Connected Bradford

Data Access User Guide

Guidance

October 2023

DRAFT - Version 0v1

Version control

| Date | Version no. | Reviewed by | Comments |
| --- | --- | --- | --- |
| 12/10/23 | 0v1 |  | Initial draft |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Approval

| Approved by | Date |
| --- | --- |
|  |  |

Contents

[1 Introduction 5](#_Toc147937502)

[1.1 Introduction 5](#_Toc147937503)

[1.2 What is Connected Bradford? 5](#_Toc147937504)

[1.3 What You'll Find in this Manual 5](#_Toc147937505)

[1.4 Key Features 5](#_Toc147937506)

[1.5 Your Contribution Matters 5](#_Toc147937507)

[2 Accessing the Connected Bradford Database: 6](#_Toc147937508)

[2.1 Request Access: 6](#_Toc147937509)

[2.2 Set Up Your Google YHCR Account: 6](#_Toc147937510)

[2.3 Enable Two-Factor Authentication: 6](#_Toc147937511)

[2.4 Accessing Datasets: 6](#_Toc147937512)

[2.5 Navigating the Google Cloud Console Dashboard 6](#_Toc147937513)

[3 Accessing BigQuery 7](#_Toc147937514)

[3.1 The BigQuery Dashboard 7](#_Toc147937515)

[4 Understanding Workspaces 9](#_Toc147937516)

[4.1 Project Collaboration Area (CB\_(unique 4 digit number)): 9](#_Toc147937517)

[4.2 Your Personal Workspace ("MYSPACE\_(Your initials)"): 9](#_Toc147937518)

[4.3 Connected Bradford GitHub Collaboration 10](#_Toc147937519)

[5 Other tables 12](#_Toc147937520)

[5.1 CB\_LOOKUPS: 12](#_Toc147937521)

[5.2 CB\_FDM\_MASTER: 12](#_Toc147937522)

[6 Writing SQL Queries 13](#_Toc147937523)

[6.1 Data Export and Download 13](#_Toc147937524)

[7 Advanced Analytics - Vertex Environment 14](#_Toc147937525)

[7.1 Accessing the Vertex AI Platform 14](#_Toc147937526)

[7.2 Setting Up Vertex Notebooks 14](#_Toc147937527)

[8 Analysing Data in Vertex Notebooks 17](#_Toc147937528)

[8.1 Using R in Jupyter 17](#_Toc147937529)

[8.2 Running Python Code for Analysis 17](#_Toc147937530)

[8.3 Listing the tables in BigQuery 19](#_Toc147937531)

[8.4 Using OpenCodeList with Juypter Notebooks 19](#_Toc147937532)

[8.5 How to create a Cohort Using My Codelist 20](#_Toc147937533)

[9 Useful links 21](#_Toc147937534)

# Introduction

Welcome to the Connected Bradford User Manual!

## Introduction

Are you a dedicated researcher or analyst with approved access to the invaluable resources of the Connected Bradford database? If so, you've come to the right place. This user manual is designed to be your trusted companion on this data-driven journey.

## What is Connected Bradford?

Connected Bradford is a dynamic and comprehensive database hosted by Bradford Teaching Hospitals NHS Foundation Trust. This repository of knowledge resides within the Secure Data Environment (SDE), underpinned by the reliable infrastructure of Google Cloud.

In our rapidly evolving world, the power of data has become indispensable. Researchers and analysts like you play a pivotal role in harnessing this data to drive meaningful change. The Connected Bradford database is your gateway to a wealth of information, insights, and opportunities to make a positive impact on healthcare and beyond.

## What You'll Find in this Manual

This manual is your key to unlocking the potential of Connected Bradford. Here, you'll find detailed instructions, best practices, and tips for navigating and utilising the database effectively.

## Key Features

User Access: Learn how to access and log in to the Connected Bradford database securely via Google Cloud.

Data Exploration: Discover how to explore the wealth of data available.

Analysis Techniques: Get insights into advanced analytical methods and tools for extracting valuable information.

GitHub: Get more information on the data or analytical techniques in the Connected Bradford GitHub which is located here: <https://github.com/ConnectedBradford>

## Your Contribution Matters

Connected Bradford is not just a database; it's a community of dedicated individuals striving to make a difference. Your work here has the potential to transform healthcare, improve patient outcomes, and contribute to our collective understanding of health-related challenges.

