Module 3

# Exploring your Data with SQL

*In this module we will:*

- **Compare Common Data Exploration Techniques**
- Learn How to Code High Quality Standard SQL
- Explore Google BigQuery Public Datasets

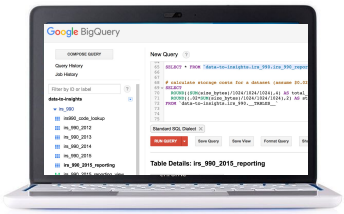
© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

In this module we will compare data exploration techniques and focus on writing SQL in BigQuery on our course dataset.

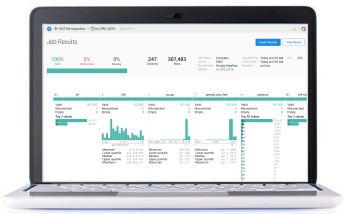
# Options for exploring a dataset are complimentary

## SQL + Web UI




- Flexible, Fast, and Familiar
- Requires SQL knowledge

## Data Preparation Tools




- GUI for Exploring Columns and Rows
- Fast Summary Statistics

## Visualization Tools

- Visually Shape and Re-Shape Quickly
- See Data a Different Way

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

The three primary options for exploring a dataset are:

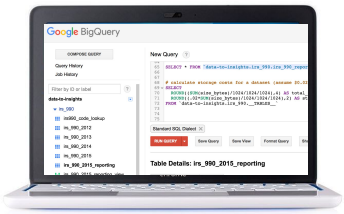
- Write some SQL and use the UI
- Use a Data Preparation or BI Tool
- Visualize your raw data with another Tool

SQL syntax hasn't changed all that significantly since the 1980s and, out of the three listed, is the most core to your skillset as a data analyst.

As for data preparation and visualization tools, we will also cover those later in this course.

# Writing SQL is a fast and familiar way to explore datasets

**SQL + Web UI**




- Flexible, Fast, and Familiar
- Requires SQL knowledge

**Data Preparation Tools**



- GUI for Exploring Columns and Rows
- Fast Summary Statistics

**Visualization Tools**



- Visually Shape and Re-Shape Quickly
- See Data a Different Way

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Let’s focus on exploring through SQL first

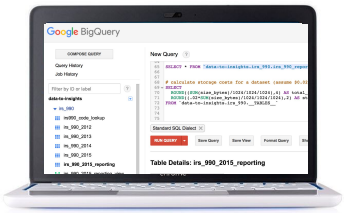
# **Steps** to Explore Data through SQL

**Question**

Give me a list of all organizations sorted by the revenue lowest to highest

Ask Good Questions

**Dataset**




Know Your Data

**SQL**
**Structured Query Language**

```
SELECT
  name,
  revenue
FROM dataset
ORDER BY revenue;
```

Write Good SQL

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Exploring a dataset through SQL is more than just writing good code. You need to know what destination you're heading towards and the general layout of your data. Good data analysts will explore how the dataset is structured even before writing a single line of code.

Questions:
What type of data am I interested in?
Financial Nonprofit Organizations

What specifically about that data?
Organization revenue

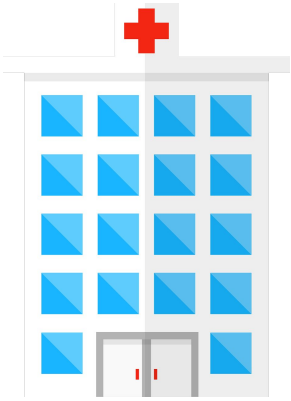
How do I want to receive the results?
Sorted by highest revenue first

Dataset:
What datasets do I have available to me? Do I need to find and upload my own?
How is the data structured? Are there multiple tables? Important fields?
How much data is there to explore?

SQL:
How do I translate my question into a SQL query?
Is my data clean?

What fields should I focus on?
Do I need to perform any aggregations?
Do I need data from multiple tables?

# Brainstorm Questions for our Nonprofit Dataset



Open this example form:
https://goo.gl/5ZWCko

What are potential insights about nonprofit organizations?

- Revenue? Expenses? Employees?
- Trends over time?

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Example Form for Kaiser Foundation Health Plan 2015:
https://goo.gl/d92L8c (only works clicking as a hyperlink, issues pasting the URL in directly to Chrome)

Image (hospital) cc0:
https://pixabay.com/get/eb36b9072af7083ed1584d05fb0938c9bd22ffd41cb1134393f4c57bae/hospital-2388239_1280.png

Module 3

# Exploring your Data with SQL

*In this module we will:*

- Compare Common Data Exploration Techniques
- **Learn How to Code High Quality Standard SQL**
- **Explore Google BigQuery Public Datasets**

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

# Enable standard SQL in your queries

**We will be using Standard SQL:**

- Add `#standardSQL` as a comment in first line of your query
- Show Options → Disable Legacy SQL
- Keep this Reference Handy: [https://cloud.google.com/bigquery/docs/reference/standard-sql/](https://cloud.google.com/bigquery/docs/reference/standard-sql/)

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Why Standard SQL?
More flexibility in the language
Strong performance benefits
DML support

Read more here:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

# Use Run Selected to only run a portion of your query

## Highlight and then Ctrl or Command (mac) + E

New Query

```
# standardSQL
SELECT
  language,
  wikimedia_project,
  title,
  SUM(views) AS views
FROM
  `bigquery-samples.wikipedia_benchmark.Wiki10B`
WHERE
  title LIKE '%Google%'
GROUP BY
  wikimedia_project,
  title,
  language;

SELECT 'just run this line - ctrl or command (mac) + E';
```

Standard SQL Dialect

RUN QUERY | Save Query | Save View | Format Query | Show Options

Run Selected

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

# Use backticks ` around table names in standard SQL

```
#standardSQL
SELECT column
FROM `project.dataset.table`
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Use backticks `` and not brackets [ ] for table names

## What is wrong with the below query?

