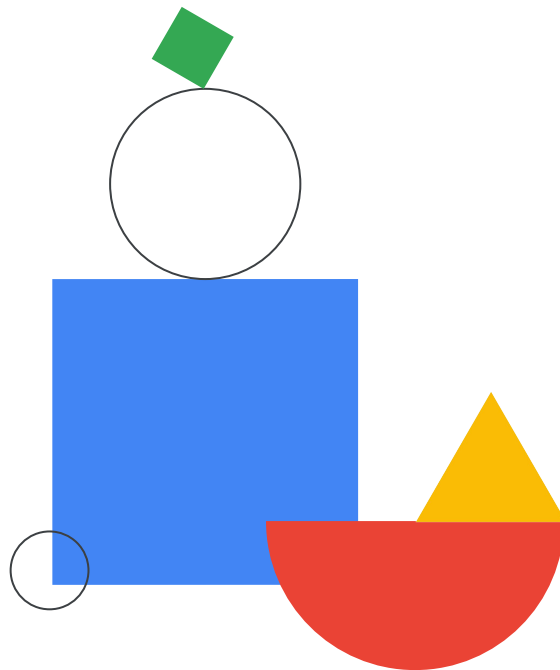


# Introduction to Data Analytics & AI on Google Cloud

From Raw Data to Real Insights

Q1 2026

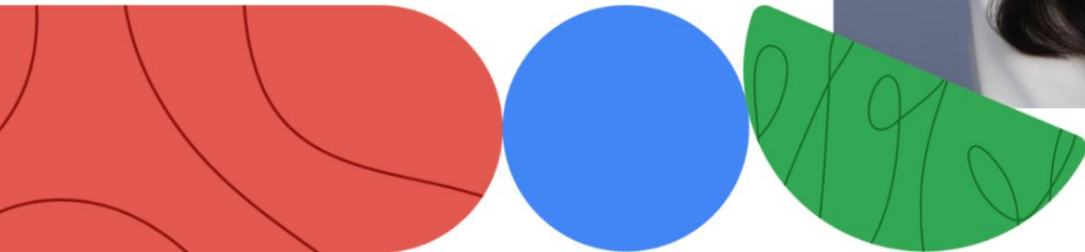




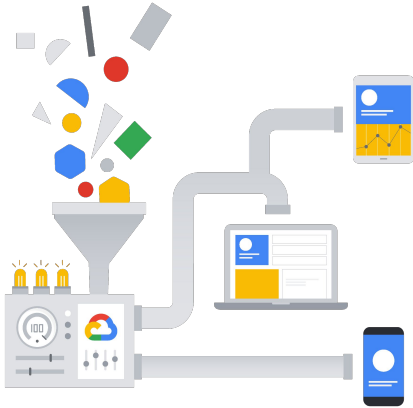
# Paggie Chen

Trainer

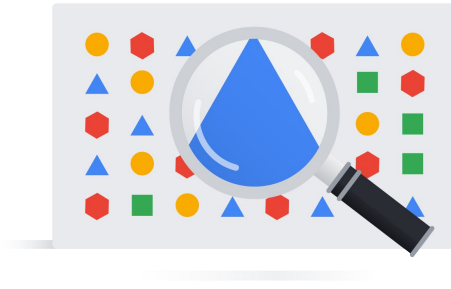
Google Cloud



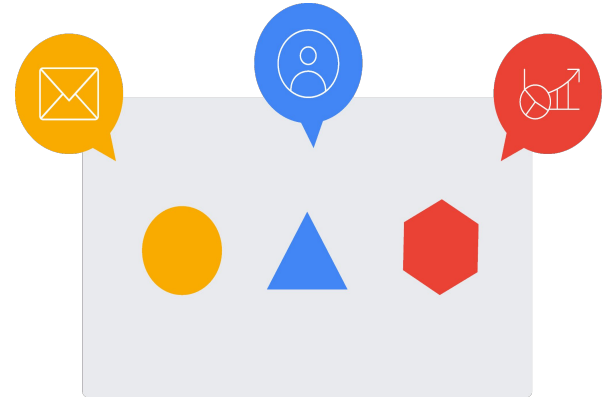
# What is Data Analytics?



