

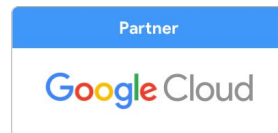
Google BigQuery



Giuliano Ribeiro

<https://about.me/giulianobr>  @giulianobr

- **Google Cloud Solutions Architect**
 - G Suite
 - Google App Engine
 - BigQuery
 - Java
 - Build and Release automation
- **Working:**
 - ilegra +14y
 - Google technologies +6y





Compute

From virtual machines with proven price/performance advantages to a fully managed app development platform.

Compute Engine

App Engine

Kubernetes Engine



Big Data

Fully managed data warehousing, batch and stream processing, data exploration, Hadoop/Spark, and reliable messaging.

BigQuery

Cloud Dataflow

Cloud Dataproc



Internet of Things

Intelligent IoT platform that unlocks business insights from your global device network

Cloud IoT Core



Developer Tools

Develop and deploy your applications using our command-line interface and other developer tools.

Cloud SDK

Container Registry

Container Builder



Storage and Databases

Scalable, resilient, high performance object storage and databases for your applications.

Cloud Storage

Cloud SQL

Cloud Bigtable



Data Transfer

Online and offline transfer solutions for moving data quickly and securely.

Google Transfer Appliance

Cloud Storage Transfer Service

Google BigQuery Data Transfer



Cloud AI

Fast, large scale and easy to use AI services.

Cloud AutoML ^{Alpha}

Cloud Machine Learning Engine

Cloud Job Discovery



Identity & Security

Control access and visibility to resources running on a platform protected by Google's security model.

Cloud IAM

Cloud Identity-Aware Proxy

Cloud Data Loss Prevention API



Networking

State-of-the-art software-defined networking products on Google's private fiber network.

Cloud Virtual Network

Cloud Load Balancing

Cloud CDN



API Platform & Ecosystems

Cross-cloud API platform enabling businesses to unlock the value of data, deliver modern applications, and power ecosystems.

Apigee API Platform

API Monetization

Developer Portal



Management Tools

Monitoring, logging, and diagnostics and more, all in an easy to use web management console or mobile app.

Stackdriver Overview

Monitoring

Logging



Google Cloud Platform

Before BigQuery ...

High amount of data

Need to be fast and precise

Dremel internal massively parallel query service

Google has been using Dremel in production since 2006

BigQuery is a new version and the public implementation of Dremel



So what is BigQuery?

It is the Google's data warehouse solution in the cloud.

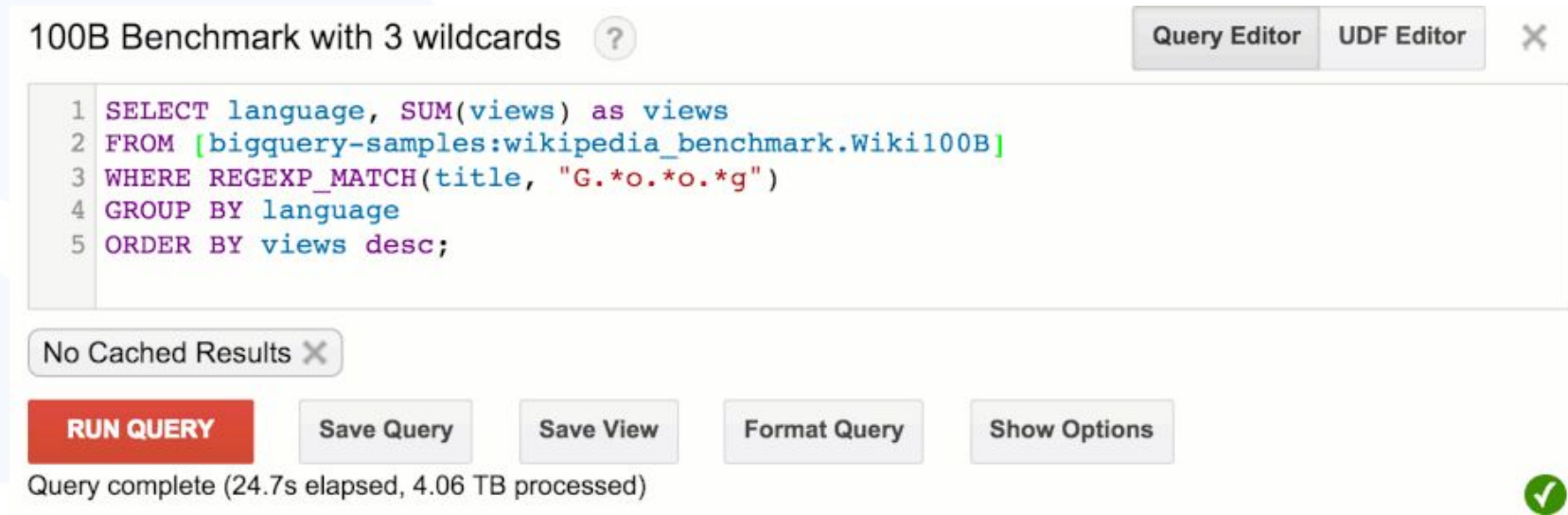
Petabyte scale

Fully managed

Low Cost analytics



How powerful is BigQuery?