```
#standardSQL
SELECT totrevenue
FROM `irs_990_2015`
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

What’s missing in the above query?

# Read BigQuery error messages for helpful tips

```
#standardSQL
SELECT totrevenue
FROM `irs_990_2015`
```

**Error:** Table name "irs_990_2015" cannot be resolved: dataset name is missing.

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

BigQuery will try and help you out with Query Validator errors

# What are a few things wrong with this query?

```
#standardSQL
SELECT totrevenue
FROM
`bigquery-public-data.irs_990.irs_990_2015`
```

How many Total Revenue records does this return for 2015 filings?

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

This is a wasteful query that returns only one column of revenue for all organizations' filings for 2015. Without aggregations or filters or limits or sorting, the results won't tell us anything.

# Add a LIMIT clause to limit results

```
#standardSQL
SELECT totrevenue
FROM `bigquery-public-data.
irs_990.irs_990_2015`
LIMIT 10
```

... added a LIMIT clause....

Results | Explanation | Job Information

| Row | totrevenue |
|---|---|
| 1 | 5707740 |
| 2 | 10161249 |
| 3 | 14678254 |
| 4 | 1944403 |
| 5 | 7062794 |
| 6 | 285763404 |
| 7 | 1490277 |
| 8 | 2176491 |
| 9 | 422950 |
| 10 | 161055 |

Table | JSON

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

We've limited the results but they're still pretty useless

# Add an ORDER BY clause to sort results

```
#standardSQL
SELECT totrevenue
FROM `bigquery-public-data.
irs_990.irs_990_2015`
ORDER BY totrevenue DESC
LIMIT 10
```

LIMIT is the last clause in your query. order of the clauses matters greatly in SQL.

Results | Explanation | Job Information

| Row | totrevenue |
|---|---|
| 1 | 45409123226 |
| 2 | 20796549014 |
| 3 | 11091388129 |
| 4 | 10098163008 |
| 5 | 9890722789 |
| 6 | 9475129863 |
| 7 | 9021585970 |
| 8 | 8655129029 |
| 9 | 7523260077 |
| 10 | 6740015230 |

Table | JSON

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

What does DESC mean in the context of this query? Which results will be shown first?

Now we've got the top 10 revenue (although it's a little hard to read without commas in the numbers)

# Add a FORMAT function to format results

```
#standardSQL
SELECT FORMAT("%'d",totrevenue)
FROM
`bigquery-public-data.irs_990.irs_990_2015`
ORDER BY totrevenue DESC
LIMIT 10
```

... format the presentation of the column result ...

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

FORMAT here is a SQL function (you can tell because it's wrapped around an input value -- total revenue in this case. A SQL function accepts parameters, perform some sort of action, and return a result.

d = decimal
i = integer

https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators#supported-format-specifiers

# **SQL Functions Perform Actions** on Inputs

```
FORMAT("%'d",totrevenue)
```

Function = Performs an **Action**

Parameters = **Inputs** you provide

```
SELECT FORMAT("%'d",1000)
#Returns 1,000
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

# Result: Add a FORMAT Function to **Format Results**

Results | Explanation | Job Information

| Row | f0_ | |
|---|---|---|
| 1 | 45,409,123,226 | |
| 2 | 20,796,549,014 | |
| 3 | 11,091,388,129 | |
| 4 | 10,098,163,008 | |
| 5 | 9,890,722,789 | |
| 6 | 9,475,129,863 | |
| 7 | 9,021,585,970 | |
| 8 | 8,655,129,029 | |
| 9 | 7,523,260,077 | |
| 10 | 6,740,015,230 | |

Table | JSON

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

# Use the Right Function for the Right Job

- **String Manipulation Functions - FORMAT( ) ...**
- Aggregation Functions
- Data Type Conversion Functions
- Date Functions
- Statistical Functions
- Analytic Functions
- User-defined Functions

BigQuery Functions Reference

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

We will be introducing these to you in context over the rest of this course

Data Type Conversion

Aggregation = perform calculations over a set of values (like SUM, COUNT, MIN, MAX)

String Manipulation = make every letter uppercase, pull the left 5 characters, format

Statistical = standard deviation, variance, and more

Analytic = perform aggregations over a subset or window of data

User-defined = write your own function in SQL or even Javascript

# Add a FORMAT function to format results

Results | Explanation | Job Information

| Row | f0_ | |
|---|---|---|
| 1 | 45,409,123,226 | |
| 2 | 20,796,549,014 | |
| 3 | 11,091,388,129 | |
| 4 | 10,098,163,008 | |
| 5 | 9,890,722,789 | |
| 6 | 9,475,129,863 | |
| 7 | 9,021,585,970 | |
| 8 | 8,655,129,029 | |
| 9 | 7,523,260,077 | |
| 10 | 6,740,015,230 | |

Table | JSON

wait... what happened to our column name?

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Now we've got pretty commas. But wait, what is f0_ ?

# Create aliases using the AS keyword

Results | Explanation | Job Information

| Row | revenue |
|---|---|
| 1 | 45,409,123,226 |
| 2 | 20,796,549,014 |
| 3 | 11,091,388,129 |
| 4 | 10,098,163,008 |
| 5 | 9,890,722,789 |
| 6 | 9,475,129,863 |
| 7 | 9,021,585,970 |
| 8 | 8,655,129,029 |
| 9 | 7,523,260,077 |
| 10 | 6,740,015,230 |

Table | JSON

