



# How to use Google Cloud in a datathon

Google Cloud



*Use this slide if you have one analysis project for all teams*

All teams will be saving their work in this project

# Google Cloud project structure

Data-hosting project

*insert name of project here*

Google Cloud Storage

Data bucket 1

Data bucket 2

Data bucket 3

Google Cloud BigQuery analytical warehouse

Dataset 1

Dataset 2

Dataset 3



Analysis project

*insert name of project here*

SQL through BigQuery

R through RStudio

python in Jupyter notebooks through Colab



*You can delete the extra Data buckets and Datasets, so that this better reflects your datathon data-hosting project*

*Use this slide if you have one analysis project for each team*

Each team will be saving their work in a project like this

# Google Cloud project structure

Data-hosting project

*insert name of project here*

Google Cloud Storage

Data bucket 1

Data bucket 2

Data bucket 3

Google Cloud BigQuery analytical warehouse

Dataset 1

Dataset 2

Dataset 3



Analysis project

*insert names of team projects here*

SQL through BigQuery

R through RStudio

python in Jupyter notebooks through Colab



*You can delete the extra Data buckets and Datasets, so that this better reflects your datathon data-hosting project*

# What is BigQuery?

Google Cloud Platform's enterprise data warehouse for analytics

**Save**  
Fully-managed service with low TCO

**Secure**  
Data encrypted at rest and in motion with fine-grained access control



## Simple

- Analyze data from across Google Cloud Platform using SQL

## Scalable

- Multi-petabyte horizontal scalability, high performance

## Shareable

- Public and commercial datasets to enhance analysis

BI/Visualisation

tableau

looker

MicroStrategy

Qlik Q

ZOOMDATA

iCharts

CHARTIO

Yellowfin

bime by zendesk

Dundas BI

OWOX

re dash

*Use this set of slides if you are using the **old** BigQuery UI*

# Old BigQuery UI

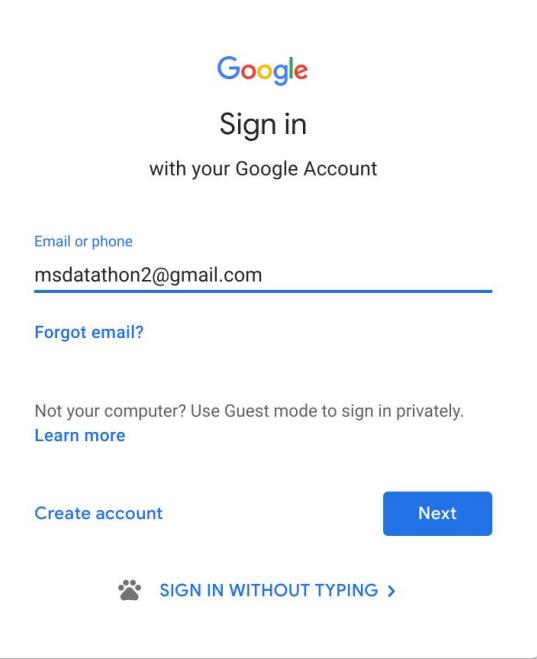
# First time in BigQuery

Enter [bigquery.cloud.google.com](https://bigquery.cloud.google.com)  
& log in with your Google ID  
(Gmail or G Suite university domain)



Sign in - Google Accounts

https://accounts.google.com/signin/v2/identifier?service=bigquery&passive=1209600&conti...



The image shows a screenshot of a web browser displaying the Google Sign-in page for BigQuery. The URL in the address bar is https://accounts.google.com/signin/v2/identifier?service=bigquery&passive=1209600&conti... The page features the Google logo and the text "Sign in with your Google Account". A text input field is pre-filled with the email "msdatathon2@gmail.com". Below the input field are links for "Forgot email?" and "Not your computer? Use Guest mode to sign in privately. Learn more". At the bottom, there are buttons for "Create account" and "Next", and a link for "SIGN IN WITHOUT TYPING >". The page also includes standard footer links for English (United States), Help, Privacy, and Terms.

English (United States) ▾

Help Privacy Terms

# First time in BigQuery

Log in with your Google ID  
(Gmail or G Suite university domain)



Sign in - Google Accounts

https://accounts.google.com/signin/v2/sl/pwd?service=bigquery&passive=1209600&continu...

Google Welcome

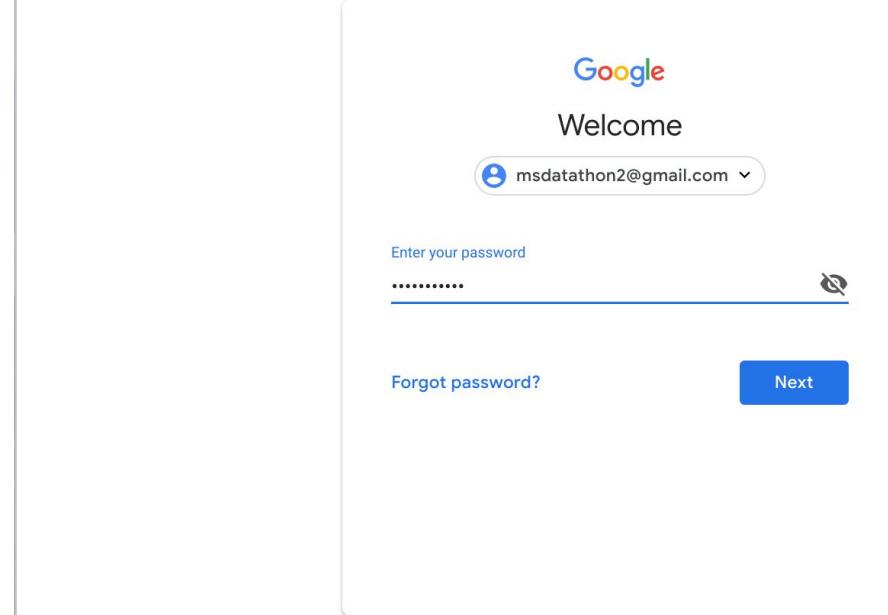
msdatathon2@gmail.com

Enter your password

.....

Forgot password? Next

English (United States) ▾ Help Privacy Terms

A screenshot of a Google sign-in page for BigQuery. The page features the Google logo and the word "Welcome". A user's email address, "msdatathon2@gmail.com", is displayed in a dropdown menu. Below the email is a password input field with placeholder text "Enter your password" and a redacted password. To the right of the input field is a "Next" button. At the bottom of the page are links for "Forgot password?", "Help", "Privacy", and "Terms". The status bar at the top shows the URL "https://accounts.google.com/signin/v2/sl/pwd?service=bigquery&passive=1209600&continu...".