The screenshot shows the BigQuery Query Editor interface. At the top, the title bar reads "100B Benchmark with 3 wildcards" with a help icon. To the right are tabs for "Query Editor" and "UDF Editor", and a close button. The main area contains a SQL query with line numbers 1 through 5. Below the query editor is a status bar that says "No Cached Results" with a close icon. Underneath are five buttons: "RUN QUERY" (highlighted in red), "Save Query", "Save View", "Format Query", and "Show Options". At the bottom, a message states "Query complete (24.7s elapsed, 4.06 TB processed)" followed by a green checkmark icon.

100B Benchmark with 3 wildcards ? Query Editor UDF Editor X

```
1 SELECT language, SUM(views) as views
2 FROM [bigquery-samples:wikipedia_benchmark.Wiki100B]
3 WHERE REGEXP_MATCH(title, "G.*o.*o.*g")
4 GROUP BY language
5 ORDER BY views desc;
```

No Cached Results X

RUN QUERY Save Query Save View Format Query Show Options

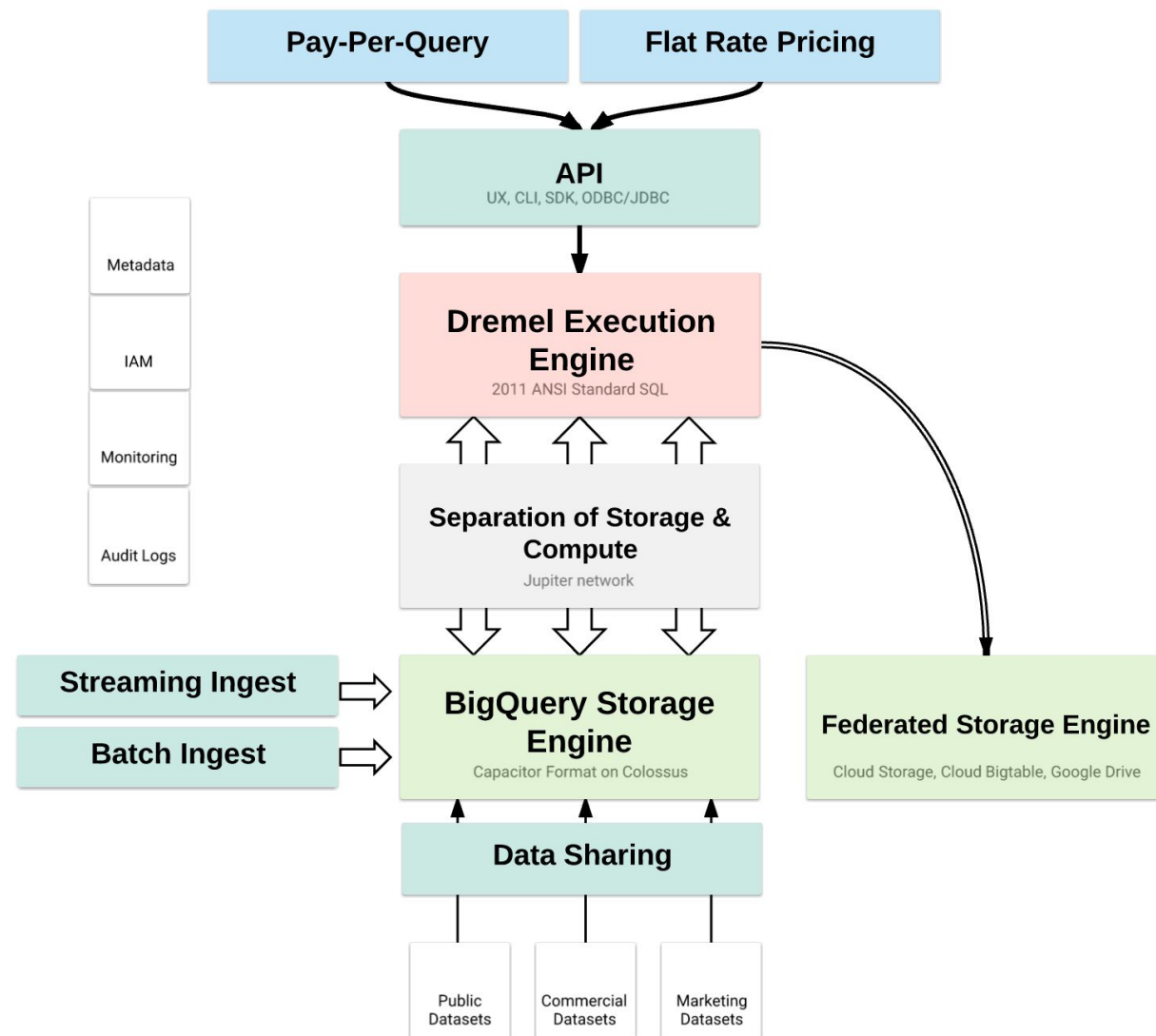
Query complete (24.7s elapsed, 4.06 TB processed) ✓