Capture & Manage



Insights



Data-driven Decisions

# Agenda

01 The Data Analytics Lifecycle Overview

---

02 Ingest

---

03 Process

---

04 Store

---

## Analyze

Introduction to BigQuery

Querying External Data Sources or Public Datasets **Demo**

Control access with Policy Tag **Demo**

Building Pipelines in BigQuery - from low-code to coding method **Demo**

Visualization & Business Intelligence with Looker & Looker Studio **Demo**

Derive insights from your data in BigQuery (Agent, Gemini Enterprise) **Demo**

---

06

## Activate

Introduction AI on Google Cloud

BigQuery ML **Demo**

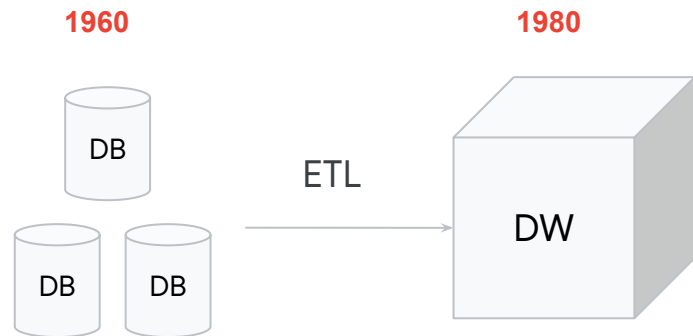
---

07

Governance

---

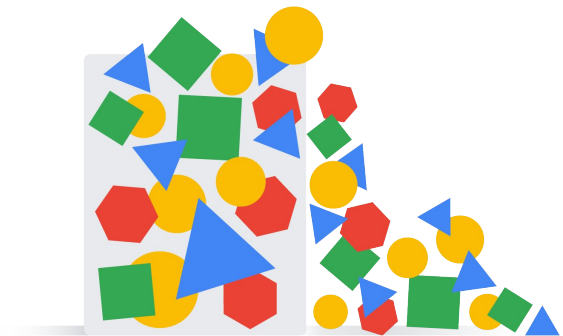
1980  
A **data warehouse** is like a highly organized library.



2010

# Data lake

A data lake is a centralized repository designed to store, process, and secure large amounts of structured, semistructured, and unstructured data. It can store data in its native format and process any variety of it, ignoring size limits.



