716-compute-instance?authus...
(env) user !super-188716-compute-instance:~/sc17/cats$ sh run_step_2b_get_images.sh $PROJECT dataset catinfo $BUCKET catimages
No handlers could be found for logger "oauth2client.contrib.multistore_file"
sh run_step_2b_get_images.sh $PROJECT dataset catinfo $BUCKET catimages

Secure | https://ssh.cloud.google.com/projects/super-188716/zones/us-east1-c/instances/super-188716-compute-instance?authus...

```
INFO:root:2017-12-20T18:30:34.043Z: JOB_MESSAGE_DETAILED: (1f4f7cec0d1c28ca): Fusing consumer resize_and_pad_images
 into filter_bad_or_absent_images
INFO:root:2017-12-20T18:30:34.074Z: JOB_MESSAGE_DETAILED: (1f4f7cec0d1c2b48): Fusing consumer filter_bad_or_absent_
images into fetch_images_from_urls
INFO:root:2017-12-20T18:30:34.097Z: JOB_MESSAGE_DETAILED: (1f4f7cec0d1c2dc6): Fusing consumer write_images_to_stora
ge into resize_and_pad_images
INFO:root:2017-12-20T18:30:34.121Z: JOB_MESSAGE_DEBUG: (1f4f7cec0d1c2044): Workflow config is missing a default res
ource spec.
INFO:root:2017-12-20T18:30:34.149Z: JOB_MESSAGE_DEBUG: (1f4f7cec0d1c22c2): Adding StepResource setup and teardown t
o workflow graph.
INFO:root:2017-12-20T18:30:34.182Z: JOB_MESSAGE_DEBUG: (1f4f7cec0d1c2540): Adding workflow start and stop steps.
INFO:root:2017-12-20T18:30:34.224Z: JOB_MESSAGE_DEBUG: (1f4f7cec0d1c27be): Assigning stage ids.
INFO:root:2017-12-20T18:30:34.392Z: JOB_MESSAGE_DEBUG: (54dc4c9c134f910d): Executing wait step start3
INFO:root:2017-12-20T18:30:34.458Z: JOB_MESSAGE_BASIC: (54dc4c9c134f9124): Executing operation read_rows_from_cat_i
nfo_table+fetch_images_from_urls+filter_bad_or_absent_images+resize_and_pad_images+write_images_to_storage
INFO:root:2017-12-20T18:30:34.503Z: JOB_MESSAGE_DEBUG: (348e57263115b477): Starting worker pool setup.
INFO:root:2017-12-20T18:30:34.539Z: JOB_MESSAGE_BASIC: (348e57263115b6ad): Starting 50 workers in us-central1-f...
INFO:root:2017-12-20T18:30:34.573Z: JOB_MESSAGE_BASIC: (903c4cdd1507102b): BigQuery query issued as job: "dataflow_
job_257583136778705608". You can check its status with the bq tool: "bq show -j --project_id=super-188716 dataflow_
job_257583136778705608".
INFO:root:Job 2017-12-20_10_30_29-8607060276961813628 is in state JOB_STATE_RUNNING
INFO:root:2017-12-20T18:30:42.350Z: JOB_MESSAGE_DETAILED: (6199526dfc07429f): Autoscaling: Raised the number of wor
kers to 0 based on the rate of progress in the currently running step(s).
INFO:root:2017-12-20T18:30:58.331Z: JOB_MESSAGE_DETAILED: (6199526dfc0744c4): Autoscaling: Raised the number of wor
kers to 46 based on the rate of progress in the currently running step(s).
INFO:root:2017-12-20T18:30:58.350Z: JOB_MESSAGE_DETAILED: (6199526dfc07470a): Resized worker pool to 46, though goa
l was 50.  This could be a quota issue.
INFO:root:2017-12-20T18:31:03.652Z: JOB_MESSAGE_DETAILED: (6199526dfc074022): Autoscaling: Raised the number of wor
kers to 48 based on the rate of progress in the currently running step(s).
INFO:root:2017-12-20T18:31:03.680Z: JOB_MESSAGE_DETAILED: (6199526dfc074268): Resized worker pool to 48, though goa
l was 50.  This could be a quota issue.
INFO:root:2017-12-20T18:31:08.996Z: JOB_MESSAGE_DETAILED: (6199526dfc0744a3): Autoscaling: Raised the number of wor
kers to 50 based on the rate of progress in the currently running step(s).
```