# First time in BigQuery

Log in with your Google ID  
(Gmail or G Suite university domain)



Google BigQuery

https://bigquery.cloud.google.com/welcome

Try the new UI

?

ms

Google BigQuery

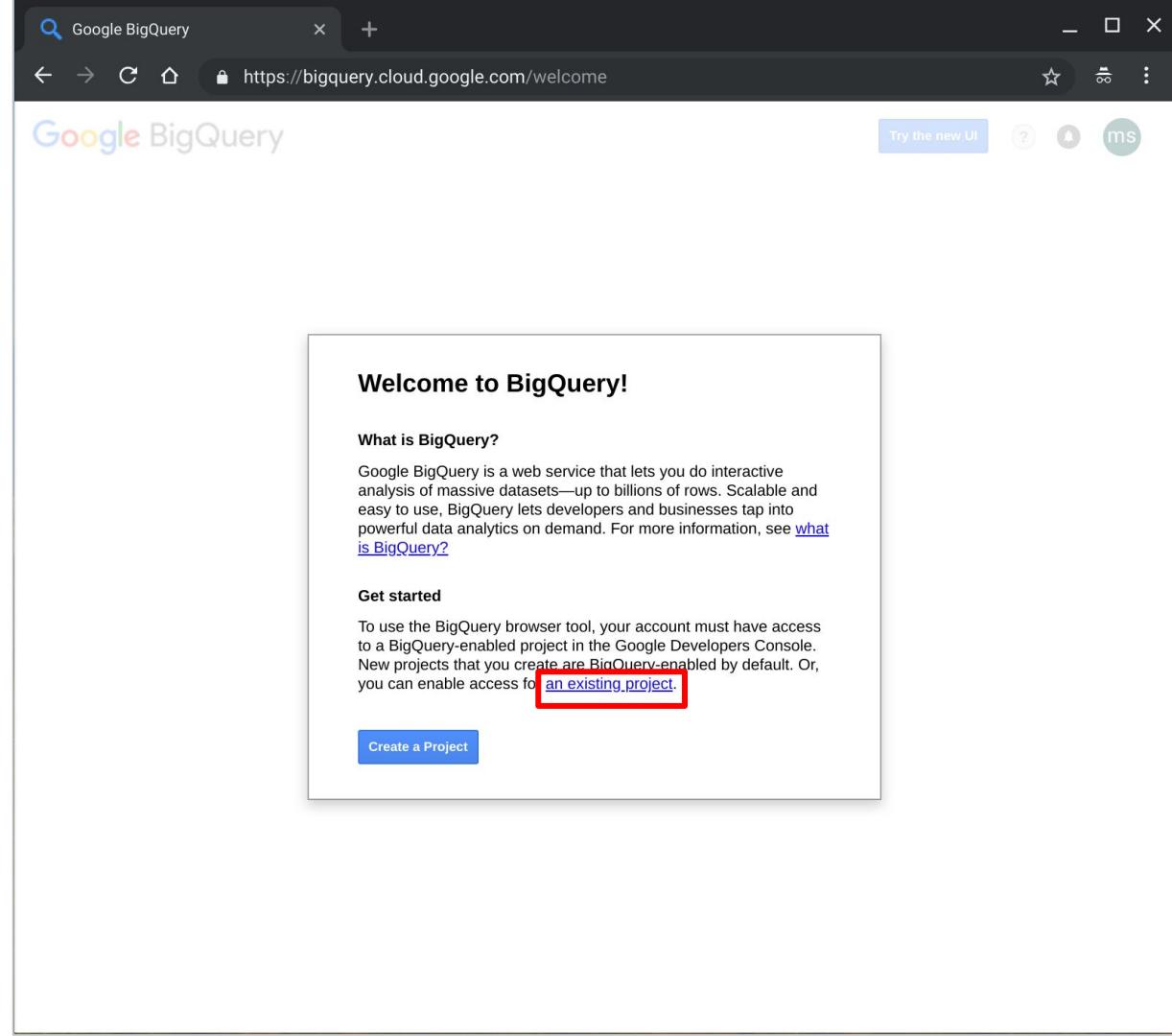
Loading BigQuery...

Waiting for bigquery.cloud.google.com...

A screenshot of a web browser window titled "Google BigQuery". The URL in the address bar is "https://bigquery.cloud.google.com/welcome". On the right side of the page, there is a blue button labeled "Try the new UI", a question mark icon, a bell icon, and a green circular icon with the letters "ms". The main content area shows the text "Loading BigQuery..." above a blue progress bar. At the bottom of the page, a dark bar displays the text "Waiting for bigquery.cloud.google.com...".

# First time in BigQuery

Access an existing project



The screenshot shows a web browser window for Google BigQuery. The address bar displays the URL <https://bigquery.cloud.google.com/welcome>. The main content area features the Google BigQuery logo and a large white box containing the "Welcome to BigQuery!" message. Inside this box, there are sections for "What is BigQuery?", "Get started", and a "Create a Project" button. A red box highlights the link "an existing project." under the "Get started" section.

Google BigQuery

Welcome to BigQuery!

**What is BigQuery?**

Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand. For more information, see [what is BigQuery?](#)

**Get started**

To use the BigQuery browser tool, your account must have access to a BigQuery-enabled project in the Google Developers Console. New projects that you create are BigQuery-enabled by default. Or, you can enable access for [an existing project.](#)

Create a Project

# First time in BigQuery

Accept the conditions



Enable an API - Google API Console +

>Your free trial is waiting: activate now to get \$300 credit to explore Google Cloud products. [Learn more](#)

[DISMISS](#) [ACTIVATE](#)

Google APIs Select a project  [Search](#)

Register your application for BigQuery API in Google API Console

Google API Console allows you to manage your application and monitor API usage.

You have no existing projects. A new project named "My Project" will be created.

Terms of Service  I agree to the [Google Cloud Platform Terms of Service](#), and the terms of service of [any applicable services and APIs](#).

Country of residence

Please email me updates regarding feature announcements, performance suggestions, feedback surveys and special offers.

Yes  No

[Agree and continue](#)

https://console.developers.google.com/flows/enableapi?apiid=bigquery&pli=1

# First time in BigQuery

Accept the conditions



Enable an API - Google API Cons X +

>Your free trial is waiting: activate now to get \$300 credit to explore Google Cloud products. [Learn more](#)

DISMISS ACTIVATE

Google APIs Select a project

Register your application for BigQuery API in Google API Console

Google API Console allows you to manage your application and monitor API usage.

You have no existing projects. A new project named "My Project" will be created.

**Terms of Service**

I agree to the [Google Cloud Platform Terms of Service](#), and the terms of service of [any applicable services and APIs](#).

Country of residence

United States

Please email me updates regarding feature announcements, performance suggestions, feedback surveys and special offers.

Yes  No

Agree and continue

The screenshot shows a web browser window for the Google API Console. The URL is https://console.developers.google.com/flows/enableapi?apiid=bigquery. The page title is "Enable an API - Google API Console". A banner at the top encourages activating a free trial with \$300 credit. Below the banner, there's a navigation bar with "Google APIs" and "Select a project" dropdowns, and a search bar. The main content area is titled "Register your application for BigQuery API in Google API Console" and explains that a new project named "My Project" will be created. It features two red-bordered sections: one for accepting the "Terms of Service" (which includes a checkbox for agreeing to the Google Cloud Platform Terms of Service and any applicable service APIs) and another for opting in to receive updates (with a radio button for "Yes" and one for "No", where "No" is selected). At the bottom is a blue "Agree and continue" button.