01

Structured data

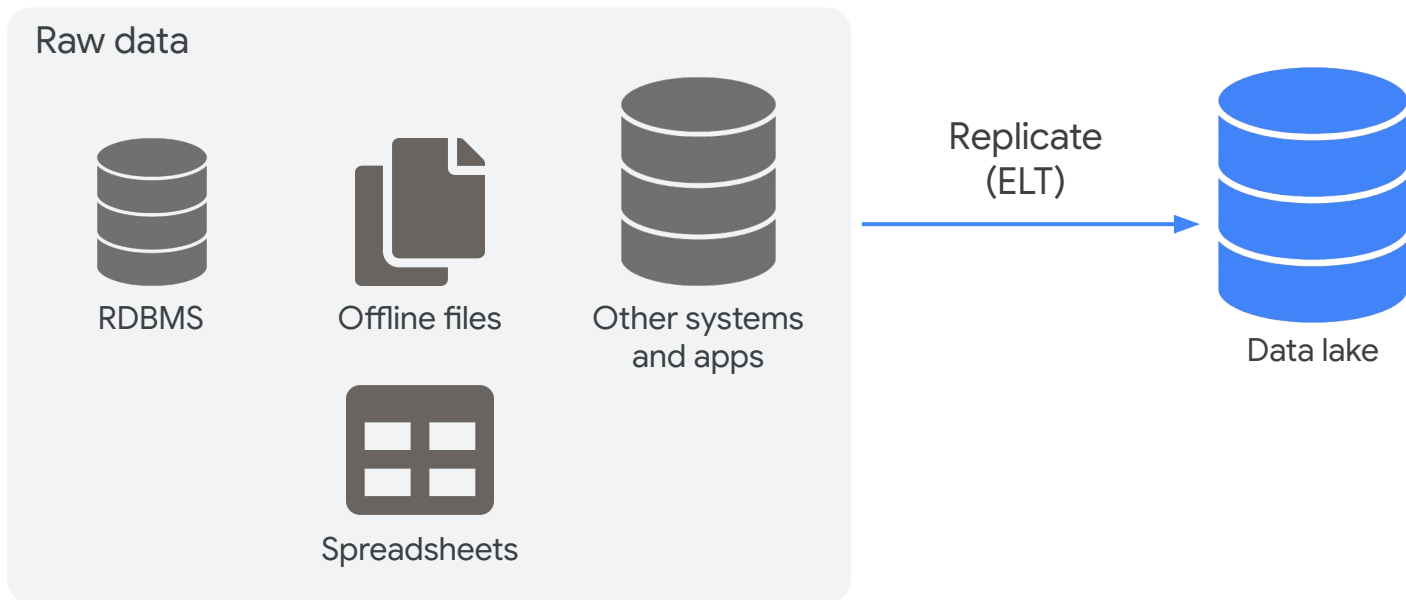
02

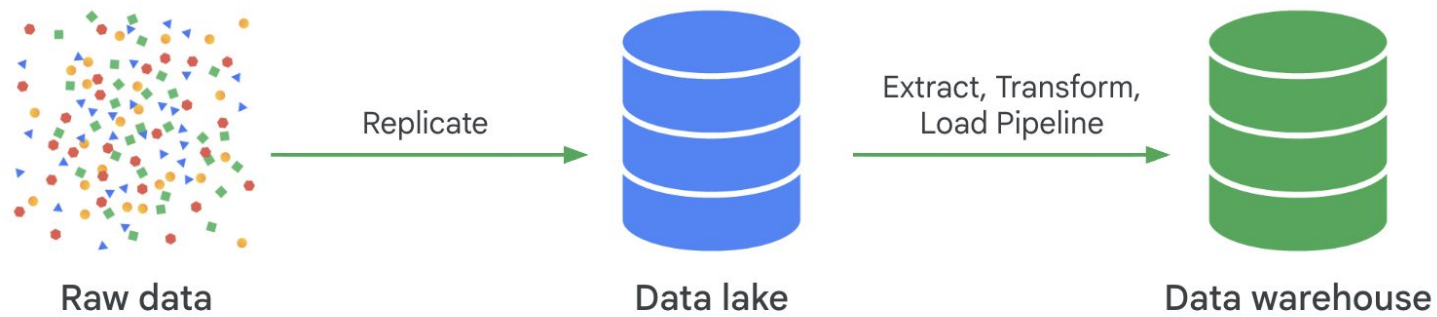
Semistructured data

03

Unstructured data

# A data lake brings together data from across the enterprise into a single location





# Data warehouse vs. data lake

## Traditional data storage

- Data warehouse: Ideal for structured, organized data used for business insights.
- Data lake: A cost-effective place for huge amounts of raw data in any format.