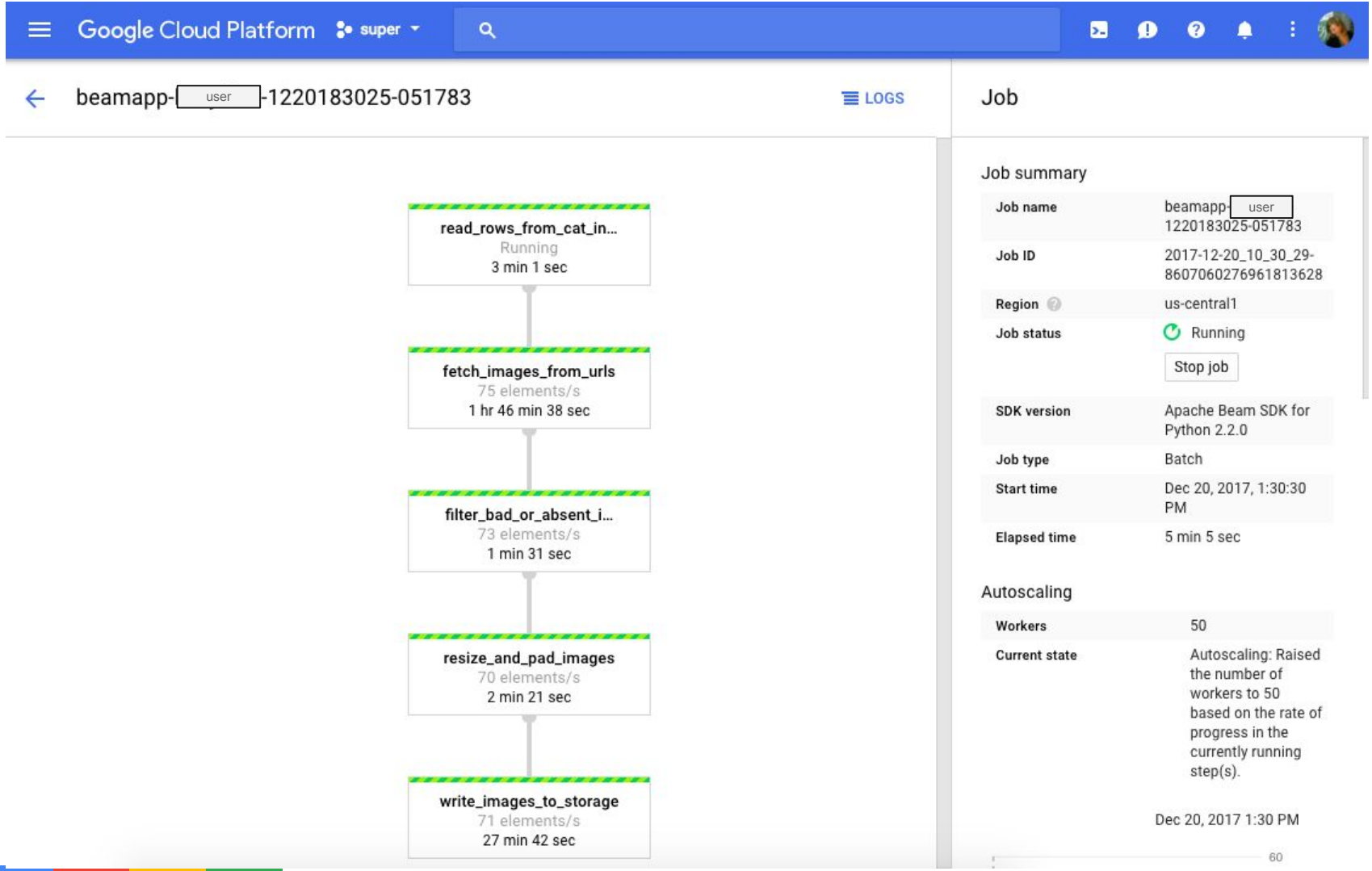

Google Cloud Platform
super
Error Reporting
TOOLS
Container Registry
Source Repositories
Deployment Manager
Endpoints
BIG DATA
BigQuery
Pub/Sub
Dataproc
Dataflow
ML Engine
IoT Core
Genomics
rowser
CREATE BUCKET
REFRESH
DELETE
SHOW INFO PANEL
Filter by prefix...
Columns
uckets
Name
Default storage class
Location
Lifecycle
Labels
Requester pays
super-188716-bucket
Regional
US-EAST1
None
Off