```
SELECT totrevenue AS revenue
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Now our column name is properly showing but 45 Billion in revenue is not that useful if we don't know which Organization it belongs to. Let's add some more fields

# **Pitfall:** Unintentionally affecting the data type

Results | Explanation

| Row | revenue |
|---|---|
| 1 | 45,409,123,226 |
| 2 | 20,796,549,014 |
| 3 | 11,091,388,129 |
| 4 | 10,098,163,008 |
| 5 | 9,890,722,789 |
| 6 | 9,475,129,863 |
| 7 | 9,021,585,970 |
| 8 | 8,655,129,029 |
| 9 | 7,523,260,077 |
| 10 | 6,740,015,230 |

Table | JSON

## Beware of stylistic formatting in SQL.

Your output is now treated like a string. This makes math operations on this calculated field more difficult.

It's best to save stylistic elements for your visualization tool

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Consider doing formatting in just your final query or, better yet, in your data visualization platform

# Aliases do not exist yet when filtering in WHERE

```
#standardSQL
SELECT
  (totrevenue - totfuncexpns) AS income
FROM `bigquery-public-data.irs_990.irs_990_2015`
WHERE income > 0   ← Does not exist yet, will error in WHERE clause
ORDER BY income DESC ← Does exist, allowed in ORDER BY, GROUP BY, HAVING
LIMIT 10
```

… **Error: income undefined** …

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Newly defined aliases in the SELECT statement cannot be used yet for filtering but are allowed in ORDER BY, GROUP BY, HAVING

If you do not want to repeat the calculation in the WHERE clause, consider starting with a sub-query and then filtering (we will discuss later)

# Add new fields in SELECT clause to return more data

```
#standardSQL
SELECT
   totrevenue AS revenue
   ein,
   operateschools170cd AS is_school,
FROM
`bigquery-public-data.irs_990.irs_990_2015`
ORDER BY totrevenue DESC
LIMIT 10
```

Can you spot the errors?
Hint: It's one character

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Be sure to include a comma after each field

No trailing commas after the last column

# Add new fields in SELECT clause to return more data

```
#standardSQL
SELECT
   totrevenue AS revenue,
   ein,
   operateschools170cd AS is_school
FROM
`bigquery-public-data.irs_990.irs_990_2015`
ORDER BY totrevenue DESC
LIMIT 10
```

Correctly shown with commas

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Commas are fixed

# Add new fields in SELECT clause to return more data

Results | Explanation | Job Information

| Row | revenue | ein | is_school |
|---|---|---|---|
| 1 | 45409123226 | 941340523 | N |
| 2 | 20796549014 | 941105628 | N |
| 3 | 11091388129 | 900656139 | N |
| 4 | 10098163008 | 208295721 | N |
| 5 | 9890722789 | 900424876 | N |
| 6 | 9475129863 | 390123480 | N |
| 7 | 9021585970 | 390123480 | N |
| 8 | 8655129029 | 941196203 | N |
| 9 | 7523260077 | 912153073 | N |
| 10 | 6740015230 | 42103580 | Y |

Table | JSON

EIN (Employer Identification Number) is the unique identifier for that charity

Is_school is a flag field indicating whether that charity is a school

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Is there any potential for data validation that we could perform here on the is_school field? Yes! We could ensure all values are within the known range (Y or N)

# **Pitfall:** Using **SELECT *** to Return All Columns

**Avoid using * (star)** to return all columns.

Selecting only the columns you need greatly increases query speed and helps with readability.

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

SELECT * is performance intensive and there are better ways to explore a subset of your data.

But how will you then know which columns to choose? Is there a faster way to explore table fields?

## **Demo:** Use BigQuery UI to quickly explore schemas

In your query, hold ctrl or command (mac) to **highlight table names**

Click a **`table_name`**

View the **Schema**

Click on **Column names** to automatically add into your query

Next, Click on **Preview** to see a sample of data values

© 2019 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Use this sample query

```
#standardSQL
SELECT
  totrevenue AS revenue,
  ein,
  operateschools170cd AS is_school
FROM
  `bigquery-public-data.irs_990.irs_990_2015`
ORDER BY totrevenue DESC
LIMIT 10
```

Benefits:
Faster way to see all fields in a table
Helps prevent field misspelling

# Filter rows using the WHERE clause

```
#standardSQL
SELECT
   totrevenue AS revenue,
   ein,
   operateschools170cd AS is_school
FROM
`bigquery-public-data.irs_990.irs_990_2015`
WHERE operateschools170cd = 'Y'
ORDER BY totrevenue DESC
LIMIT 10
```

why didn't we write this as is_school = 'Y'?

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Note: You cannot use is_school in the where clause as it is not defined there yet but you could in the ORDER BY (and HAVING clause)

The 'Y' value is case-sensitive and, since it's a string, must be escaped with quotes. We will be using single quotes as our standard.