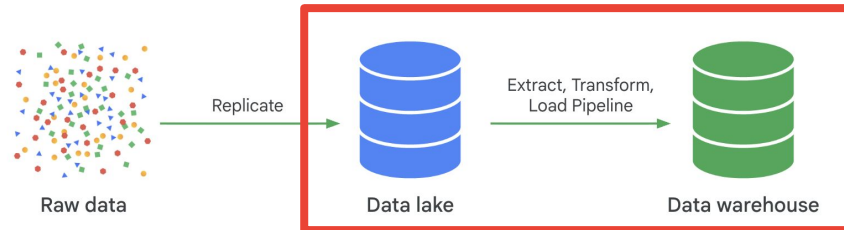
## Challenges of separation

- Data silos: Hard to analyze different data types together.
- Complex extract, transform, load (ETL): Required moving and duplicating data, increasing costs and delays.
- Governance Issues: Managing security and access across two systems was complicated.

2020

# Data lakehouse architecture

Proprietary + Confidential



01

Combines low-cost, flexible data lake storage.

02

Offers powerful querying, transaction management, and governance features of a data warehouse.

03

Delivers a single, unified system for modern data architecture.

BigLake

# Agenda

## 01 The Data Analytics Lifecycle Overview

---

## 02 Ingest

---

## 03 Process

---

## 04 Store

---

### Analyze

Introduction to BigQuery

Querying External Data Sources or Public Datasets **Demo**

## 05

Control access with Policy Tag **Demo**

Building Pipelines in BigQuery - from low-code to coding method **Demo**

Visualization & Business Intelligence with Looker & Looker Studio **Demo**

Derive insights from your data in BigQuery (Agent, Gemini Enterprise) **Demo**

---

## 06

### Activate

Introduction AI on Google Cloud

BigQuery ML **Demo**

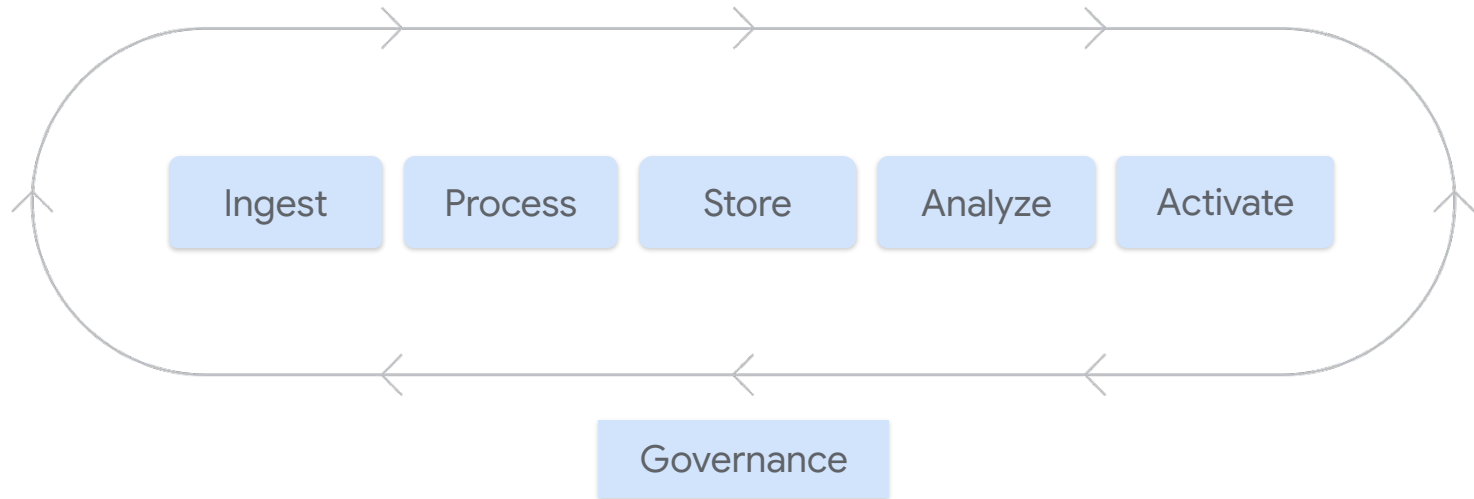
---

## 07

### Governance

---

# Data analytics lifecycle



→ It's an iterative process. You will often go back and forth.