# First time in BigQuery

Enter [bigquery.cloud.google.com](https://bigquery.cloud.google.com)  
in the URL address bar again



Enable an API - Google API Cons... +

>Your free trial is waiting: activate now to get \$300 credit to explore Google Cloud products. [Learn more](#)

DISMISS ACTIVATE

Google APIs Select a project ▾

The API is enabled

The project has been created and BigQuery API has been enabled.

Next, to use the API you'll need the right credentials.

Go to credentials

https://console.developers.google.com/flows/enableapi?apiid=bigquery

# Working in BigQuery

Find the dataset you want to query  
We'll use the Medicare public dataset  
in this walkthrough

The new UI is in beta, so not all  
features work exactly the same yet:  
stick to the existing UI for the  
datathon



Google BigQuery

https://bigquery.cloud.google.com/welcome/aerobic-entropy-219517

Try the new UI ? ms

## Welcome to BigQuery!

Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand.

To get started, try one of the following options:

- Read our [BigQuery Quickstart guide](#)
- Run a query against our sample data by clicking "Compose Query"
- Create a new dataset and load some of your own data into a table using the  menu on the left
- Learn about [cost control options in BigQuery](#)
- For more information on the UI, see the [BigQuery Web UI guide](#)

COMPOSE QUERY

Query History  
Job History  
Scheduled Queries  
Transfers

Filter by ID or label

My Project

No datasets found in this project.  
Please create a dataset or select a new project from the menu above.

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
- ▶ nyc-tlc:green
- ▶ nyc-tlc:yellow

# Working in BigQuery

Find the dataset you want to query  
We'll use the Medicare public dataset  
in this walkthrough



Google BigQuery

https://bigquery.cloud.google.com/welcome/aerobic-entropy-219517

Try the new UI

?

ms

## Welcome to BigQuery!

Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand.

To get started, try one of the following options:

- Read our [BigQuery Quickstart guide](#)
- Run a query against our sample data by clicking "Compose Query"
- Create a new dataset and load some of your own data into a table using the  menu on the left

Learn about [cost control options in BigQuery](#)

[BigQuery Web UI guide](#)

My Project (aerobic-entropy-219517)

Create new dataset

Switch to project

Customer-Managed Encryption

Refresh

Display project...

Manage projects...

COMPOSE QUERY

Query History

Job History

Scheduled Queries

Transfers

Filter by ID or label

No datasets found in this project.  
Please create a dataset or select a new project from the menu above.

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](#)
- ▶ [nyc-tlc:green](#)
- ▶ [nyc-tlc:yellow](#)

# Working in BigQuery

Find the dataset you want to query  
We'll use the Medicare public dataset  
in this walkthrough

**bigquery-public-data**



The screenshot shows the Google BigQuery web interface. At the top, there's a search bar with "Google BigQuery" and a URL bar with the address <https://bigquery.cloud.google.com/welcome/aerobic-entropy-219517>. On the left, there's a sidebar with "COMPOSE QUERY" at the top, followed by "Query History", "Job History", "Scheduled Queries", and "Transfers". Below that is a "Filter by ID or label" input field and a dropdown menu labeled "My Project". The main content area has a heading "Welcome to BigQuery!" and text about its capabilities. It also lists options to get started: "Read our [BigQuery Quickstart guide](#)", "Run a query against our sample data by clicking "Compose Query"" (with a cursor icon over "Compose Query"), "Create a new dataset and load some of your own data into a table using the  menu on the left", and "Learn about [cost control options in BigQuery](#)". A modal window titled "Display Project" is open in the center. It contains a "Project ID" input field with "bigquery-public-data" typed into it, a radio button group where the first option ("Display project in navigation panel") is selected, and two buttons at the bottom: "OK" and "Cancel".

# Working in BigQuery

Find the dataset you want to query  
We'll use the Medicare public dataset  
in this walkthrough

The screenshot shows the Google BigQuery web interface. At the top, there's a search bar with 'Google BigQuery', a URL bar with 'https://bigquery.cloud.google.com/welcome/aerobic-entropy-219517', and a 'Try the new UI' button. The main header says 'Google BigQuery'. On the left, there's a sidebar with 'COMPOSE QUERY' at the top, followed by 'Query History', 'Job History', 'Scheduled Queries', and 'Transfers'. Below that is a 'Filter by ID or label' input field and a 'My Project' section which says 'No datasets found in this project. Please create a dataset or select a new project from the menu above.' A dropdown menu is open, showing 'Bigquery-public-data' which is highlighted with a red box. Underneath it, there's a 'Public Datasets' section with several entries: 'gdel-bq:hathitrustbooks', 'gdel-bq:internetarchivebooks', 'lookerdata:cde', 'nyc-tlc:green', and 'nyc-tlc:yellow'. To the right, there's a 'Welcome to BigQuery!' section with a brief introduction, a 'To get started' section with five bullet points, and a 'Read our [BigQuery Quickstart guide](#)' link.

**Welcome to BigQuery!**

Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand.

To get started, try one of the following options:

- Read our [BigQuery Quickstart guide](#)
- Run a query against our sample data by clicking "Compose Query"
- Create a new dataset and load some of your own data into a table using the  menu on the left
- Learn about [cost control options in BigQuery](#)
- For more information on the UI, see the [BigQuery Web UI guide](#)

[BigQuery Quickstart guide](#)

[cost control options in BigQuery](#)

[BigQuery Web UI guide](#)

# Working in BigQuery

Find the dataset you want to query  
We'll use the Medicare public dataset  
in this walkthrough



Google BigQuery

https://bigquery.cloud.google.com/welcome/aerobic-entropy-219517

Try the new UI

?

ms

## Google BigQuery

COMPOSE QUERY

Query History

Job History

Scheduled Queries

Transfers

Filter by ID or label

My Project

No datasets found in this project.  
Please create a dataset or select a new project from the menu above.

bigquery-public-data

- austin\_311
- austin\_bikeshare
- austin\_crime
- austin\_incidents
- austin\_waste
- baseball
- bitcoin\_blockchain
- bls
- census\_bureau\_construction
- census\_bureau\_international
- census\_bureau\_usa
- chicago\_crime
- chicago\_taxi\_trips

### Welcome to BigQuery!

Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand.