Thank you for being part of this vital initiative. Your commitment to data-driven research is paving the way for a brighter future. Let's embark on this journey together, armed with knowledge, expertise, and a shared vision of a healthier world.

# Accessing the Connected Bradford Database:

## Request Access:

The Connected Bradford team will request access if you have an approved Expression of Interest (EOI and data sharing agreement.

You will receive an email from [jira@yhcrservice.atlassian.net](mailto:jira@yhcrservice.atlassian.net) confirming that access is being established and instructions on how to activate your Google YHCR account.

Note that the activiation link expires after 48 hours. If the link has expired, you will need to email the Connected Bradford team at [cBradford@bthft.nhs.uk](mailto:cBradford@bthft.nhs.uk) and request for it to be reset.

## Set Up Your Google YHCR Account:

Open a web browser and visit the Google Cloud Console.

Sign in with your Google YHCR account associated with Google Cloud.

## Enable Two-Factor Authentication:

Ensure you have activated your Google YHCR account.

Enable two-factor authentication for added security. Google will ask you to set-up two-factor authentication via text message of the Google Authenticator app.

## Accessing Datasets:

Once your access is approved and two-factor authentication is enabled, you can access the permitted datasets as agreed in your EOI. Further information is in Section 3 of this document.

Top of Form

## Navigating the Google Cloud Console Dashboard

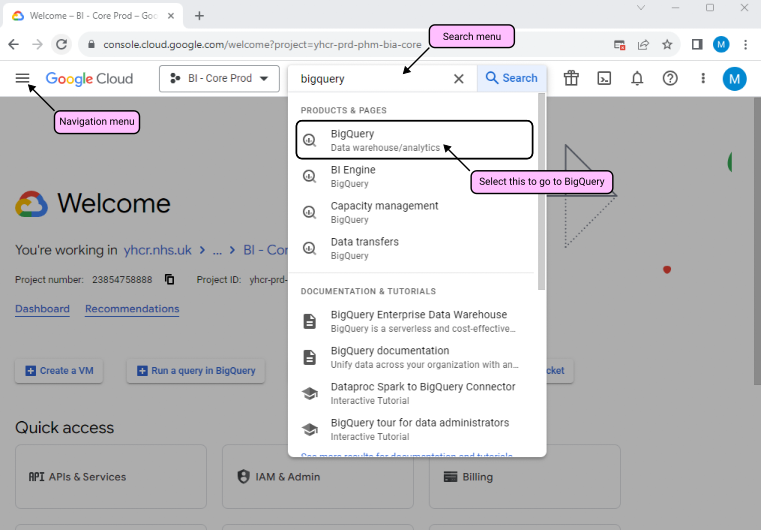
This section provides an overview of the Google Cloud Console dashboard, highlighting key components such as navigation menu, project selection, and notifications.

# Accessing BigQuery

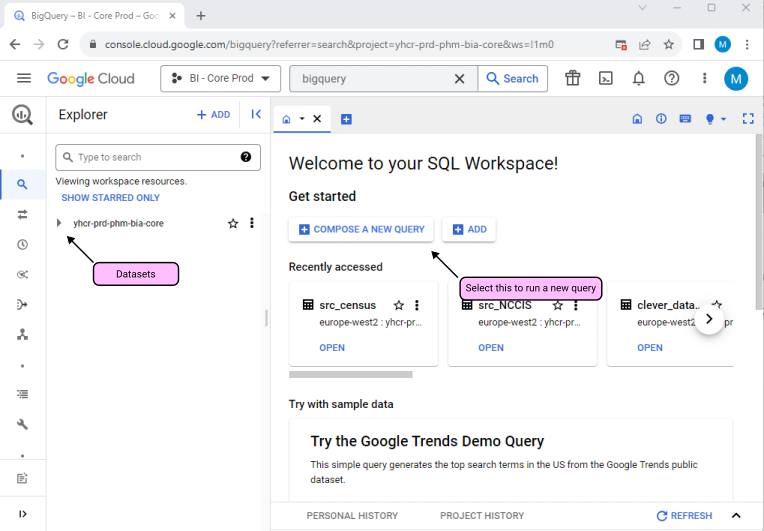
Ok, you've got your account activated, now let's go to BigQuery. BigQuery is a powerful and versatile data warehouse and analytics platform offered by Google Cloud. It is designed to handle large datasets such as the Connected Bradford database and perform fast, SQL-like queries on them.