# Agenda

01 The Data Analytics Lifecycle Overview

---

02 Ingest

---

03 Process

---

04 Store

---

Analyze

Introduction to BigQuery

Querying External Data Sources or Public Datasets **Demo**

05 Control access with Policy Tag **Demo**

Building Pipelines in BigQuery - from low-code to coding method **Demo**

Visualization & Business Intelligence with Looker & Looker Studio **Demo**

Derive insights from your data in BigQuery (Agent, Gemini Enterprise) **Demo**

---

06 Activate

Introduction AI on Google Cloud

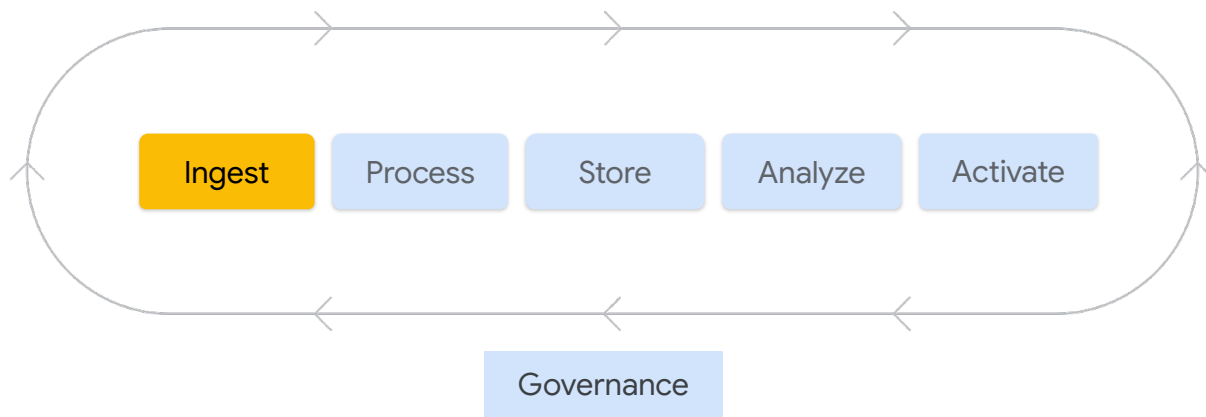
BigQuery ML **Demo**

---

07 Governance

---

# Data analytics lifecycle: Ingest



**Objective:** Breakdown data silos, make all data reliably accessible when needed.

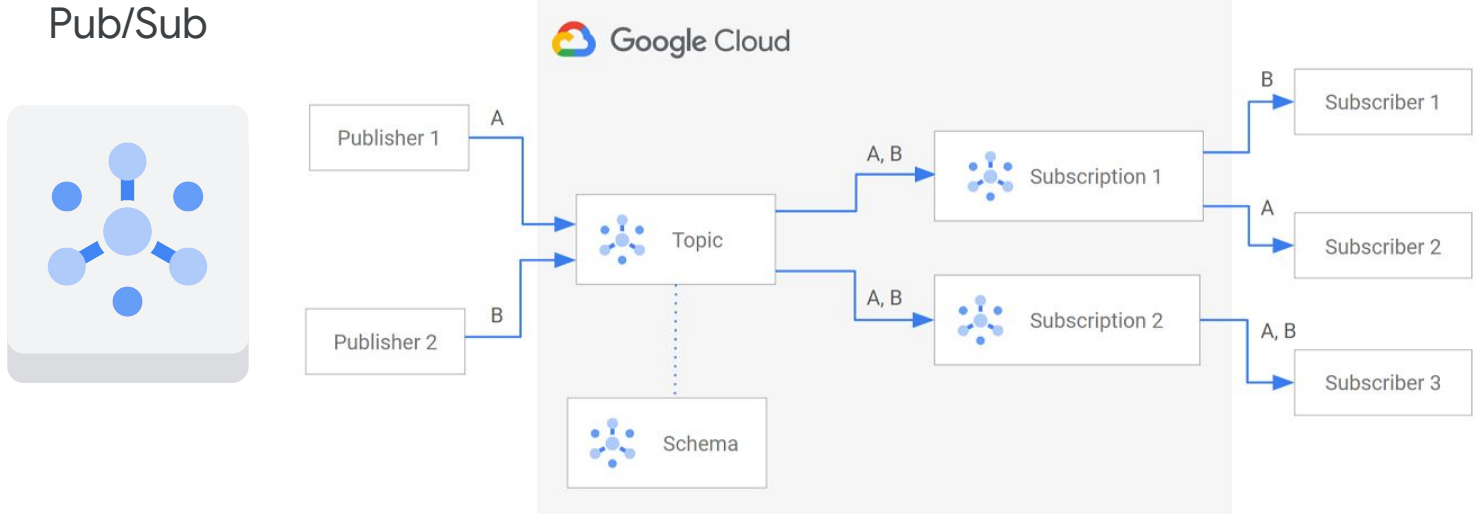
**This layer focuses on:**

- Latency requirements (real-time vs near real-time vs batch)