To get started, try one of the following options:

- Read our [BigQuery Quickstart guide](#)
- Run a query against our sample data by clicking "Compose Query"
- Create a new dataset and load some of your own data into a table using the  menu on the left
- Learn about [cost control options in BigQuery](#)
- For more information on the UI, see the [BigQuery Web UI guide](#)

# Working in BigQuery

Drill on the Medicare dataset to see the tables



Google BigQuery

https://bigquery.cloud.google.com/welcome/aerobic-entropy-219517

Try the new UI

COMPOSE QUERY

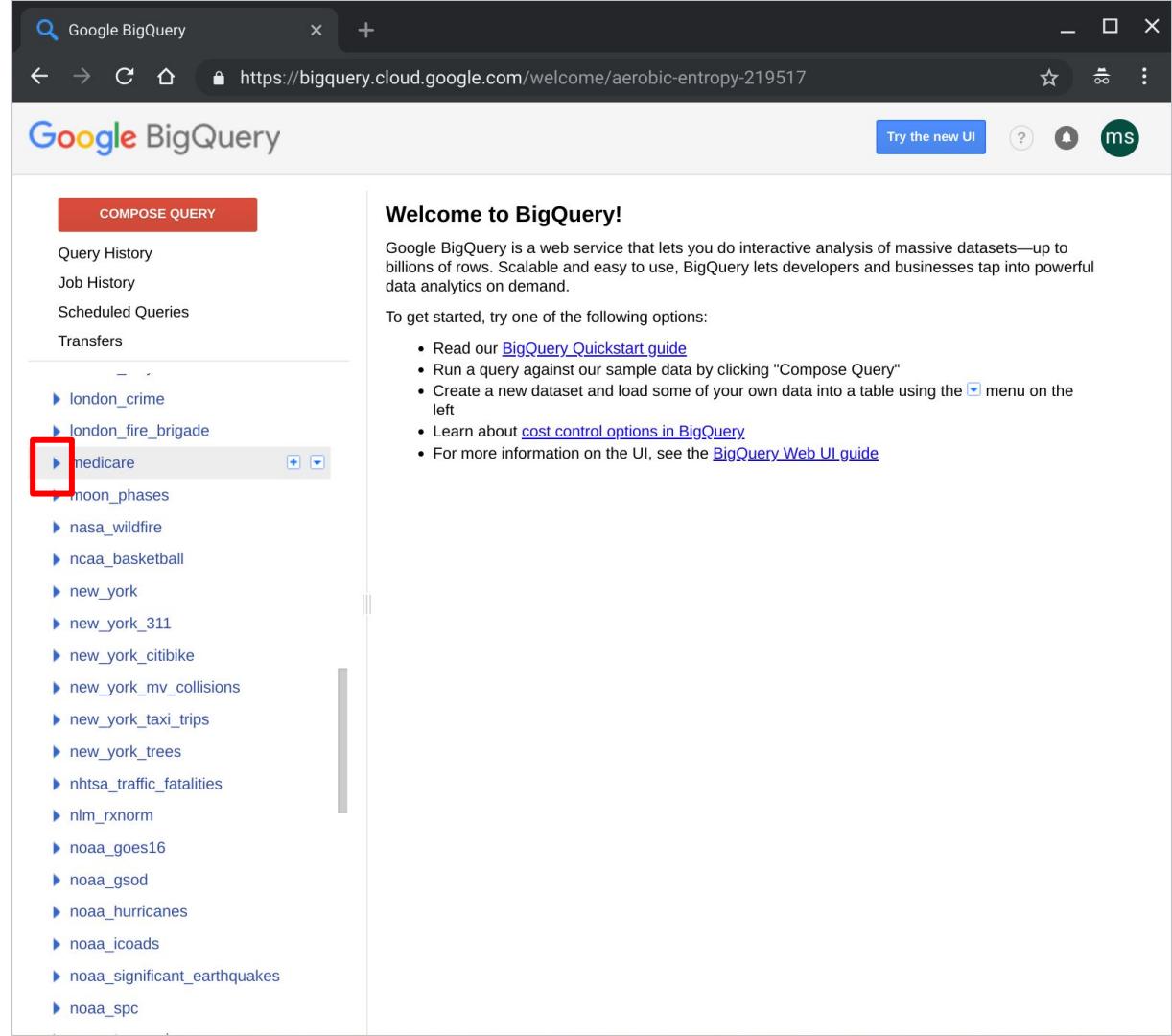
Welcome to BigQuery!

Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand.

To get started, try one of the following options:

- Read our [BigQuery Quickstart guide](#)
- Run a query against our sample data by clicking "Compose Query"
- Create a new dataset and load some of your own data into a table using the  menu on the left
- Learn about [cost control options in BigQuery](#)
- For more information on the UI, see the [BigQuery Web UI guide](#)

london\_crime  
london\_fire\_brigade  
**medicare**    
nhsn\_phases  
nasa\_wildfire  
ncaa\_basketball  
new\_york  
new\_york\_311  
new\_york\_citibike  
new\_york\_mv\_collisions  
new\_york\_taxi\_trips  
new\_york\_trees  
nhtsa\_traffic\_fatalities  
nlm\_rxnorm  
noaa\_goes16  
noaa\_gsod  
noaa\_hurricanes  
noaa\_icoads  
noaa\_significant\_earthquakes  
noaa\_spc



# Working in BigQuery

Here are the tables in the Medicare dataset

Click on a table name to see its schema



The screenshot shows the Google BigQuery web interface. At the top, there's a search bar with "Google BigQuery", a URL bar with "https://bigquery.cloud.google.com/welcome/aerobic-entropy-219517", and a navigation bar with icons for back, forward, refresh, and a dropdown menu.

The main area has a "COMPOSE QUERY" button at the top left. To its right is a sidebar with links: "Query History", "Job History", "Scheduled Queries", and "Transfers". Below these are sections for "london\_crime", "london\_fire\_brigade", and "medicare". The "medicare" section is expanded, showing a list of tables: "inpatient\_charges\_2011", "inpatient\_charges\_2012", "inpatient\_charges\_2013", "inpatient\_charges\_2014", "outpatient\_charges\_2011", "outpatient\_charges\_2012", "outpatient\_charges\_2013", "outpatient\_charges\_2014", "part\_d\_prescriber\_2014", "physicians\_and\_other\_suppli...", "physicians\_and\_other\_suppli...", "physicians\_and\_other\_suppli...", "moon\_phases", "nasa\_wildfire", "ncaa\_basketball", "new\_york", and "new\_york\_311". The table "inpatient\_charges\_2011" is highlighted with a red box.

The right side of the interface features a "Welcome to BigQuery!" section with a brief introduction and a list of starting options:

- Read our [BigQuery Quickstart guide](#)
- Run a query against our sample data by clicking "Compose Query"
- Create a new dataset and load some of your own data into a table using the  menu on the left
- Learn about [cost control options in BigQuery](#)
- For more information on the UI, see the [BigQuery Web UI guide](#)

# Working in BigQuery

Here is the schema for the `inpatient_charges_2011` table

Click on Preview to see a sample of records



Google BigQuery

Table Details: `inpatient_charges_2011`

Preview

Field	Type	Nullable	Description
<code>drg_definition</code>	STRING	REQUIRED	The code and description identifying the MS-DRG. MS-DRGs are a classification system that groups similar clinical conditions (diagnoses) and the procedures furnished by the hospital during the stay
<code>provider_id</code>	INTEGER	REQUIRED	The CMS Certification Number (CCN) of the provider billing for outpatient hospital services
<code>provider_name</code>	STRING	NULLABLE	The name of the provider
<code>provider_street_address</code>	STRING	NULLABLE	The street address in which the provider is physically located
<code>provider_city</code>	STRING	NULLABLE	The city in which the provider is physically located
<code>provider_state</code>	STRING	NULLABLE	The state in which the provider is physically located
<code>provider_zipcode</code>	INTEGER	NULLABLE	The zip code in which the provider is physically located
<code>hospital_referral_region_description</code>	STRING	NULLABLE	The Hospital Referral Region (HRR) in which the provider is physically located
<code>total_discharges</code>	INTEGER	NULLABLE	The number of discharges billed by the provider for inpatient hospital services
<code>average_covered_charges</code>	FLOAT	NULLABLE	The provider's average charge for services covered by Medicare for all discharges in the MS-DRG. These will vary from hospital to hospital because of differences in hospital charge structures
			The average total payments to all providers for the MS-DRG including the MDSRG amount, teaching...