# Filter rows using the WHERE clause

Results | Explanation | Job Information

| Row | revenue | ein | is_school |
|---|---|---|---|
| 1 | 6740015230 | 42103580 | Y |
| 2 | 6000839000 | 231352685 | Y |
| 3 | 5717023246 | 941156365 | Y |
| 4 | 5569004000 | 520595110 | Y |
| 5 | 5133788413 | 135562308 | Y |
| 6 | 4623485966 | 951642394 | Y |
| 7 | 4560196033 | 416011702 | Y |
| 8 | 4477633568 | 60646973 | Y |
| 9 | 4471027733 | 135598093 | Y |
| 10 | 4368738915 | 150532082 | Y |

Table | JSON

only return school filings

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

# Use the Right Function for the Right Job

- String Manipulation Functions - FORMAT( )
- **Aggregation Functions - SUM( ) COUNT( ) AVG( ) MAX( ) ...**
- Data Type Conversion Functions
- Date Functions
- Statistical Functions
- Analytic Functions
- User-defined Functions

BigQuery Functions Reference

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

We will be introducing these to you in context over the rest of this course

Data Type Conversion

Aggregation = perform calculations over a set of values (like SUM, COUNT, MIN, MAX)

String Manipulation = make every letter uppercase, pull the left 5 characters, format

Statistical = standard deviation, variance, and more

Analytic = perform aggregations over a subset or window of data

User-defined = write your own function recipe in SQL or even Javascript

# Perform calculations over values with aggregation

```
#standardSQL
SELECT
   SUM(totrevenue) AS total_2015_revenue,
   AVG(totrevenue) AS avg_revenue,
   COUNT(ein) AS nonprofits,
   COUNT(DISTINCT ein) AS nonprofits_distinct,
   MAX(noemplyeesw3cnt) AS num_employees
FROM
`bigquery-public-data.irs_990.irs_990_2015`
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

# Perform calculations over values with aggregation

Results | Explanation | Job Information

| Row | total_2015_revenue | avg_revenue | nonprofits | nonprofits_distinct | num_employees |
|---|---|---|---|---|---|
| 1 | 2344355088288 | 7952843.417467928 | 294782 | 275077 | 787050 |

Table | JSON

More SQL Aggregation Functions

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Is it possible to round average revenue?

# Embed functions inside of other functions

```
#standardSQL
SELECT
   SUM(totrevenue) AS total_2015_revenue,
   ROUND(AVG(totrevenue),2) AS avg_revenue,
   COUNT(ein) AS nonprofits,
   COUNT(DISTINCT ein) AS nonprofits_distinct,
   MAX(noemplyeesw3cnt) AS num_employees
FROM
`bigquery-public-data.irs_990.irs_990_2015`
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

ROUND( ) is actually a math function

Not only can you stuff functions inside each other but you can also stuff SQL queries inside each other. More on that later in the course!

# Embed functions inside of other functions

Results | Explanation | Job Information

| Row | total_2015_revenue | avg_revenue | nonprofits | nonprofits_distinct | num_employees |
|---|---|---|---|---|---|
| 1 | 2344355088288 | 7952843.42 | 294782 | 275077 | 787050 |

Table | JSON

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

# Investigate uniqueness with COUNT(DISTINCT field)

Results | Explanation | Job Information

| Row | total_2015_revenue | avg_revenue | nonprofits | nonprofits_distinct | num_employees |
|---|---|---|---|---|---|
| 1 | 2344355088288 | 7952843.42 | 294782 | 275077 | 787050 |

Table | JSON

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Why is there a difference between count and count distinct? This means duplicates are present for our function field: ein.

# Create aggregation groupings with GROUP BY

```
#standardSQL
SELECT
  ein, # not aggregated
  COUNT(ein) AS ein_count # aggregated
FROM
`bigquery-public-data.irs_990.irs_990_2015`
GROUP BY ein
ORDER BY ein_count DESC
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Let's count the occurrences of each EIN in our 2015 filings dataset.

Here we include ein (a non aggregated field) and the COUNT(ein) which is an aggregation.

Any non aggregated fields in your select statement must be included in your GROUP BY clause. In our case, this is ein.

# **Pitfall:** Forgetting to **Group Non-Aggregated Fields**

Do not forget to **use a GROUP BY** if you are using a mix of aggregated and non-aggregated fields

```
SELECT
   company,
   SUM(revenue) AS total
FROM table
GROUP BY company
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Note that you can group by an index instead of the column name if you prefer (e.g. 1). It is general best practice to order the fields in your select statement and put aggregations either at the beginning or end to enhance readability.

# Locate duplicate records with COUNT + GROUP BY

Results | Explanation | Job Information

| Row | ein | ein_count |
|---|---|---|
| 1 | 680612926 | 7 |
| 2 | 830345294 | 7 |
| 3 | 256412267 | 7 |
| 4 | 364710020 | 7 |
| 5 | 362235151 | 7 |
| 6 | 611749317 | 7 |
| 7 | 20430344 | 7 |
| 8 | 582535749 | 7 |
| 9 | 262946183 | 7 |
| 10 | 262152334 | 7 |
| 11 | 20786314 | 7 |
| 12 | 161387890 | 7 |
| 13 | 208367574 | 7 |

Table | JSON

There are many charities that have more than one record for the tax filing year 2015. This is highly unusual.