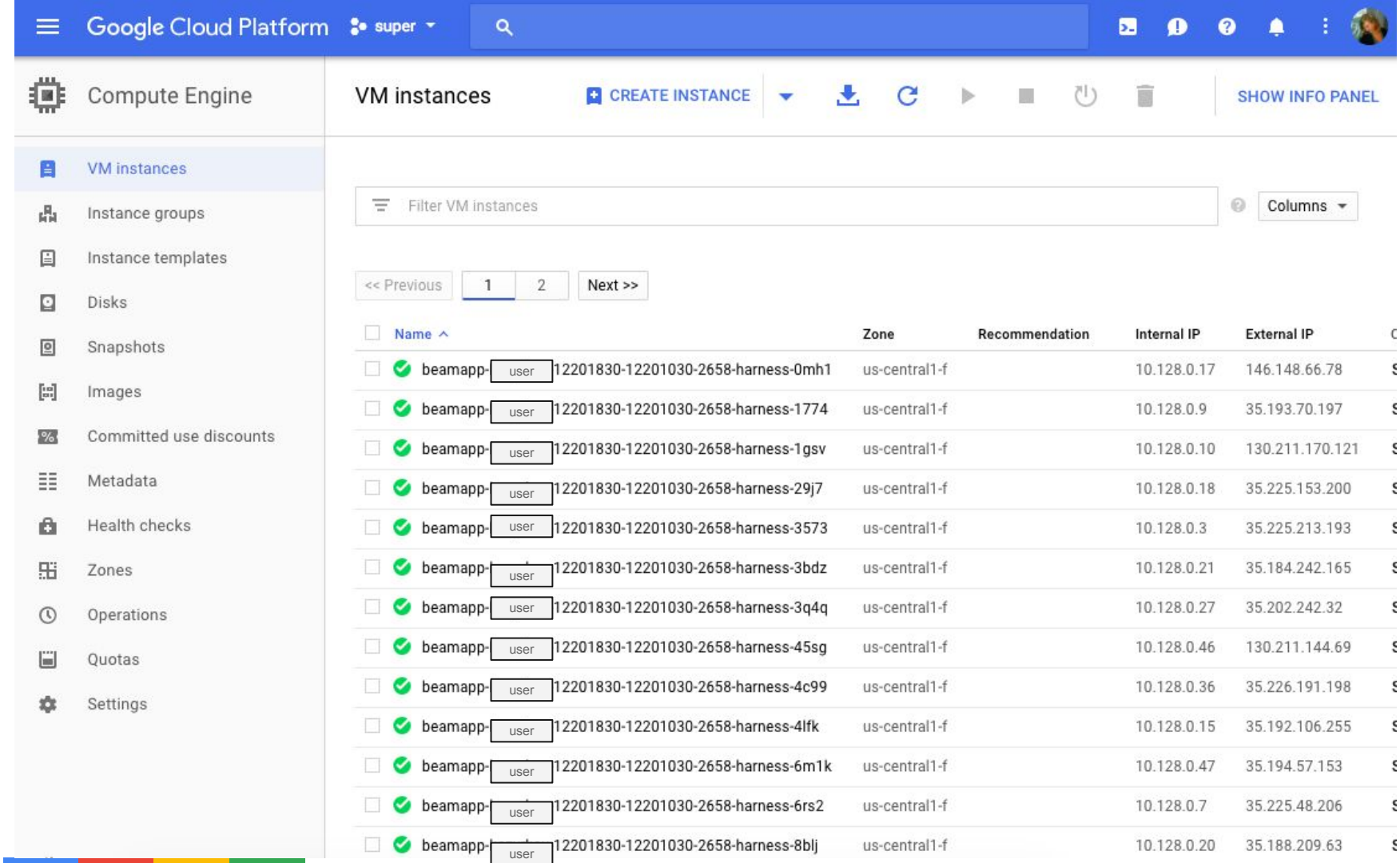
Dataflow | Jobs

+ CREATE JOB FROM TEMPLATE

Filter jobs

Name	Type	End time	Elapsed time	Start time	Status	ID	Region
beamapp-user1220183025-051783	Batch	–	4 min 49 sec	Dec 20, 2017, 1:30:30 PM	Running	2017-12-20_10_30_29-8607060276961813628	us-central1


Google Cloud Platform
super
beamapp-user-1220183025-051783
LOGS
read_rows_from_cat_in...
Running
3 min 1 sec
fetch_images_from_urls
75 elements/s
1 hr 46 min 38 sec
filter_bad_or_absent_i...
73 elements/s
1 min 31 sec
resize_and_pad_images
70 elements/s
2 min 21 sec
write_images_to_storage
71 elements/s
27 min 42 sec
Job
Job summary
Job name
beamapp-user-1220183025-051783
Job ID
2017-12-20_10_30_29-8607060276961813628
Region
us-central1
Job status
Running
Stop job
SDK version
Apache Beam SDK for Python 2.2.0
Job type
Batch
Start time
Dec 20, 2017, 1:30:30 PM
Elapsed time
5 min 5 sec
Autoscaling
Workers
50
Current state
Autoscaling: Raised the number of workers to 50 based on the rate of progress in the currently running step(s).
Dec 20, 2017 1:30 PM
60