# Working in BigQuery

Here are the first few records in the inpatient\_charges\_2011 table

Now let's query the table



Google BigQuery

Table Details: inpatient\_charges\_2011

Query Table (highlighted with a red box)

Row	drg_definition
1	418 - LAPAROSCOPIC CHOLECYSTECTOMY W/O C.D.E. W CC
2	481 - HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W CC
3	473 - CERVICAL SPINAL FUSION W/O CC/MCC
4	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC
5	065 - INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W CC
6	195 - SIMPLE PNEUMONIA & PLEURISY W/O CC/MCC
7	329 - MAJOR SMALL & LARGE BOWEL PROCEDURES W MCC
8	207 - RESPIRATORY SYSTEM DIAGNOSIS W VENTILATOR SUPPORT 96+ HOURS
9	392 - ESOPHAGITIS, GASTROENT & MISC DIGEST DISORDERS W/O MCC
10	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC
11	552 - MEDICAL BACK PROBLEMS W/O MCC
12	682 - RENAL FAILURE W MCC
13	389 - G.I. OBSTRUCTION W CC
14	308 - CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS W MCC
15	191 - CHRONIC OBSTRUCTIVE PULMONARY DISEASE W CC
16	603 - CELLULITIS W/O MCC
17	918 - POISONING & TOXIC EFFECTS OF DRUGS W/O MCC
18	640 - MISC DISORDERS OF NUTRITION,METABOLISM,FLUIDS/ELECTROLYTES W MCC
19	069 - TRANSIENT ISCHEMIA
20	870 - SEPTICEMIA OR SEVERE SEPSIS W MV 96+ HOURS

Table JSON First < Prev Rows 1 - 20 of 163065 Next > Last

# Working in BigQuery

To query the table, you write SQL

The **first time** you query a table in a session, you should uncheck the Legacy SQL dialect  
(This is one of those things that you won't have to do in the new UI)

Let's do this: Show the options



The screenshot shows the Google BigQuery web interface. On the left, a sidebar lists various datasets and tables, with "medicare" expanded to show tables like "inpatient\_charges\_2011" through "outpatient\_charges\_2014". On the right, the main area has a "New Query" tab open, displaying a SQL query: "SELECT FROM [bigquery-public-data:medicare.inpatient\_charges\_2011] LIMIT 1000". Below the query are buttons for "RUN QUERY", "Save Query", "Save View", "Format Query", and "Schedule Query". A red box highlights the "Show Options" button, which is located just below the run query button. Further down, under "Table Details: inpatient\_charges\_2011", there are buttons for Refresh, Query Table, Copy Table, Export Table, and Delete Table. The "Preview" tab is selected, showing a table with 7 rows of medical procedure codes and descriptions. The first few rows are:

Row	drg_definition	provider_id
1	418 - LAPAROSCOPIC CHOLECYSTECTOMY W/O C.D.E. W CC	20001
2	481 - HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W CC	20001
3	473 - CERVICAL SPINAL FUSION W/O CC/MCC	20001
4	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	20001
5	065 - INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W CC	20001
6	195 - SIMPLE PNEUMONIA & PLEURISY W/O CC/MCC	20001
7	329 - MAJOR SMALL & LARGE BOWEL PROCEDURES W MCC	20001

At the bottom, there are buttons for "Table" and "JSON", and links for "First", "< Prev", "Rows 1 - 7 of 163065", "Next >", and "Last".

# Working in BigQuery

Uncheck the Legacy SQL dialect



The screenshot shows the Google BigQuery web interface. On the left, there's a sidebar with links like "COMPOSE QUERY", "Query History", "Job History", "Scheduled Queries", "Transfers", and a expanded section for "medicare" containing tables such as "inpatient\_charges\_2011", "inpatient\_charges\_2012", "inpatient\_charges\_2013", "inpatient\_charges\_2014", "outpatient\_charges\_2011", "outpatient\_charges\_2012", "outpatient\_charges\_2013", "outpatient\_charges\_2014", "part\_d\_prescriber\_2014", "physicians\_and\_other\_suppli...", "physicians\_and\_other\_suppli...", "physicians\_and\_other\_suppli...", "moon\_phases", "nasa\_wildfire", "ncaa\_basketball", "new\_york", and "new\_york\_311".

The main area is titled "New Query" and contains the following code:

```
1 SELECT FROM [bigquery-public-data:medicare.inpatient_charges_2011] LIMIT 1000
```

Below the query, there are several configuration sections:

- Destination Table:** A "Select Table" button with the note "Ctrl + Enter: run query, Tab or Ctrl + Space: autocomplete. No table selected". Options include "Write if empty" (selected), "Append to table", and "Overwrite table".
- Write Preference:** Radio buttons for "Allow Large Results" (unchecked), "Flatten Results" (checked), "Use Cached Results" (checked), "Interactive" (selected), and "Batch".
- Results Size:** A dropdown menu set to "Project Default".
- Results Schema:** A dropdown menu set to "Project Default".
- Query Caching:** A dropdown menu set to "Default".
- Query Priority:** A dropdown menu set to "Unspecified".
- UDF Source URIs:** A "Edit" button.
- Maximum Bytes Billed:** A dropdown menu set to "Project Default".
- SQL Dialect:** A dropdown menu with "Legacy SQL" selected, indicated by a red box.
- Destination Encryption:** A dropdown menu set to "Default".
- Processing Location:** A dropdown menu set to "Unspecified".

At the bottom, there are buttons for "RUN QUERY", "Save Query", "Save View", "Format Query", and "Schedule Query".

**Table Details: inpatient\_charges\_2011**

Buttons for "Refresh", "Query Table", "Copy Table", "Export Table", and "Delete Table".

# Working in BigQuery

Now hide the options



Google BigQuery

COMPOSE QUERY

New Query

1 | SELECT FROM [bigquery-public-data:medicare.inpatient\_charges\_2011] LIMIT SQL

Query Editor UDF Editor

Try the new UI

Destiny Table

Write Preference

Results Size

Results Schema

Query Caching

Query Priority

UDF Source URIs

Maximum Bytes Billed

SQL Dialect

Destination Encryption

Processing Location

RUN QUERY Save View Format Query Schedule Query

Hide Options

Table Details: inpatient\_charges\_2011

Refresh Query Table Copy Table Export Table Delete Table

A screenshot of the Google BigQuery web interface. On the left, there's a sidebar with links like 'COMPOSE QUERY', 'Query History', 'Job History', 'Scheduled Queries', 'Transfers', and a list of datasets and tables. The main area shows a 'New Query' editor with a single line of SQL: 'SELECT FROM [bigquery-public-data:medicare.inpatient\_charges\_2011] LIMIT SQL'. To the right of the query are various configuration options: 'Destiny Table' (with 'Select Table' button), 'Write Preference' (radio buttons for 'Write if empty', 'Append to table', 'Overwrite table'), 'Results Size', 'Results Schema', 'Query Caching', 'Query Priority', 'UDF Source URIs', 'Maximum Bytes Billed', 'SQL Dialect' (checkbox for 'Use Legacy SQL' which is highlighted with a red box), 'Destination Encryption', and 'Processing Location' (dropdowns for 'Default' and 'Unspecified'). At the bottom are buttons for 'RUN QUERY', 'Save View', 'Format Query', and 'Schedule Query'. A 'Hide Options' button is located at the bottom left of the configuration area.