Next let's count how often this happens in total

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Looks like some EINs have 7 records in the 2015 tax filings table. This is unusual.

Can we count just the total EINs with duplicates (e.g. exclude those with an ein_count of only 1)?

# Filter aggregations with HAVING clause

```
#standardSQL
SELECT
  ein,
  COUNT(ein) AS ein_count
FROM
`bigquery-public-data.irs_990.irs_990_2015`
GROUP BY ein
HAVING ein_count > 1
ORDER BY ein_count DESC
```

Notice that we can use the alias here for filtering

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Yes - introducing the HAVING clause which allows you to filter on aggregations.

Much like the WHERE clause filters out rows before aggregation, the HAVING clause filters out aggregated rows after aggregation

# 17,994 EINs with more than one filing in 2015

Results | Explanation | Job Information

| Row | ein | ein_count |
|---|---|---|
| 1 | 262946183 | 7 |
| 2 | 352123470 | 7 |
| 3 | 363645106 | 7 |
| 4 | 341565721 | 7 |
| 5 | 651192755 | 7 |
| 6 | 611749317 | 7 |
| 7 | 208367574 | 7 |
| 8 | 261511165 | 7 |
| 9 | 364710020 | 7 |
| 10 | 20430344 | 7 |
| 11 | 591800795 | 7 |
| 12 | 256412267 | 7 |
| 13 | 582535749 | 7 |

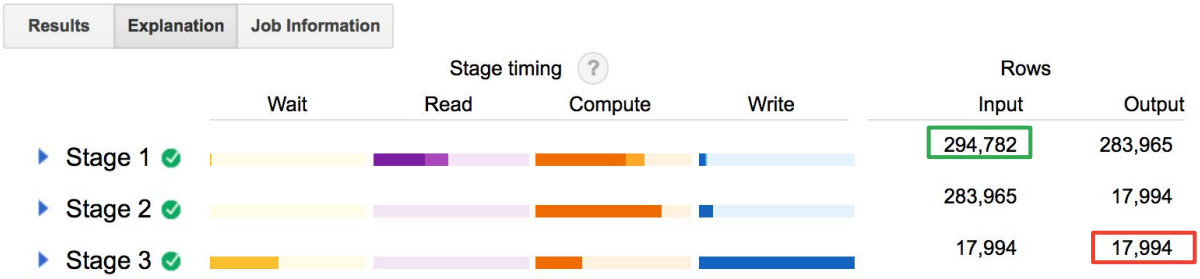
Table | JSON

First < Prev Rows 1 - 13 of 17994

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

# Track input and output counts with Explanation tab




© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

A different way to check how many records we are processing is by clicking on the Explanation tab in the BigQuery UI after running a query.

We started with 294K nonprofits for 2015 filings and filtered down to roughly 18K that have more than one filing for 2015.

The query stages represent how BigQuery mapped out the work required to perform the query job. We'll discuss this more in the Performance Module.

# Explore further by filtering on one nonprofit

Results | Explanation | Job Information

| Row | ein | ein_count |
|---|---|---|
| 1 | 262946183 | 7 |
| 2 | 352123470 | 7 |
| 3 | 363645106 | 7 |
| 4 | 341565721 | 7 |
| 5 | 651192755 | 7 |
| 6 | 611749317 | 7 |
| 7 | 208367574 | 7 |
| 8 | 261511165 | 7 |
| 9 | 364710020 | 7 |
| 10 | 20430344 | 7 |
| 11 | 591800795 | 7 |
| 12 | 256412267 | 7 |
| 13 | 582535749 | 7 |

Table | JSON

First < Prev Rows 1 - 13 of 17994

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Let’s drilldown on one specific ein to see what’s going on with these 7 filings for 2015.

# Explore further by filtering on one nonprofit

```
#standardSQL
SELECT *
FROM
`bigquery-public-data.irs_990.irs_990_2015`
WHERE ein = 262152334
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Note that we're using all columns because we the row count (7) is known.

The EIN is stored as an integer which is why we don't need to escape WHERE clause matching operator

# Data quality Insight: Multiple tax periods filed in 2015

Results | Explanation | Job Information

| Row | elf | ein | tax_pd | subseccd | s501c3or4947a1cd |
|---|---|---|---|---|---|
| 1 | P | 262152334 | 200912 | 3 | N |
| 2 | P | 262152334 | 201112 | 3 | N |
| 3 | P | 262152334 | 201312 | 3 | N |
| 4 | P | 262152334 | 200812 | 3 | N |
| 5 | P | 262152334 | 201412 | 3 | N |
| 6 | P | 262152334 | 201212 | 3 | N |
| 7 | P | 262152334 | 201012 | 3 | N |

Table | JSON

**Question:**
What could we do if we only wanted the latest 2015 filing (for tax period 2014)?

There are multiple right answers.

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

How could you extract just the rows for 2014?

Possible options:

Take only the MAX tax_pd record for each (but may end up not being 2014)

Parse out and convert the tax period as a true date and then EXTRACT the year (and filter the year for 2014)

Use a substring to pull out year (good short term, bad long term solution)

# Example Solution: parse, convert, filter for 2014