Google Cloud Platform
super
Compute Engine
VM instances
CREATE INSTANCE
SHOW INFO PANEL
VM instances
Instance groups
Instance templates
Disks
Snapshots
Images
Committed use discounts
Metadata
Health checks
Zones
Operations
Quotas
Settings
Filter VM instances
Columns
<< Previous
1
2
Next >>
Name
Zone
Recommendation
Internal IP
External IP
beamapp-user-12201830-12201030-2658-harness-0mh1 us-central1-f 10.128.0.17 146.148.66.78
beamapp-user-12201830-12201030-2658-harness-1774 us-central1-f 10.128.0.9 35.193.70.197
beamapp-user-12201830-12201030-2658-harness-1gsv us-central1-f 10.128.0.10 130.211.170.121
beamapp-user-12201830-12201030-2658-harness-29j7 us-central1-f 10.128.0.18 35.225.153.200
beamapp-user-12201830-12201030-2658-harness-3573 us-central1-f 10.128.0.3 35.225.213.193
beamapp-user-12201830-12201030-2658-harness-3bdz us-central1-f 10.128.0.21 35.184.242.165
beamapp-user-12201830-12201030-2658-harness-3q4q us-central1-f 10.128.0.27 35.202.242.32
beamapp-user-12201830-12201030-2658-harness-45sg us-central1-f 10.128.0.46 130.211.144.69
beamapp-user-12201830-12201030-2658-harness-4c99 us-central1-f 10.128.0.36 35.226.191.198
beamapp-user-12201830-12201030-2658-harness-4lfk us-central1-f 10.128.0.15 35.192.106.255
beamapp-user-12201830-12201030-2658-harness-6m1k us-central1-f 10.128.0.47 35.194.57.153
beamapp-user-12201830-12201030-2658-harness-6rs2 us-central1-f 10.128.0.7 35.225.48.206
beamapp-user-12201830-12201030-2658-harness-8blj us-central1-f 10.128.0.20 35.188.209.63