# Working in BigQuery

Finally query the table again to get the SQL form for a query in the right SQL dialect

You won't need to do this 2 step process again in this session



Google BigQuery

COMPOSE QUERY

New Query

Query Editor UDF Editor SQL

1 SELECT FROM [bigquery-public-data:medicare.inpatient\_charges\_2011] LIMIT

Standard SQL Dialect

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

Table Details: inpatient\_charges\_2011

Refresh Query Table Copy Table Export Table Delete Table

Schema Details Preview

Row	drg_definition	provider_id
1	418 - LAPAROSCOPIC CHOLECYSTECTOMY W/O C.D.E. W CC	20001 PROVIDER
2	481 - HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W CC	20001 PROVIDER
3	473 - CERVICAL SPINAL FUSION W/O CC/MCC	20001 PROVIDER
4	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	20001 PROVIDER
5	065 - INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W CC	20001 PROVIDER
6	195 - SIMPLE PNEUMONIA & PLEURISY W/O CC/MCC	20001 PROVIDER

Table JSON First < Prev Rows 1 - 6 of 163065 Next > Last

A screenshot of the Google BigQuery web interface. On the left, there's a sidebar with links like 'COMPOSE QUERY', 'Query History', 'Job History', 'Scheduled Queries', 'Transfers', and a list of datasets including 'london\_crime', 'london\_fire\_brigade', 'medicare' (which is expanded to show tables like 'inpatient\_charges\_2011', 'inpatient\_charges\_2012', etc.), 'moon\_phases', 'nasa\_wildfire', 'ncaa\_basketball', 'new\_york', and 'new\_york\_311'. The main area shows a 'New Query' window with a SQL code editor containing a simple SELECT statement. Below it is a 'Table Details' section for the 'inpatient\_charges\_2011' table, showing its schema and some sample data rows. A red box highlights the 'Query Table' button in the table details section.

# Working in BigQuery

Now you're ready to write SQL queries in BigQuery

The format for the table name is  
`project.dataset.table`

(those are back-ticks)



Google BigQuery

New Query

```
1 SELECT FROM `bigquery-public-data.medicare.inpatient_charges_2011` LIMIT 1000
```

Query Editor UDF Editor SQL

Standard SQL Dialect

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

Table Details: inpatient\_charges\_2011

Refresh Query Table Copy Table Export Table Delete Table

Schema Details Preview

Row	drg_definition	provider_id
1	418 - LAPAROSCOPIC CHOLECYSTECTOMY W/O C.D.E. W CC	20001 PROVIDER
2	481 - HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W CC	20001 PROVIDER
3	473 - CERVICAL SPINAL FUSION W/O CC/MCC	20001 PROVIDER
4	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	20001 PROVIDER
5	065 - INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W CC	20001 PROVIDER
6	195 - SIMPLE PNEUMONIA & PLEURISY W/O CC/MCC	20001 PROVIDER

Table JSON First < Prev Rows 1 - 6 of 163065 Next > Last

A screenshot of the Google BigQuery web interface. On the left, there's a sidebar with links like 'COMPOSE QUERY', 'Query History', 'Job History', 'Scheduled Queries', 'Transfers', and a list of datasets including 'london\_crime', 'london\_fire\_brigade', 'medicare' (which is expanded to show tables like 'inpatient\_charges\_2011', 'inpatient\_charges\_2012', etc.), 'moon\_phases', 'nasa\_wildfire', 'ncaa\_basketball', 'new\_york', and 'new\_york\_311'. The main area shows a 'New Query' window with a single line of SQL: 'SELECT FROM `bigquery-public-data.medicare.inpatient\_charges\_2011` LIMIT 1000'. Below the query are buttons for 'RUN QUERY', 'Save Query', 'Save View', 'Format Query', and 'Schedule Query'. Underneath the query editor is a section titled 'Table Details: inpatient\_charges\_2011' with buttons for 'Refresh', 'Query Table', 'Copy Table', 'Export Table', and 'Delete Table'. This section also includes tabs for 'Schema', 'Details', and 'Preview'. A preview table shows six rows of data with columns for 'Row', 'drg\_definition', and 'provider\_id'. The data includes various medical procedures and their corresponding provider IDs. At the bottom, there are buttons for 'Table' and 'JSON', and navigation links for 'First', '< Prev', 'Rows 1 - 6 of 163065', 'Next >', and 'Last'.

# Working in BigQuery

Type your SQL into the Query Editor

Check for confirmation that your query will execute



Google BigQuery

New Query

```
1 SELECT provider_state, SUM(total_discharges)
2 FROM `bigquery-public-data.medicare.inpatient_charges_2011`
3 GROUP BY provider_state
4 ORDER BY SUM(total_discharges) DESC
```

Valid: This query will process 1.87 MB when run.

Standard SQL Dialect

RUN QUERY Save Query Save View Format Query Schedule Query

Show Options

Table Details: inpatient\_charges\_2011

Row	drg_definition	provider_id
1	418 - LAPAROSCOPIC CHOLECYSTECTOMY W/O C.D.E. W CC	20001 PROVIDER
2	481 - HIP & FEMUR PROCEDURES EXCEPT MAJOR JOINT W CC	20001 PROVIDER
3	473 - CERVICAL SPINAL FUSION W/O CC/MCC	20001 PROVIDER
4	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	20001 PROVIDER
5	065 - INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W CC	20001 PROVIDER

Table JSON First <Prev Rows 1 - 5 of 163065 Next > Last

A screenshot of the Google BigQuery web interface. The left sidebar shows a tree view of datasets and tables, with 'medicare' expanded to show 'inpatient\_charges\_2011' selected. The main area contains a 'New Query' editor with a SQL code block. Below the code, a green box indicates the query will process 1.87 MB. A red box highlights the 'RUN QUERY' button. To the right, a preview of the 'inpatient\_charges\_2011' table is shown with 5 rows of data. A red box also highlights the top-right corner of the preview area. At the bottom, navigation links like 'Table' and 'JSON' are visible, along with page navigation controls.

# Working in BigQuery

This query shows the number of Medicare discharges for all DRGs grouped together by state



Google BigQuery

New Query

```
1 SELECT provider_state, SUM(total_discharges)
2 FROM `bigquery-public-data.medicare.inpatient_charges_2011`
3 GROUP BY provider_state
4 ORDER BY SUM(total_discharges) DESC
```

Valid: This query will process 1.87 MB when run.

