**BigQuery on Google Cloud Platform**

**BigQuery Product Info**

<https://cloud.google.com/bigquery/>

**Architecture**

* data is horizontally and vertically partitioned
* clusters run on dozens of thousands of machines, with hundreds of thousands of cores

**BigQuery Syntax**

<https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax>

**Public Data Sets** (maintained and paid for by Google, free to use up to 1TB of query results)

<https://cloud.google.com/bigquery/public-data/>

**Open Images Dataset**

<https://ai.googleblog.com/2016/09/introducing-open-images-dataset.html#1>

* contains approximately 9 million URLs and metadata for images that have been annotated with labels spanning more than 6,000 categories
* each image has approximately 8 labels assigned

**Overview of Selected Tables and Attributes**

*dict:*

*label\_name*

*label\_display\_name*

*images:*

*image\_id*

*original\_url*

*labels:*

*image\_id*

*label\_name*

*source*

*confidence*

**How to Access Tables in a FROM Clause**

`bigquery-public-data.*data-set-name*.*table-name*`

**Queries to Write**

Give the SQL and a screen shot of the answer (no more than the first 10 if there are many). Make sure your screen shot includes the elapsed time, GB processed, and number of rows in the answer.

1. How many labels do images have on average? (hint: you can use the division operator / in your SELECT statement)

SELECT COUNT(\*)/COUNT(DISTINCT images.image\_id)

FROM `bigquery-public-data.open\_images.images` AS images JOIN `bigquery-public-data.open\_images.labels` AS labels

ON images.image\_id = labels.image\_id

10.24 Labels/Image

1. How many labels were created by machines?

SELECT COUNT(\*)

FROM `bigquery-public-data.open\_images.labels` AS labels

WHERE source = 'machine'

81035623

1. Give the display label name, original URL, source, and confidence for a label of your choosing. (hint: use my sample query 2 with a different display label name)

select d.label\_display\_name, i.original\_url, l.source, l.confidence

from `bigquery-public-data.open\_images.dict` d,

`bigquery-public-data.open\_images.labels` l,

`bigquery-public-data.open\_images.images` i

where d.label\_name = l.label\_name

and l.image\_id = i.image\_id

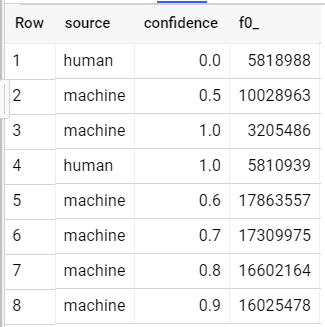
and d.label\_display\_name like '%Newt%';

1. How many images are there for each source and confidence level pairs?

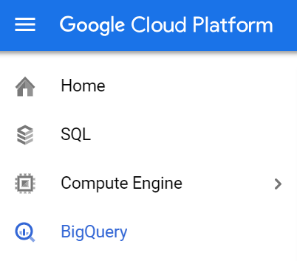
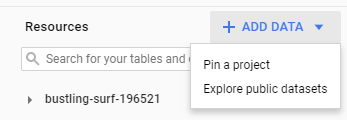
SELECT DISTINCT source, confidence, COUNT(l.image\_id)

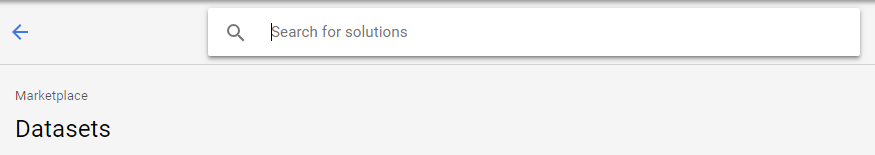
from `bigquery-public-data.open\_images.labels` l

GROUP BY source, confidence

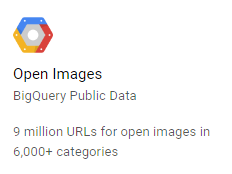
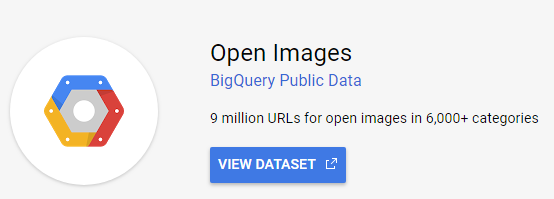


**Explore the Data Set**

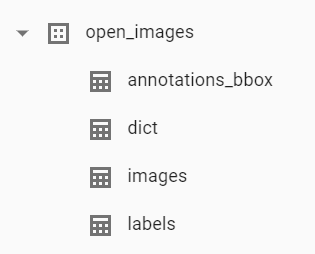
1. Select *BigQuery* in the Google Cloud Platform menu (under the 3 bars in the upper left).
2. Find *Resources* and click on the blue down arrow next to *ADD DATA.* Select *Explore public datasets*.
3. Use the search bar to search for *Open Images*.