- About 330 100MB/sec dedicated hard-drives to read 1TB of data
- A 330 Gigabit network to shuffle the 1.25 TB of data
- 3,300 cores to uncompress 1TB of data and process 100 billion regular expressions at 1 μ sec per

How it is possible?

Because BigQuery is

- NoOps
- Serverless
- Truly Cloud Native
- EASY



Some features

- **SQL 2011** standard
- ODBC and JDBC drivers
- Table partition
- External sources
- Batch & Streaming ingestion
 - CSV
 - JSON
 - Datastore backups
 - AVRO
 - ORC,
 - Parquet

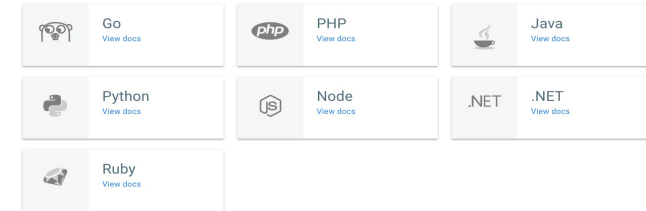
Some features

- Petabyte ready
- Global available - US/EU/JP
- Pay only for storage & processing
- Zero administration for performance and scale
- Enterprise-grade Data Sharing
- Encrypted at rest and the wire

How to use

- Web UI
- CLI (bq)
- API
- JDBC/ODBC

- # How to use
- Web UI
 - CLI (bq)
 - API
 - JDBC/ODBC



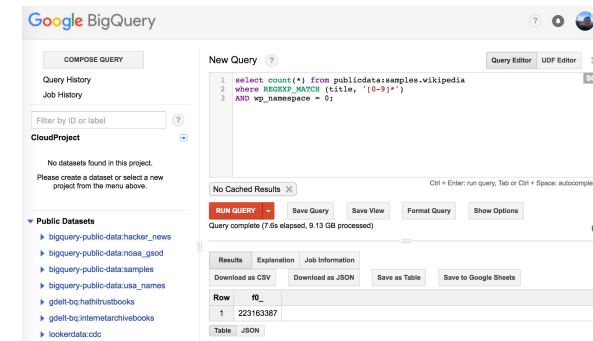
```
OSX % bq shell
Welcome to BigQuery! (Type help for more information.)
> select count(*) from publicdata:samples.shakespeare;
Waiting on bqjob_r5083c9926faec936_0000015d662f1545_1 ... (0s) Current status: DONE

+-----+
| f0_   |
+-----+
| 164656 |
+-----+

> SELECT count(*) FROM [bigquery-public-data:usa_names.usa_1910_current] where name = 'Richard'
Waiting on bqjob_r1f510cc5735f9291_0000015d66326565_1 ... (0s) Current status: DONE

+-----+
| f0_   |
+-----+
| 5987  |
+-----+

>
```