Standard SQL Dialect

RUN QUERY Save Query Save View Format Query Schedule Query

Show Options Query complete (1.2s elapsed, cached)

Results Details

Download as CSV Download as JSON Save as Table Save to Google Sheets

Row	provider_state	f0
1	FL	536859
2	TX	479939
3	CA	474979
4	NY	435557
5	IL	361603
6	PA	314936
7	MI	295552

Table JSON First < Prev Rows 1 - 7 of 51 Next > Last

A screenshot of the Google BigQuery web interface. On the left, there's a sidebar with links like 'COMPOSE QUERY', 'Query History', 'Job History', 'Scheduled Queries', and 'Transfers'. Below that is a list of datasets and tables, including 'london\_crime', 'london\_fire\_brigade', 'medicare' (with sub-tables like 'inpatient\_charges\_2011' through 'outpatient\_charges\_2014', 'part\_d\_prescriber\_2014', 'physicians\_and\_other\_suppli...', and 'moon\_phases'), 'nasa\_wildfire', 'ncaa\_basketball', 'new\_york', and 'new\_york\_311'. The main area shows a 'New Query' window with a SQL code editor containing a query to sum total discharges by provider state. Below the editor, a message says 'Valid: This query will process 1.87 MB when run.' and 'Query complete (1.2s elapsed, cached)'. The results section displays a table with 7 rows of data, showing state abbreviations and their corresponding total discharges. At the bottom, there are buttons for 'Table' and 'JSON', and navigation links for the results.

# BigQuery SQL syntax

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

It's normal SQL but extremely fast  
with support for user-defined  
functions and embedded hierarchies



## Standard SQL Query Syntax

Query statements scan one or more tables or expressions and return the computed result rows.  
This topic describes the syntax for SQL queries in BigQuery.

### SQL Syntax

```
query_statement:
  [ WITH with_query_name AS ( query_expr ) [, ...] ]
  query_expr

query_expr:
  { select | ( query_expr ) | query_expr set_op query_expr }
  [ ORDER BY expression { ASC | DESC } [, ...] ]
  [ LIMIT count [ OFFSET skip_rows ] ]

select:
  SELECT [ { ALL | DISTINCT } ]
    { expression.* [ EXCEPT ( column_name [, ...] ) ] }
    [ REPLACE ( expression [ AS ] column_name [, ...] ) ]
    | expression [ [ AS ] alias ] [, ...]
  [ FROM from_item [, ...] ]
  [ WHERE bool_expression ]
  [ GROUP BY { expression [, ...] | ROLLUP ( expression [, ...] ) } ]
  [ HAVING bool_expression ]
  [ WINDOW window_name AS ( window_definition ) [, ...] ]

set_op:
  UNION { ALL | DISTINCT } | INTERSECT DISTINCT | EXCEPT DISTINCT

from_item:
  table_name [ [ AS ] alias ] [ FOR SYSTEM TIME AS OF timestamp_expression ] |
  join |
  ( query_expr ) [ [ AS ] alias ] |
  field_path |
  { UNNEST( array_expression ) | UNNEST( array_path ) | array_path }
    [ [ AS ] alias ] [ WITH OFFSET [ [ AS ] alias ] ] |
  with_query_name [ [ AS ] alias ]
}

join:
  from_item [ join_type ] JOIN from_item
  [ { ON bool_expression | USING ( join_column [, ...] ) } ]

join_type:
  { INNER | CROSS | FULL [OUTER] | LEFT [OUTER] | RIGHT [OUTER] }
```



Internal: Count: 10, Average: 3.8

Contents	
SQL Syntax	
SELECT list	
SELECT *	
SELECT expression	
SELECT expression.*	
SELECT modifiers	
Aliases	
Analytic functions	
FROM clause	
Syntax	
Subqueries	
Aliases	
JOIN types	
Syntax	
[INNER] JOIN	
CROSS JOIN	
FULL [OUTER] JOIN	
LEFT [OUTER] JOIN	
RIGHT [OUTER] JOIN	
ON clause	
USING clause	
Sequences of JOINs	
WHERE clause	
Syntax	
GROUP BY clause	
Syntax	
HAVING clause	
Syntax	
Mandatory aggregation	
ORDER BY clause	
Syntax	
Set operators	
Syntax	
UNION	
INTERSECT	
EXCEPT	
LIMIT clause and OFFSET clause	
Syntax	
WITH clause	
Aliases	
Explicit alias syntax	
Explicit alias visibility	

# Working in BigQuery

To materialize a view



Google BigQuery

Try the new UI ? ms

COMPOSE QUERY

New Query ?

Query Editor UDF Editor SQL

```
1 SELECT provider_state, SUM(total_discharges)
2 FROM `bigquery-public-data.medicare.inpatient_charges_2011`
3 GROUP BY provider_state
4 ORDER BY SUM(total_discharges) DESC
```

Valid: This query will process 1.87 MB when run.

Standard SQL Dialect X

RUN QUERY Save Query Save View Format Query Schedule Query

Show Options Query complete (1.2s elapsed, cached)

Results Details

Download as CSV Download as JSON Save as Table Save to Google Sheets

Row	provider_state	f0_
1	FL	536859
2	TX	479939
3	CA	474979
4	NY	435557
5	IL	361603
6	PA	314936
7	MI	295552

Table JSON First < Prev Rows 1 - 7 of 51 Next > Last

A screenshot of the Google BigQuery web interface. On the left, there's a sidebar with links like 'COMPOSE QUERY', 'Query History', 'Job History', 'Scheduled Queries', 'Transfers', and a list of datasets including 'london\_crime', 'london\_fire\_brigade', 'medicare' (which is expanded to show 'inpatient\_charges\_2011' through 'outpatient\_charges\_2014'), 'part\_d\_prescriber\_2014', 'physicians\_and\_other\_suppli...', 'physicians\_and\_other\_suppli...', 'physicians\_and\_other\_suppli...', 'moon\_phases', 'nasa\_wildfire', 'ncaa\_basketball', 'new\_york', and 'new\_york\_311'. The main area shows a 'New Query' section with a SQL code editor containing a query to sum total discharges by provider state. Below the code, a message says 'Valid: This query will process 1.87 MB when run.' A red box highlights the 'Save View' button in the query toolbar. The results section shows a table with 7 rows of data, including columns for Row, provider\_state, and f0\_. At the bottom, there are buttons for Table and JSON, and navigation links for the results page.

# Working in BigQuery

Save the View in your datathon analysis project, and give it a name



Google BigQuery

https://bigquery.cloud.google.com/results/aerobic-entropy-219517:US.bquijob\_25baebde\_16...

