Google Cloud Platform

Cloud BigQuery

udemy GCP Gurus

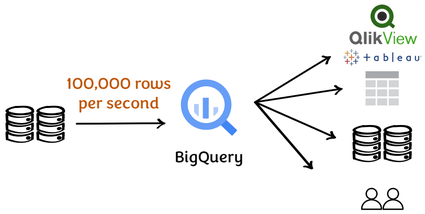
Store data from Any

Source
- Any Size
- From Anywhere

Retrieve/Query Data
- Super fast...
- Charge based on amount of data processed

Google Cloud → BigQuery

BigQuery is a fully managed, AI-ready data analytics platform that helps you maximize value from your data and is designed to be multi-engine, multi-format, and multi-cloud.



BigQuery Partitions

- Use for low cardinality
- Columns based partitions

BigQuery Clustering

- Use for high cardinality
- Less expensive
- Data is stored sorted within partition
- No Need to manually Shared tables

Product	Price
A	$120
B	$30
C	$80
D	$900
E	$60
F	$10

($0-$49.99)

Product	Price
F	$10
B	$30

($50-$99.99)

Product	Price
E	$60
C	$80

($100+)

Product	Price
A	$120
D	$900

Advanced Strategies for Partitioning and Clustering in BigQuery

BigQuery Best Practices

Query Performance

Input data and data sources (I/O): How many bytes does your query read?

Communication between nodes (shuffling): How many bytes does your query pass to the next stage? How many bytes does your query pass to each slot?

Computation: How much CPU work does your query require?

Outputs (materialization): How many bytes does your query write?

Query anti-patterns: Are your queries following SQL best practices?

Storage Optimizations
- Use the expiration settings to remove unneeded tables and partitions
- Take advantage of long-term storage
- Use the pricing calculator to estimate storage costs

BigQuery Limits

Query Jobs, Load Jobs, Export Jobs, DataSet limits, Table Limits → BigQuery → UDF Limits, Data Manipulation's Statements, BigQueryML Limits, Streaming inserts, API Requests

Features

- Flexible Data Ingestion
- Global Availability
- Security & Permissions
- Cost Controls
- Highly Available
- Fully Integrated
- Connect with Google Products
- Automatic Data Transfer Service

Flexible Data Ingestion

Load your data from Google Cloud Storage or Google Cloud Datastore, or stream it into BigQuery at 100,000 rows per second to enable real-time analysis of your data.

Global Availability

You have the option to store your BigQuery data in European locations while continuing to benefit from a fully managed service, now with the option of geographic data control, without low-level cluster maintenance headaches.

Security & Permissions

You have full control over who has access to the data stored in Google BigQuery. Shared datasets will not impact your cost or performance (those you share with pay for their own queries).

Fully Integrated

In addition to SQL queries, you can easily read and write data in BigQuery via Cloud Dataflow, Spark, and Hadoop.

Cost Controls

BigQuery provides cost control mechanisms that enable you to cap your daily costs to an amount that you choose. For more information, see Cost Controls.

Connect with Google Products

You can automatically export your data from Google Analytics Premium into BigQuery, visualize it using Google Data Studio and analyze datasets stored in Google Cloud Storage.

Highly Available

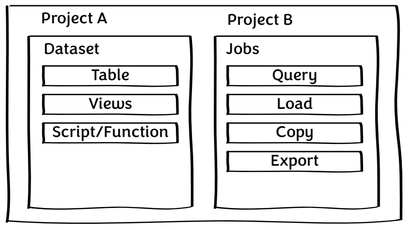
Transparent data replication in multiple geographies means your data is available and durable even in the case of extreme failure modes.

Automatic Data Transfer Service

The BigQuery Data Transfer Service automatically transfers data from partner SaaS applications to Google BigQuery on scheduled, managed basis.

	Common Database Design Constraints	Advanced Big Query .. ...
IO	• Disks • Attached Storage scaling limits	• 100k disks in parallel • Separation of storage and compute
CPU	Not getting faster enough	1000's CPU's
Network	Coordination bottleneck	Jupiter Networks

BigQuery Resources



Datasets :
- Datasets are the top-level containers used to organize and control access to the BigQuery tables and views.
- Datasets frequently map to schemas in standard relational databases and data warehouses.

Jobs :
- Jobs are actions that BigQuery runs on your behalf to load data, export data, query data, or copy data.
- Jobs are not linked to the same project that the data is stored in. However, the location where the job can execute is linked to the dataset location.

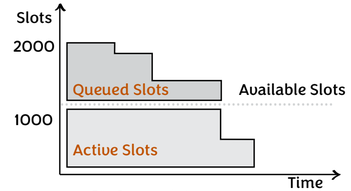
Fundamentals

Tables contain your data in BigQuery, along with a corresponding table schema that describes field names, types, and other information. BigQuery also supports views, virtual tables defined by a SQL query.

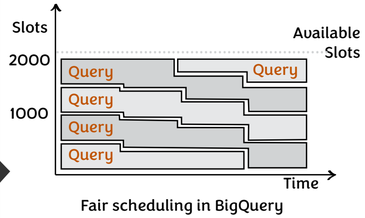
BigQuery Creates Tables:
- Loading data into a new table
- Running a query
- Copying a table

Slots

- Unit of Computational Capacity
- Automatically calculates for query
- Flat rate pricing can be used to purchase dedicated number of slots



BigQuery slots queued up if demand exceeds availability



Fair scheduling in BigQuery

Estimation

- You can get estimate of query before submitting it
- Online in Console or with –dryrun in bg