- Data completeness and delivery guarantees
- Replayability and backfill capability
- Source reliability and failure handling

**Related Tools:**

- Pub/Sub
- Dataproc
- Dataflow
- Cloud Storage

# Data analytics lifecycle: Ingest



# Data analytics lifecycle: Ingest

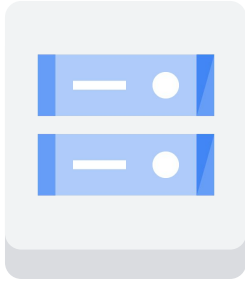
Pub/Sub



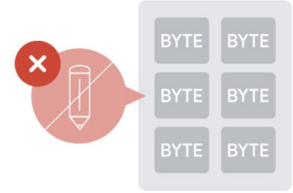
- ✓ Ensures at-least-once delivery
- ✓ No provisioning is required
- ✓ APIs are open
- ✓ Global by default
- ✓ Offers end-to-end encryption

# Data analytics lifecycle: Ingest/Store

Cloud Storage



Object



Cannot directly edit



Bucket

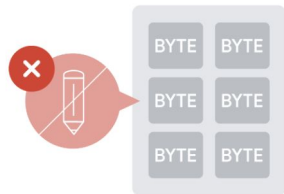
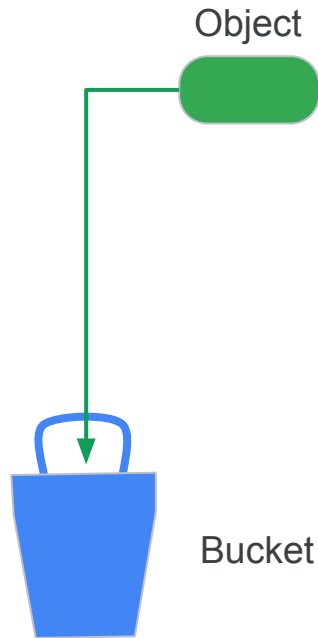
# Cloud Storage is an object storage service

## Objects

- Immutable
- Maximum object size: 5 TiB
- Object versioning
- Object Lifecycle Management
- Access by URL