Google BigQuery

COMPOSE QUERY

New Query

Query Editor UDF Editor SQL

```
1 SELECT provider_state, SUM(total_discharges)
2 FROM `bigquery-public-data.medicare.inpatient_charges_2011`
3 GROUP BY provider_state
4 ORDER BY SUM(total_discharges) DESC
```

Try the new UI

Save View

Project: My Project (aerobic-entropy-219517)

Dataset: datathon-analysis

Table ID: results\_table\_from\_query

OK Cancel

Row	provider_state	f0
1	FL	536859
2	TX	479939
3	CA	474979
4	NY	435557
5	IL	361603
6	PA	314936
7	MI	295552

Table JSON First < Prev Rows 1 - 7 of 51 Next > Last

# Working in BigQuery

Create a view

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



Secure | https://bigquery.cloud.google.com/results/london-datathon:US.bquijob\_4e12bada\_165b4c41e56

Try the new UI

New Query

```
1 SELECT *
2 FROM `physionet-data.mimicii_clinical.admissions`
3 WHERE admission_type = 'EMERGENCY' AND discharge_location = 'DEAD/EXPIRED'
```

COMPOSE QUERY

Query History

Job History

Transfers

mimicii\_clinical

- admissions
- callout
- caregivers
- chartevents
- cptevents
- d\_cpt
- d\_icd\_diagnoses
- d\_icd\_procedures
- d\_items
- d\_labitems
- datetimeevents
- diagnoses\_icd
- drgcodes
- icustays
- inputevents\_cv
- inputevents\_mv
- labevents
- microbiologyevents
- outputevents
- patients
- prescriptions
- procedureevents\_mv
- procedures\_icd

Save View

Project: london-datathon (london-datathon)

Dataset: temp

Table ID: cohort\_emer\_dead

OK Cancel

EATHTIME	ADMISSION_TYPE	EM						
9-13T04:35:00	EMERGENCY	EM						
2-22T04:20:00	EMERGENCY	EM						
1-27T02:20:00	EMERGENCY	EM						
3-18T20:03:00	EMERGENCY	EM						
5	14750	11999	144909	2108-08-09T10:35:00	2108-08-11T04:51:00	2108-08-11T04:51:00	EMERGENCY	EM
6	12319	10069	146672	2188-02-08T11:15:00	2188-02-27T18:41:00	2188-02-27T18:41:00	EMERGENCY	EM
7	14467	11804	131416	2134-04-16T21:09:00	2134-04-19T19:43:00	2134-04-19T19:43:00	EMERGENCY	EM
8	14636	11903	137399	2157-08-25T19:50:00	2157-08-25T07:26:00	2157-08-25T07:26:00	EMERGENCY	EM
9	12144	9922	134192	2109-06-07T17:43:00	2109-06-08T14:56:00	2109-06-08T14:56:00	EMERGENCY	EM
10	14099	11503	141190	2152-04-01T10:53:00	2152-04-02T21:15:00	2152-04-02T21:15:00	EMERGENCY	EM
11	15797	12879	199845	2186-01-13T15:45:00	2186-01-25T14:38:00	2186-01-25T14:38:00	EMERGENCY	EM
12	15041	12240	124994	2134-09-22T22:52:00	2134-09-28T04:13:00	2134-09-28T04:13:00	EMERGENCY	EM

Table JSON

First < Prev Rows 1 - 12 of 5434 Next > Last

US | +1 | 17:03 | ☰

# Accessing RStudio

RStudio is hosted on Google Compute Engine

<http://35.242.128.79:8787>

Log in with your wifi credentials

The RStudio tutorial is accessed from the right pane

A screenshot of a web browser window showing an RStudio session running on Google Compute Engine. The browser tabs include "Web Authentication", "BBC - Home", "Compute Engine - london", and "RStudio". The RStudio interface shows the "Console" tab with R version 3.3.3 startup messages, the "Environment" tab indicating an empty environment, and the "Files" tab where a file named "bigrquery\_tutorial.Rmd" is listed and highlighted with a red box. The status bar at the bottom right shows "US" and the time "08:18".

Not secure | 35.242.128.79:8787

R version 3.3.3 (2017-03-06) -- "Another Canoe"  
Copyright (C) 2017 The R Foundation for Statistical Computing  
Platform: x86\_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

>

Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Home

Name	Size	Modified
bigrquery_tutorial.Rmd	9.3 KB	Sep 8, 2018, 12:13 AM
R		

US | 08:18 +2 08:18

# Working in Colab

Free Jupyter notebook environment  
in the cloud, stored on Google Drive

[https://colab.research.google.com/  
github/GoogleCloudPlatform/health  
care/blob/master/datathon/cms\\_m  
edicare/bigquery\\_tutorial.ipynb](https://colab.research.google.com/github/GoogleCloudPlatform/healthcare/blob/master/datathon/cms_medicare/bigquery_tutorial.ipynb)

Run the code



Make a copy of  
the notebook

Share it with others  
in your team

The screenshot shows a Google Colab notebook titled "Datathon Tutorial". The top navigation bar includes "File" (highlighted with a green box), "Edit", "View", "Insert", "Runtime", "Tools", and "Help". Below the menu is a toolbar with "CODE", "TEXT", "CELL UP", "CELL DOWN", "COPY TO DRIVE", "CONNECT" (with a dropdown arrow), and "EDITING". A "SHARE" button is also visible in the top right corner, highlighted with a green box. The main content area displays the title "Datathon Tutorial" and a brief description: "The aim of this tutorial is to get you familiarized with BigQuery to query/filter/aggregate/export data with Python." A section titled "Prerequisites" follows, stating that a valid Google account is required to log in to Google Cloud Platform. It also mentions that users can create an account at <https://accounts.google.com>. A note for datathon organizers is included about registering their account for data access. A "Setup" section is expanded, providing instructions on how to run queries by creating a copy of the notebook and sharing it with teammates. It also explains the "SHARE" button and the "CONNECTED" status indicator. Below the setup instructions, a code cell is shown with the following Python code:

```
# Import libraries
import numpy as np
import os
import pandas as pd
import matplotlib.pyplot as plt
```

# Accessing BigQuery with other tools

If the analysis tool you want to use does not have a native BigQuery connector, you may be able to use a JDBC/ODBC connector to connect to BigQuery.

Matlab example:

<https://github.com/GoogleCloudPlatform/healthcare/blob/master/datathon/anzics18/tutorial.matlab>

# Find the tutorials

Hosted on Github, no need for an account

[https://github.com/GoogleCloudPlatform/healthcare/tree/master/datathon/cms\\_medicare/README.md](https://github.com/GoogleCloudPlatform/healthcare/tree/master/datathon/cms_medicare/README.md)

GoogleCloudPlatform / healthcare

Code Issues 1 Pull requests 1 Projects 0 Insights

Branch: master healthcare / datathon / cms\_medicare / README.md

a-goolger Project import generated by Copybara. 1eaedd 18 hours ago

1 contributor

13 lines (7 sloc) | 1.35 KB

Raw Blame History

## Datathon Tutorials

Welcome to datathon on GCP!

We have prepared tutorials to get you started on [BigQuery](#), the tool to filter, join, aggregate and extract data from the raw datasets for analysis. In each of the tutorials, some comprehensive examples are included to show how to view the datasets, run transformations and analyze them.

- For Python users, please start from the [Python colab](#) (a copy is available in the [tutorials](#) folder as well), which is a Jupyter notebook hosted in Google Drive, and can be shared with other people for collaboration. It has the most comprehensive examples, including how to train machine learning models on the CMS Medicare dataset with [Tensorflow](#).
- For people who have experience with R, start with the [R tutorial](#), which provides an interactive interface to go through