## The BigQuery Dashboard

So, from the Google Cloud Console dashboard, click on "BigQuery" in the navigation menu or enter "BigQuery" in the Search bar.



You will see that there is a Project ID displayed as - yhcr-prd-phm-bia-core. This references the Connected Bradford database which means that you are in the right place.



To run a query, navigate to the "Compose a New Query". More information on this later.

# Understanding Workspaces

The Connected Bradford team setup two workspaces in Google CLoud for you which you can see in the Explorer panel. These are:

## Project Collaboration Area (CB\_(unique 4 digit number)):

This area is specifically designated for your project and is identified by a unique 4-digit number (e.g., CB\_1234).

It's essential when collaborating with other researchers and analysts on the same project.

### How to Use It:

When working on a shared project, upload your project-related files and documents here.

Organise files by creating folders for different aspects of your project.

Ensure that your files are appropriately named for easy identification.

Share access with your project team by providing them with the folder's link or access permissions.

## Your Personal Workspace ("MYSPACE\_(Your initials)"):

This personal space is dedicated to you and is identified by your initials (e.g., MYSPACE\_JD).

You can use it to store and manage scripts and files related to your individual work.

### How to Use It:

Keep your personal workspace organised by creating folders for different projects or tasks in case you are working on more than one project.

You can use this space for your own scripts, drafts, or any other documents you're working on.

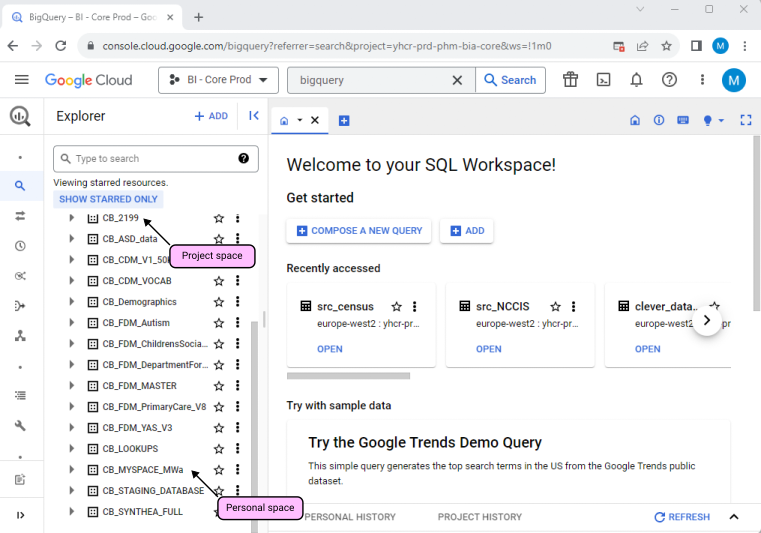
It's a private area, so you can safely store your work without worrying about sharing it with others.

### Additional Tips:

Use clear and descriptive file and folder names to make it easy to find what you need.

Regularly clean up and organise your files to maintain an efficient workspace.

Always respect privacy and access permissions, especially in the project collaboration area.



## Connected Bradford GitHub Collaboration

To get started, we encourage you to join our GitHub space. If you haven't done so already, please create a GitHub account and send your GitHub username to a member of the Connected Bradford team. We will promptly grant you access.

Our GitHub space is organised with sections for every standardized dataset that has undergone our FDM (Flexible Data Model) process. Each dataset page includes:

Explanations: Understand what each dataset contains.

Number of Unique Individuals: Find out how many individuals are included.

Tables: Explore the available data tables.

Useful Code: Access helpful code snippets.

Data Dictionaries: Review data definitions.

If you've developed code that could be beneficial for others working with a dataset, please upload a redacted version. Ensure that you remove project IDs, dataset names, and table names, replacing them with placeholders. If you have any doubts, feel free to consult a member of the Connected Bradford team.

A dedicated folder will be set up for your project on GitHub. You will receive the link for this folder via email from the Connected Bradford team. You can use this folder to upload any project outputs, including analysis scripts.

These steps are designed to streamline collaboration and make it easier for analysts to access and contribute to our valuable dataset resources. Thank you for your participation in Connected Bradford."