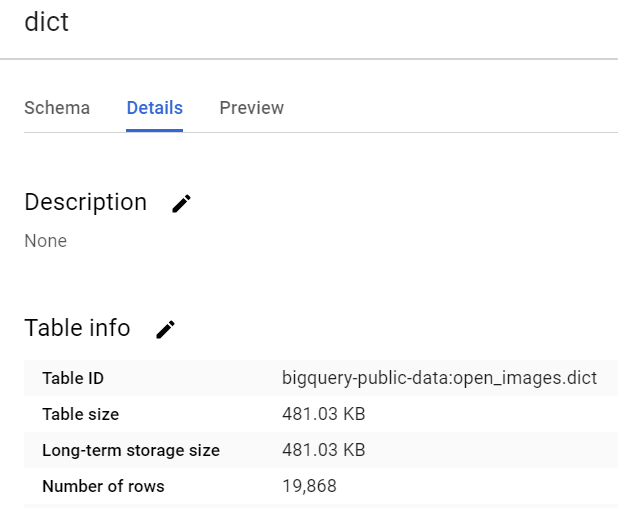
1. Select *Open Images*, then select *View Dataset*.

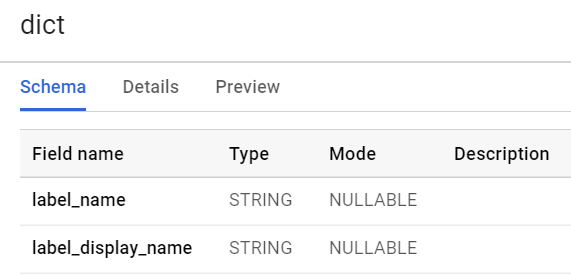
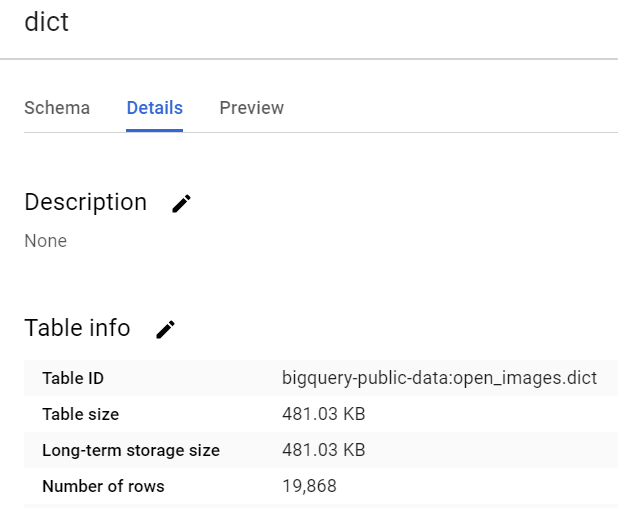
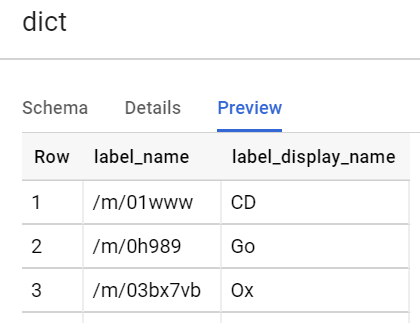


1. This will take you back to BigQuery and you’ll see a list of public data sets. Find the *open\_images* data set and expand it to show 4 tables. We will not be using the *annotations\_bbox* table.