Google Cloud Platform super

beamapp-[user]-1220183025-051783 LOGS

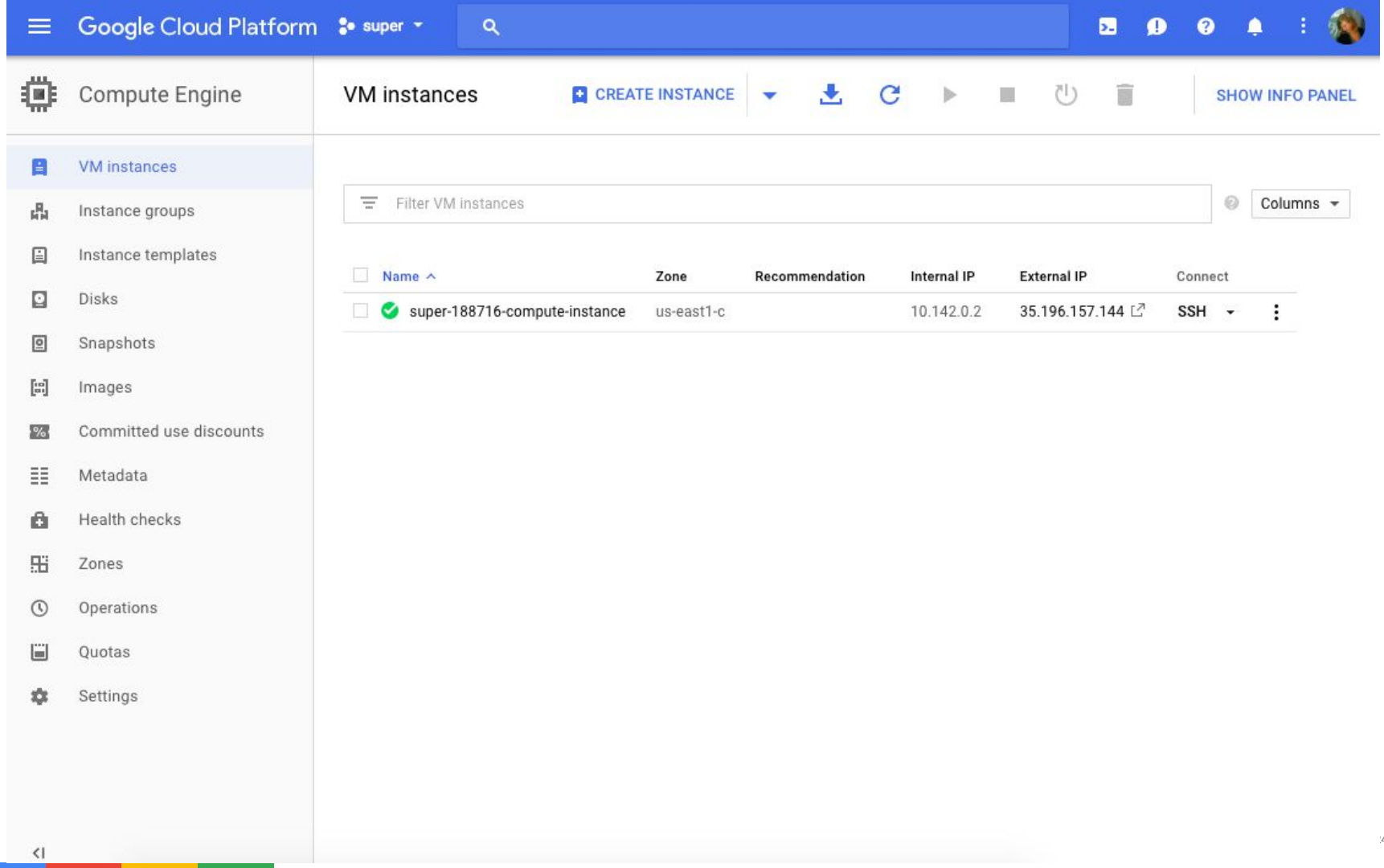
Job

Job summary

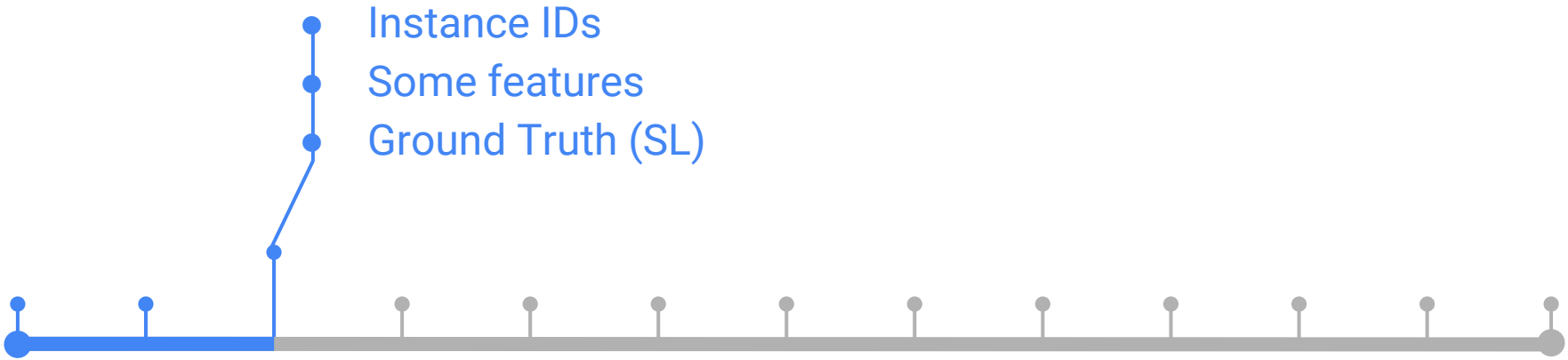
Job name	beamapp-[user]-1220183025-051783
Job ID	2017-12-20_10_30_29-8607060276961813628
Region	us-central1
Job status	Succeeded
SDK version	Apache Beam SDK for Python 2.2.0
Job type	Batch
Start time	Dec 20, 2017, 1:30:30 PM
Elapsed time	25 min 34 sec

Autoscaling

Workers	0
Current state	Autoscaling: Raised the number of workers to 50 based on the rate of progress in the currently running step(s).


Google Cloud Platform
super
Compute Engine
VM instances
CREATE INSTANCE
SHOW INFO PANEL
VM instances
Instance groups
Instance templates
Disks
Snapshots
Images
Committed use discounts
Metadata
Health checks
Zones
Operations
Quotas
Settings
Filter VM instances
Columns
Name
Zone
Recommendation
Internal IP
External IP
Connect
super-188716-compute-instance
us-east1-c
10.142.0.2
35.196.157.144
SSH

Step 2 is finished | You now have code for gathering



Google Cloud

Confidential & Proprietary

Step 2 is finished | You now have code for gathering