# Other tables

In addition to your project and personal workspaces, you will also have access to the following tables which are in the Explorer panel.

## CB\_LOOKUPS:

The CB\_LOOKUPS table contains several tables that may help with your analysis. These tables are designed to provide specific information for your research needs, and two of the most notable ones are:

a. IMD Information:

This sub-table holds data related to the Index of Multiple Deprivation (IMD), which is a crucial socioeconomic measure for understanding disparities in different areas.

It can be a valuable resource when assessing the impact of socioeconomic factors on health outcomes or other research objectives.

b. CTV3 to SNOMED codes:

This table identifies mapping of the CTV3 to SNOMED codes and can be useful when using different datasets such as primary care and secondary care.

More information on the CB Lookups is on the GitHub page - follow this link:

<https://github.com/ConnectedBradford/CB_LOOKUPS>

## CB\_FDM\_MASTER:

The CB\_FDM\_MASTER table is a comprehensive repository of de-identified data for all patients within the Connected Bradford database. It serves as the primary source for demographic information and is considered the 'master' dataset.

This dataset is created by integrating demographic data from various sources, starting with the primary care database, followed by the secondary care dataset, and possibly other datasets like education, etc.

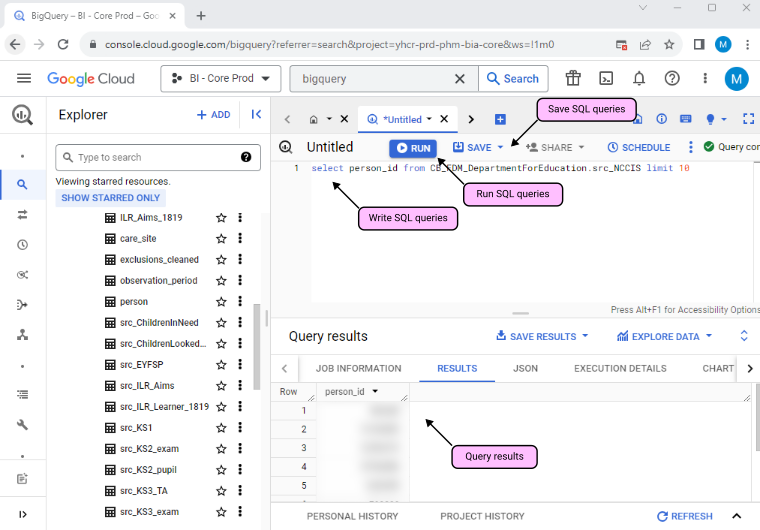
Within the CB\_FDM\_MASTER table, you'll find a rich set of demographic details, including but not limited to ethnicity, gender, and information related to birth and death events. The table assigns a unique identifier to each patient for easy reference and cross-referencing in your research.

Please be aware that the 'person\_id' is a unique identifier assigned to each individual in all datasets. This identifier is crucial for linking individuals across different datasets.

Top of Form

# Writing SQL Queries

Now that your Google account is activated and you have access to your project and personal workspace, including the master demographic table and lookup tables, it's time to start running SQL queries. Below, you'll find instructions on how to navigate the query interface, execute queries, view results, and save your work. Keep in mind that this is an introductory package, and we will delve into more advanced analytical tools in the next section.



## Data Export and Download

Please remember that you are not allowed to download data to your local machine or another server. You should use the designated folder to store scripts in your workspaces.

# Advanced Analytics - Vertex Environment

## Accessing the Vertex AI Platform

Vertex AI Workbench is a Jupyter notebook-based development environment for the entire data science workflow. You can interact with Vertex AI and other Google Cloud services from within a Vertex AI Workbench instance's Jupyter notebook.

For example, Vertex AI Workbench lets you:

- Access and explore your data from within a Jupyter notebook by using BigQuery and Cloud Storage integrations.

- Use SQL, R or Python

- Automate recurring updates to your model by using scheduled executions of your notebook's code that run on Vertex AI.

- Process data quickly by running a notebook on a Dataproc cluster.

- Run a notebook as a step in a pipeline by using Vertex AI Pipelines.

You can access Vertex Workbench [here.](https://console.cloud.google.com/vertex-ai/workbench/user-managed?authuser=2&project=yhcr-prd-phm-bia-core)

From the Google Cloud Console dashboard, click on "Vertex AI" in the navigation menu.