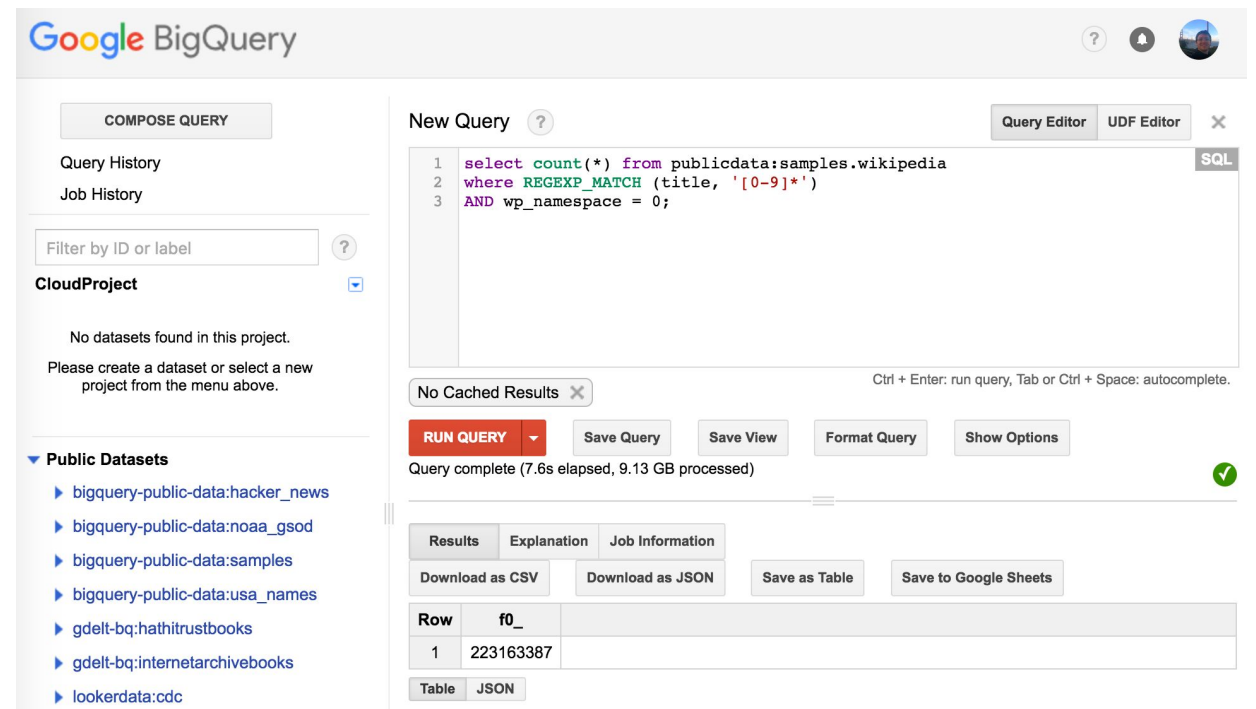
How to use

- Command line (bq) - DEMO

```
OSX ~ bq shell
Welcome to BigQuery! (Type help for more information.)
> select count(*) from publicdata:samples.shakespeare;
Waiting on bqjob_r5083c9926faec936_0000015d662ff545_1 ... (0s) Current status: DONE
+-----+
| f0_ |
+-----+
| 164656 |
+-----+
> SELECT count(*) FROM [bigquery-public-data:usa_names.usa_1910_current] where name = 'Richard'
Waiting on bqjob_r1f510cc5735f9291_0000015d66326565_1 ... (0s) Current status: DONE
+-----+
| f0_ |
+-----+
| 5987 |
+-----+
>
```

How to use

- Web UI - DEMO



The screenshot displays the Google BigQuery web interface. On the left sidebar, there are links for 'COMPOSE QUERY', 'Query History', and 'Job History'. Below these is a search bar 'Filter by ID or label' and a section for 'CloudProject' which states 'No datasets found in this project. Please create a dataset or select a new project from the menu above.' At the bottom of the sidebar is a list of 'Public Datasets' including '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', and 'lookerdata:cdc'.

The main area is titled 'New Query' and contains a SQL query editor with the following code:

```
1 select count(*) from publicdata:samples.wikipedia
2 where REGEXP_MATCH (title, '[0-9]*')
3 AND wp_namespace = 0;
```

Below the query editor, it shows 'No Cached Results' and a status bar indicating 'Query complete (7.6s elapsed, 9.13 GB processed)'. There are buttons for 'RUN QUERY', 'Save Query', 'Save View', 'Format Query', and 'Show Options'. At the bottom, there are tabs for 'Results', 'Explanation', and 'Job Information', along with buttons for 'Download as CSV', 'Download as JSON', 'Save as Table', and 'Save to Google Sheets'. The 'Results' tab is active, showing a table with one row:

| Row | f0_ |
|-----|-----------|
| 1 | 223163387 |

At the very bottom, there are tabs for 'Table' and 'JSON'.

Who is using?

- The New York Times
- HSBC
- Motorola
- Spotify
- Evernote



The
New York
Times



Spotify case



Neville Li
@sinisa_lyh

Seguindo



Em resposta a [@sinisa_lyh](#) [@frankyaorenjie](#) e 2 outros

We query 500PBs in BigQuery per month, with zero cluster management and operations. A lot of them feed into Scio/Dataflow jobs.

 Traduzir Tweet

01:10 - 21 de abr de 2017

6 Retweets 10 Curtidas



2



6



10



| Common query types | Hive / Hadoop | BigQuery |
|---|---------------|---------------|
| KPIs by specified ad hoc parameters | ~1,200 secs | ~10 - 20 secs |
| FB audience list for social targeting for AU campaign | ~4,000 secs | ~15 - 30 secs |
| Top tracks by age / gender by market | ~17,500 secs | ~500 secs |

Partners



Try it!

- First 1TB is free!
- On-demand pricing
 - Processing: \$5 per TB
 - Storage:
 - \$0.02 per GB, per month
 - \$0.01 per GB, per month for long term storage
- Flat pricing
 - \$40,000 per month

Thank you!