```
#standardSQL
SELECT
  ein,
  tax_pd,
  PARSE_DATE('%Y%m', CAST(tax_pd AS STRING)) AS tax_period
FROM `bigquery-public-data.irs_990.irs_990_2015`
WHERE
  EXTRACT(YEAR FROM
     PARSE_DATE('%Y%m', CAST(tax_pd AS STRING))
    ) = 2014
LIMIT 10
```

Tax filings are for the year prior
(2015 means for the period 2014)

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

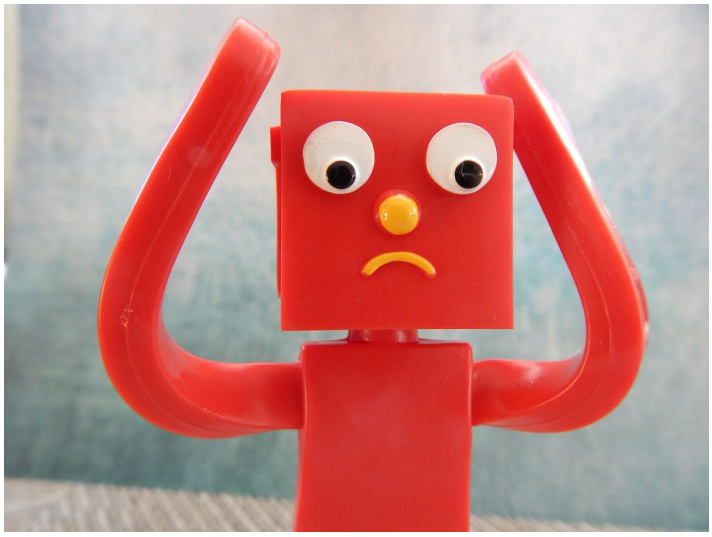
Google Cloud

Our query is starting to get a little verbose (do you see that we have to duplicate the cleansing logic?) It may be easier to clean the table and store it as a new permanent table and then query off of it. We'll cover cleaning data and storing new tables shortly.

**For now** let's look at the two new pieces we just introduced: CAST and Date Functions (like PARSE and EXTRACT)

Formatting for Dates:
https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators#supported-format-elements-for-date

# Parsing and converting dates can be hard



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Working with dates can be frustrating

Image (red guy) cc0: https://pixabay.com/en/upset-sad-confused-figurine-534103/

# Use the Right Function for the Right Job

- String Manipulation Functions - FORMAT( )
- Aggregation Functions - SUM( ) COUNT( ) AVG( ) MAX( )
- **Data Type Conversion Functions - CAST( ) ...**
- **Date Functions - PARSE_DATETIME( ) ...**
- Statistical Functions
- Analytic Functions
- User-defined Functions

BigQuery Functions Reference

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

We will be introducing these to you in context over the rest of this course

Data Type Conversion

Aggregation = perform calculations over a set of values (like SUM, COUNT, MIN, MAX)

String Manipulation = make every letter uppercase, pull the left 5 characters, format

Statistical = standard deviation, variance, and more

Analytic = perform aggregations over a subset or window of data

User-defined = write your own function recipe in SQL or even Javascript

## Comparing BigQuery data types

### Numeric Data

**Integer** (int64)
Whole numbers that can be negative (-2,-1,0,1,2,3,4)

**Float** (float64)
(1.0000000001)

### String Data

**Strings**
Text values ('dog', 'cat', '800-999-9999')

In SQL, use single quotes when dealing with strings like 'Google Inc.'

### Dates

**Dates** (datetime)
Stored in universal time format. Allowable Range: 0001-01-01 00:00:00 to 9999-12-31 23:59:59.999999

### Other

**Boolean** (Y/N)

**Array** ['apple', 'pear']

**Struct**<apple string>

BigQuery Data Types Reference

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

BigQuery data types
https://cloud.google.com/bigquery/docs/reference/standard-sql/data-types

# Using CAST to convert between data types

- `CAST("12345" AS INT64)`
  - `12345`
- `CAST("2017-08-01" AS DATE)`
  - `2017-08-01`
- `CAST(1112223333 AS STRING)`
  - `"1112223333"`
- `SELECT SAFE_CAST("apple" AS INT64)`
  - `NULL`

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

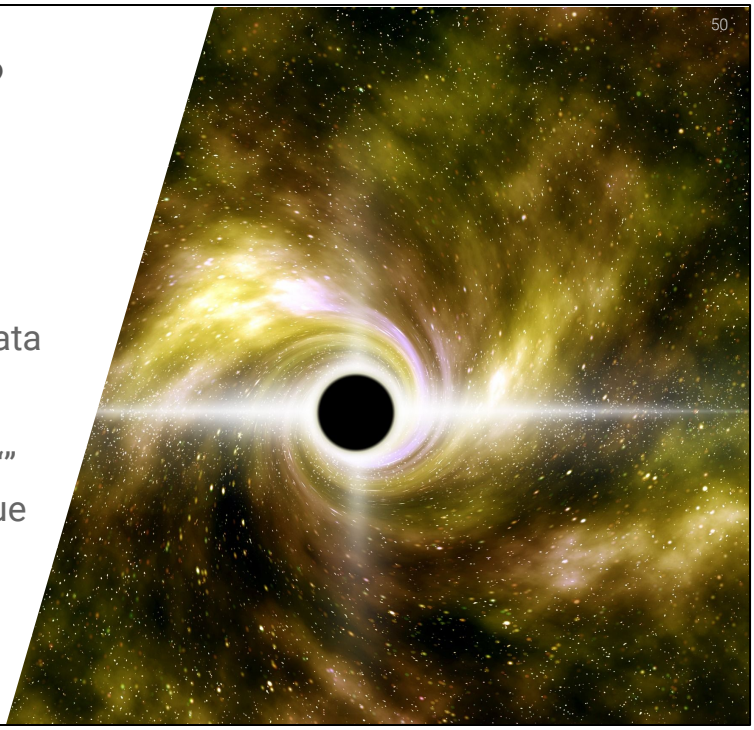
Google Cloud

Casting allows you to **treat** one data type as another. This is particularly useful if a Function expects to see an input in a specific datatype. Example, our PARSE_TIMESTAMP( ) expects a string input which is why we had to convert tax_period to a string first.

Although rare, you may run across safe_cast if you’re unsure if a column may contain “apple” in addition to “123” and “456”. If you tried cast( ) normally, this would error out. Now it just produces a NULL for that value.

# What is a NULL value?

- NULLs are valid values
- NULL is the absence of data or an empty set
- NULL is not the same as “” or a valid blank string value



Much like black holes are the absence of light, NULLs are the absence of data

NULLs may or may not be included in aggregations. By default NULLs are respected.

We will cover equivalency of NULLs when we discuss data cleanup. You cannot simply do IF VALUE = NULL since NULLS can never be equivalent to anything, not even NULL = NULL.

https://pixabay.com/get/eb31b90c2df3003ed1534705fb0938c9bd22ffd41cb113439df5c17fa2/black-hole-2483571_1920.jpg

# Handle NULL values with extreme care