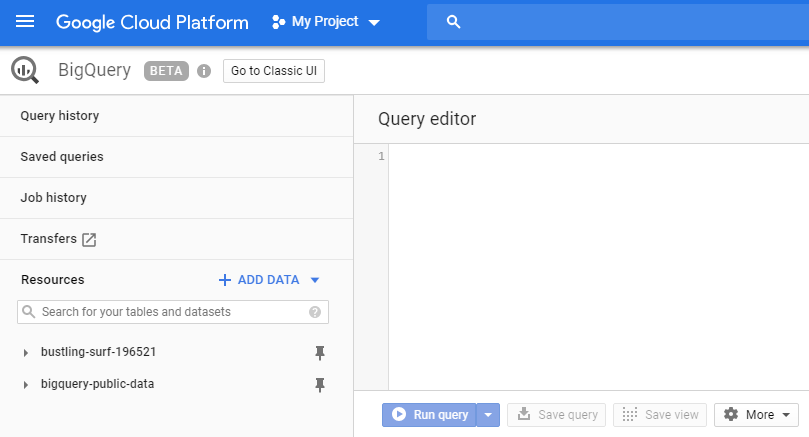


1. Click on the *dict* table (short for dictionary, maybe). Click on the *Schema*, *Details*, and *Preview* tabs. Repeat for the *images* and *labels* tables.





1. You can use the *Query editor* to type SQL queries. You can use *Query history* to see the queries that you’ve executed.



**Exploratory SQL Queries (Write, Run, and Fill in the Table Below)**

Write and run the requested sample SQL queries and fill in the data in the table below.

1. how many rows in the *dict* table?

select count(\*)

from `bigquery-public-data.open\_images.dict`;

19868 rows

1. how many rows in the *images* table?

SELECT COUNT(\*)

FROM `bigquery-public-data.open\_images.images`;

9178275 rows

1. how many rows in the *labels* table?

SELECT COUNT(\*)

FROM `bigquery-public-data.open\_images.labels`;

92665550 rows

1. how many distinct values of *source* are there?

select count (distinct source)

from `bigquery-public-data.open\_images.labels`;

2 values

1. what are the distinct values of *source*?

Human and Machine

1. what are the distinct values of *source* and *confidence*?

Human : 0.0 , 1.0

Machine: 0.5, 0.6, 0.7, 0.8, 0.9, 1.0

|  |  |  |  |
| --- | --- | --- | --- |
| **query** | **seconds elapsed** | **GB processed** | **rows in answer** |
| a. | 0.334 | 0 | 19868 rows |
| b. | 0.33 | 0 | 9178275 rows |
| c. | 0.605 | 0 | 92665550 rows |
| d. | 1.106 | 0.774 | 2 |
| e. | 1.203 | 0.774 | 2 |
| f. | 1.6 | 1.4 | 8 |

**Open Images Sample Queries to Copy and Run (Solutions Provided)**

1. How many images labeled by a machine have 100% confidence?



select count(\*) as machine\_100

from `bigquery-public-data.open\_images.labels` l

where l.source = 'machine'

and l.confidence = 1.0;

Number of Results: 3205486

2. Give the display label name, original URL, source, and confidence for images with labels like ‘Kitten.’

select d.label\_display\_name, i.original\_url, l.source, l.confidence

from `bigquery-public-data.open\_images.dict` d,

`bigquery-public-data.open\_images.labels` l,

`bigquery-public-data.open\_images.images` i

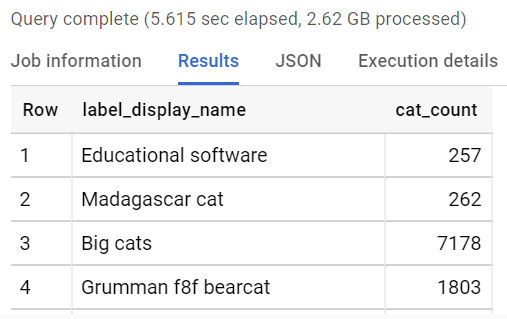
where d.label\_name = l.label\_name

and l.image\_id = i.image\_id

and d.label\_display\_name like '%Kitten%';



3. How many label display names are like cat and have more than 100 images? (the answer should have 69 rows in it)



select d.label\_display\_name, count(\*) as cat\_count

from `bigquery-public-data.open\_images.dict` d,

`bigquery-public-data.open\_images.labels` l,

`bigquery-public-data.open\_images.images` i

where d.label\_name = l.label\_name

and l.image\_id = i.image\_id

and d.label\_display\_name like '%cat%'

group by d.label\_display\_name

having cat\_count > 100;

**Additional Information**

<https://codelabs.developers.google.com/codelabs/cp100-big-query/#0>

* includes instructions for loading data from Google Cloud Storage
* can upload CSV and create the schema on the fly