.

## Setting Up Vertex Notebooks

To access your Vertex Notebook, see the screenshot below.

A screenshot of a computer

Description automatically generated

This will take you to the Workbench screen. Scroll down to find your name (Your workbook will be in the format: "mgd-firstname-surname-workspace" and click "OPEN JUPYTERLAB".

A screenshot of a computer

Description automatically generated

This will take you to the JupyterLab launcher screen. You will see this screen every time you open Jupyter Lab.

A screenshot of a computer program

Description automatically generated

The first time you use Jupyter Lab, you'll have to follow the guidance below to set-up your managed notebook. You should only ever have to do this once, but if you need to do it again, you'll be able to find a dedicated managed notebook set-up guide [here](https://github.com/ConnectedBradford/CB_GuidanceForAnalysts/blob/main/docs/Using%20Managed%20Notebooks%20-%20security%20config.doc)

Once the above Launcher screen is open, click the Terminal icon to start a new terminal session. Once here type in : "gcloud auth application-default login"

When prompted type "Y" - you will then be presented with "Enter authorisation code" message.

Select and copy (using Ctrl+C), the authorisation code and create a new tab in your browser and paste it into the address bar at the top. You may then be presented with a window that asks you to choose an account to continue to Google Auth Library. Select the "yhcr.nhs.uk account".

Click allow and this screen will appear

A screenshot of a computer

Description automatically generated

Copy the code, return to your JuypterLab and paste the code in.

Hit enter (return key) then type in the following in an R script to set up the authentication properly.

"Export - GOOGLE APPLICATION\_CREDENTIALS="(PATH)""

# Analysing Data in Vertex Notebooks

## Using R in Jupyter

You can now setup a R script as normal. BUT you must ensure that the first two lines of   
code are:

"library(bigrquery)"  
"bg\_auth()"

More information on this is covered in the document below- see link:

<https://github.com/ConnectedBradford/CB_GuidanceForAnalysts/blob/main/docs/cBdfd_newstarterguide.pdf>

Note that we need to make some amendments to this section!

## Running Python Code for Analysis

There are a couple of ways to access BigQuery via Python. One is using [google.cloud.bigquery()](https://pypi.org/project/google-cloud-bigquery/) to run SQL syntax, and the second is using Python's [magic-cell functionality](https://cloud.google.com/python/docs/reference/bigquery/latest/magics).

### Option 1: google.cloud.bigquery

##### Import libraries

From google.cloud import bigquery

##### Instantiate bigQuery client.

"client = bigquery.Client()"

The bigquery client will accept the SQL syntax as a string. For example:

Define string of SQL syntax

"my\_sql\_syntax = """ SELECT year\_of\_birth FROM yhcr-prd-phm-bia-core.CY\_MYSPACE\_CMC.person WHERE year\_of\_birth IS NOT NULL LIMIT 5 """"

Run SQL string and store in a pandas.DataFrame

"client\_query\_output = client.query(my\_sql\_syntax).to\_dataframe() client\_query\_output"

Output will look like this

| "year\_of\_birth |  |
| --- | --- |
| 1900 |  |
| 1900 |  |
| 1900 |  |
| 1900 |  |
| 1900" |  |

### Option 2: Magic Cells

Note: the below is a script that can be copied into a Python notebook. It can't be executed here, as this guide is being produced as an R Markdown file.

"%load\_ext google.cloud.bigquery

New cell

%%bigquery

SELECT

gender,

COUNT(person\_id) as person\_count

FROM yhcr-prd-phm-bia-core.CY\_MYSPACE\_RS.asc\_person\_demos

GROUP BY gender

ORDER BY gender"

Output will look like this:

| "gender | person\_count |
| --- | --- |
| FEMALE | 27888 |
| MALE | 21751 |
| UNKNOWN | 2" |

Note: to save the query as a dataframe just type the name of your data frame after "%%bigquery". e.g. "%%bigquery python\_demo" (you will not see a table output when you execute this

Using R you start by pulling in BigQuery (and any other packages you'd like to use), which is R's package to run queries via BigQuery

"library(bigrquery)"

"bq\_auth()"

For example use the Lookup dataset. Store the project ID "yhcr-prd-phm-bia-core" this is the same for all Connected Bradford projects.