```
#standardSQL
SELECT
  ein,
  street,
  city,
  state,
  zip
FROM
  `bigquery-public-data.irs_990.irs_990_ein`
WHERE
  state IS NULL
LIMIT 10;
```

NULL values cannot be equated to and thus have a special IS NULL or IS NOT NULL test

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Note that we are now querying from the Organization Details table (irs_990_ein) which has more information on each org like where it is located.

# Valid NULLs for countries not using state field

Results | Explanation | Job Information

| Row | ein | street | city | state | zip |
|---|---|---|---|---|---|
| 1 | 980121082 | APO AE | ITALY | null | 00000-0000 |
| 2 | 371658596 | 33081 AVIANO AB | AB PN | null | 00000-0000 |
| 3 | 371658596 | 33081 AVIANO AB | AB PN | null | 00000-0000 |
| 4 | 660563303 | KINGSHILL 00851 | BRITISH VIRGIN ISLANDS | null | 00000-0000 |
| 5 | 200612879 | LONDON VICTORIA | UNITED KINGDOM | null | 00000-0000 |
| 6 | 453028683 | BRIGANTINE 08203 | ALGERIA | null | 00000-0000 |
| 7 | 300254488 | BONNE TERRE MO 63628 | ALBANIA | null | 00000-0000 |
| 8 | 980683729 | TORONTO ONTARIO M3C 3P8 | CANADA | null | 00000-0000 |
| 9 | 981177602 | WHITEHORSE YUKON Y1A 6R4 | CANADA | null | 00000-0000 |
| 10 | 200318819 | HEGELALLEE 8 POTSDAM D-13367 | GERMANY | null | 00000-0000 |

Table | JSON

Not all countries will use the state field

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

This is how nulls will appear in the BigQuery results

# YYYY-MM-DD is the expected format for dates

| Date Function | Result |
|---|---|
| CURRENT_DATE([time_zone]) | Today's date |
| EXTRACT(year FROM your_date_field) | Year of date (try changing Year to Month, Day, Week etc.) |
| DATE_ADD(DATE "2008-12-25", INTERVAL 5 DAY) | Adds an interval of time |
| SELECT DATE_SUB(DATE "2008-12-25", INTERVAL 5 DAY) | Subtracts an interval of time |
| DATE_DIFF(DATE "2010-07-07", DATE "2008-12-25", DAY) | Subtracts two dates and returns the interval |
| DATE_TRUNC(DATE '2008-12-25', MONTH) | Truncates the date (e.g. 2008-12-01) |
| FORMAT_DATETIME( ) | Formats an existing date to a different date string format |
| PARSE_DATETIME( ) | Example: Turn "12/01/2017" to a proper date "2017-12-01" |

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

YYYY-MM-DD is the expected format for a date

You can also do many useful math operations on dates like add, subtract, and take the difference of.

https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators#date-functions

# Parsing string values with string functions

- `CONCAT("12345","678")`
  - `"12345678"`
- `ENDS_WITH("Apple","e")`
  - `true`
- `LOWER("Apple")`
  - `"apple"`
- `REGEXP_CONTAINS("Lunchbox",r"^*box$")`
  - `true`

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

RegEx function examples
https://cloud.google.com/bigquery/docs/reference/standard-sql/functions-and-operators#regexp_contains

# Searching for nonprofits with ‘help’ in their name

```
#standardSQL
SELECT
  ein,
  name
FROM
  `bigquery-public-data.irs_990.irs_990_ein`
WHERE
  LOWER(name) LIKE '%help%'
LIMIT 10;
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Using a String Function on name to turn everything lower case and then finding the word ‘help’ anywhere in the name

Why the LOWER( )? If we’re unsure whether the field is uppercase, lowercase, or a combination of the two.

# Searching for nonprofits with ‘help’ in their name

| Row | ein | name |
|---|---|---|
| 1 | 201515765 | PHELPS CHAMBER OF COMMERCE |
| 2 | 166052064 | PHELPS CEMETERY CORPORATION |
| 3 | 470407568 | PHELPS COUNTY FARM BUREAU INC |
| 4 | 473394369 | HILLPOINT HELPING HANDS |
| 5 | 471000644 | SENIOR HELP ALLIANCE |
| 6 | 471180492 | HELPING IMPROVE KIDS ENVIRONMENTS INC |
| 7 | 161669678 | HIS HELPING HANDS MINISTRIES INCORPORATED |
| 8 | 472894724 | LIL-BIT OF HELP |
| 9 | 471542706 | CALVARY CHAPEL HELPS INC |
| 10 | 464812552 | ONE HAND HELPING ANOTHER |

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Note that HELP can be inside of works as well

% means any character

_ underscore means a single character

# Searching for nonprofits starting with ‘help’

```
#standardSQL
SELECT
  ein,
  name
FROM
  `bigquery-public-data.irs_990.irs_990_ein`
WHERE
   LOWER(name) LIKE 'help%'
LIMIT 10;
```

Note the absence of the beginning wildcard %

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

We removed one wildcard % meaning the ‘help’ must be at the beginning of the organization name

Can also use
STARTS_WITH(LOWER(name),'help')

Or REGEX_CONTAINS( )

Or SUBSTR(4) = help

Many ways to string parse

# Searching for nonprofits starting with ‘help’

Results | Explanation | Job Information

| Row | ein | name |
|---|---|---|
| 1 | 471180492 | HELPING IMPROVE KIDS ENVIRONMENTS INC |
| 2 | 260903752 | HELPING HANDS OUTREACH MINISTRIESINC |
| 3 | 770640371 | HELPER MINISTRIES INTERNATIONAL INC |
| 4 | 731689263 | HELP THE HARVEST - HAWAII |
| 5 | 810867903 | HELPING HEROES THRIVE FOUNDATION |
| 6 | 474894747 | HELPING YOU GROW SUCCESSFULLY |
| 7 | 800743503 | HELPING HANDS OUTREACH |
| 8 | 811838250 | HELP FOR ME YOUTH AND ADULT SERVICES |
| 9 | 811638595 | HELPING HANDS OF CUMBERLAND COUNTY |
| 10 | 454896030 | HELPING HANDS FOUNDATION OF SURRY COUNTY INC |

Table | JSON

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Using a String Function on name to turn everything lower case and then finding the word ‘help’ anywhere in the name

Why the LOWER( )? If we’re unsure whether the field is uppercase, lowercase, or a combination of the two.

# Use the Right Function for the Right Job

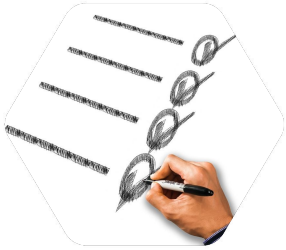
- String Manipulation Functions - FORMAT( )
- Aggregation Functions - SUM( ) COUNT( ) AVG( ) MAX( )
- Data Type Conversion Functions - CAST( )
- Date Functions - PARSE_DATETIME( )
- **Statistical Functions**
- **Analytic Functions**
- **User-defined Functions**

we will return to these three in a future module, need to build up our fundamentals first

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

# Summary: Explore your dataset with BigQuery and SQL

Explore large datasets quickly with SQL in the BigQuery UI



Comment your code and use #standardSQL for best results

Fix common SQL syntax errors by using the validator

Practice your SQL skills on over 50+ Public Datasets with example queries

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Looking back, this is the first module where we really dive into some of the core capabilities of BigQuery. As you have seen exploring massive datasets with SQL is what the tool does best and what you will need to master as a data analyst. Be sure to use standard sql mode and make best friends with the validator (it will also help you price out the bytes you will process -- which we'll cover later). Lastly, if you're looking for other datasets to explore, complete with example SQL queries, search for BigQuery Public Datasets to see the full list.

Let's test your knowledge with an interactive lab where we've written some pretty ugly and broken queries for you to fix on our IRS dataset.

Links:
https://cloud.google.com/public-datasets/

Image (check list) cc0:
https://pixabay.com/en/checklist-hand-leave-paint-hatch-2470307/
Image (check mark) cc0:
https://pixabay.com/en/successful-consent-check-mark-hook-2653403/
Image (aircraft) cc0: https://pixabay.com/en/aircraft-jet-landing-cloud-537963/
Image (taxi cab) cc0: https://pixabay.com/en/taxi-cab-traffic-cab-new-york-381233/
Image (weather) cc0: https://pixabay.com/en/lightning-storm-weather-sky-399853/
Image (doctor) cc0: https://pixabay.com/en/doctor-medical-medicine-health-563428/

# Lab 2
# Troubleshoot Common SQL Errors with BigQuery

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Lab 2 in Qwiklabs

# Troubleshoot Common SQL Errors with BigQuery

In this lab you will be faced with several broken SQL queries that are returning common errors or incorrect results in Google BigQuery.

You will practice your SQL debugging skills to fix these queries.

Identifying and fixing bugs and documenting your code are fundamental best practices

Google Cloud

Image (bug) cc0: https://pixabay.com/en/strip-bug-bug-macro-insect-red-812767/