Example dataset - change "CB\_LOOKUPS" to whichever dataset you want to query.

"targetdb1 <-'yhcr-prd-phm-bia-core.CB\_LOOKUPS'

targetdb1 <-gsub(' ','',targetdb1)

print (targetdb1)"

# Run

"[1] "yhcr-prd-phm-bia-core.CB\_LOOKUPS"

In [4]:"

If successful, you will see a printout similar to the above with the name of the dataset you've queried.

Now that we've found the dataset we're looking for, we now need to list the tables within that dataset.

Top of Form

## Listing the tables in BigQuery

This link provides information on how to list tables using BigQuery.

<https://github.com/ConnectedBradford/CB_GuidanceForAnalysts/blob/main/TrainingAndGuidance/2_ListTheTablesUsingaquery.ipynb>

## Using OpenCodeList with Juypter Notebooks

OpenCode Lists is useful if you trying to identify a cohort. OpenCode Lists generates a list of CTV3 readcodes which can then be used to identify individuals from the Connected Bradford primary care dataset associated with that CTV3 readcode.

This section will help you develop a code list from <https://www.opencodelists.org/>.

You will need to ensure that the codelist you use is the one that relates to the dataset you have access to.

In this particular example, a series of CTV3 codes have been used to build a list of codes related to “Tuberculosis”.

This needs to be downloaded to your local machine as a csv.

You will now open your notebook and upload this csv.

A screenshot of a computer

Description automatically generated

To upload this csv into a table in my project space using a R notebook



The script above reads the name of the csv. It then creates a table in your personal workspace dataset called "tuberculosis".

Now to query the Primary Care dataset to identify any person who has the tuberculosis condition, you will need to do the following:

- Join the new tuberculosis table created above, with the Primary care table and the CB Master Person table.

This script below this data into a R data frame.



There are some very useful videos on this website that detail how to use it to create a list of CTV3 readcodes.

## How to create a Cohort Using My Codelist

This link provides more information on how to create a cohort using my codelist.

<https://github.com/ConnectedBradford/CB_GuidanceForAnalysts/blob/main/TrainingAndGuidance/6_MakeACohortUsingMyCodelist.ipynb>

## Do's and Don'ts and further help with Juypter and Bigquery

This links to a document that helps understand the key differences between Juypter and Bigquery and gives some guidance on do's and don'ts for each:

<https://github.com/ConnectedBradford/CB_GuidanceForAnalysts/blob/main/docs/The%20use%20of%20Bigquery%20and%20Jypter%20Notebooks_V1.0.docx>

# Useful links

[BigQuery QuickStart](https://cloud.google.com/bigquery/docs/quickstarts) - A quickstart guide with dummy public data

[The Lazy Guide to Learning BigQuery SQL](https://codingisforlosers.com/learn-bigquery-sql/) - A smart, simple guide to the BigQuery basics, complete with tutorial videos

[Introduction to Vertex AI Workbench](https://cloud.google.com/vertex-ai/docs/workbench/introduction) - An overview of Vertex Workbench and Jupyter notebook

[Use R with BigQuery](https://cloud.google.com/vertex-ai/docs/workbench/user-managed/use-r-bigquery) - Getting started with R in BigQuery

[Quertying in the Cloud](https://medium.com/analytics-vidhya/querying-in-the-cloud-using-bigquery-with-r-and-python-576541554967) - Using R with BigQuery

[Using Stash in BigQuery to save datframes from R](https://cloud.google.com/blog/products/gcp/google-cloud-platform-for-data-scientists-using-r-with-google-bigquery-part-2-storing-and-retrieving-data-frames) - How to save dataframes and queries permanently to BigQuery

[Using IPython Magic](https://cloud.google.com/python/docs/reference/bigquery/latest/magics) - An introduction to Python magic cells

[Wotking with BigQuery and Python](https://medium.com/geekculture/working-with-bigquery-and-a-python-notebook-d4a06c008056) - Working with Pyhtonand BigQuery

[Visualising BigQuery data with Pyhton in Jupyter](https://cloud.google.com/bigquery/docs/visualize-jupyter) - A quickstart guide on visualising BigQuery data using Python in Jupyter

[Quickstarts  |  BigQuery  |  Google Cloud](https://cloud.google.com/bigquery/docs/quickstarts)