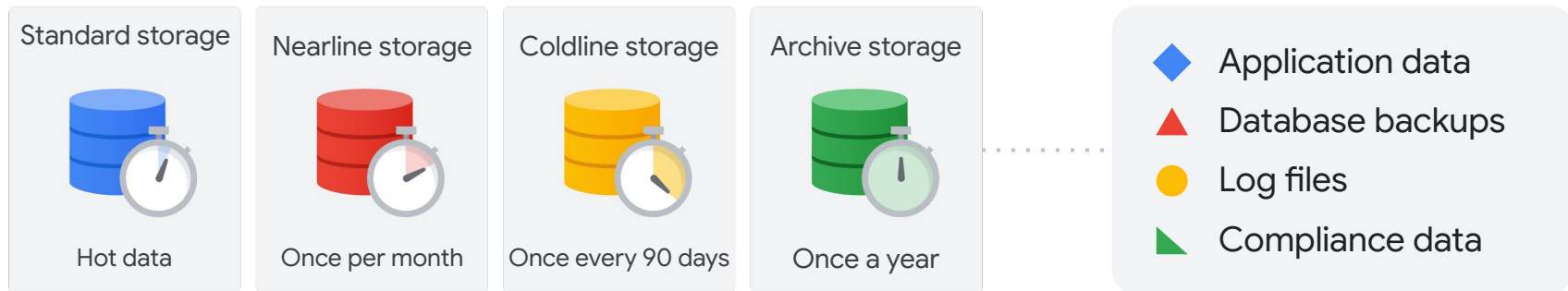
## Buckets

- Global unique name
- Zone/Region/Dual-region/Multi-region
- Storage class: Standard/Nearline/Coldline/Archive
- Autoclass



Cannot directly edit

# Cloud Storage holds your structured/unstructured data



✓ Reliability and scalability

✓ Accessed by HTTP request

✓ Retrieved by object name

# Use cases & key features

## Use cases:

- Website content
- Storing data for archiving and disaster recovery
- Distributing large data objects to users via direct download

## Key features:

- Scalable to exabytes
- Time to first byte in milliseconds
- Very high availability across all storage classes
- Single API across storage classes

# Demo: Cloud Storage

The screenshot shows the Google Cloud Storage interface for a bucket named 'coe-lecture-material'. The bucket is located in 'me-central1 (Doha)', uses 'Standard' storage class, and has 'Public access' set to 'Access granted to public principals'. The 'Protection' is set to 'Soft Delete'. The 'Objects' tab is active, displaying a list of files. The list includes files like '01. Introducing Google Cloud.pdf', '01. Introducing Google Cloud.pptx', '02. Resources and Access in the C...', '02. Resources and Access in the C...', '03. Networking in Google Cloud.pdf', '03. Networking in Google Cloud.pp...', '04. Advanced Networking.pdf', '04. Advanced Networking.pptx', '05. Google Compute Engine (GCE)...', '05. Google Compute Engine (GCE)...', '06. Container Management\_GKE ...', '06. Container Management\_GKE ...', '07. Data Architecture on Google Cl...', and '07. Data Architecture on Google Cl...'. Each file entry shows its name, size, type, creation date, and storage class.

Name	Size	Type	Created	Storage class
01. Introducing Google Cloud.pdf	4 MB	application/pdf	Jan 29, 2026, 11:54:22 AM	Standard
01. Introducing Google Cloud.pptx	27.3 MB	application/vnd.openxmlformats-officedocument.presentationml.presentation	Jan 29, 2026, 11:54:22 AM	Standard
02. Resources and Access in the C...	3.4 MB	application/pdf	Jan 29, 2026, 11:54:23 AM	Standard
02. Resources and Access in the C...	10 MB	application/vnd.openxmlformats-officedocument.presentationml.presentation	Jan 29, 2026, 11:54:23 AM	Standard
03. Networking in Google Cloud.pdf	4.5 MB	application/pdf	Jan 29, 2026, 11:54:24 AM	Standard
03. Networking in Google Cloud.pp...	17 MB	application/vnd.openxmlformats-officedocument.presentationml.presentation	Jan 29, 2026, 11:54:25 AM	Standard
04. Advanced Networking.pdf	7.1 MB	application/pdf	Jan 29, 2026, 11:54:25 AM	Standard
04. Advanced Networking.pptx	24.1 MB	application/vnd.openxmlformats-officedocument.presentationml.presentation	Jan 29, 2026, 11:54:25 AM	Standard
05. Google Compute Engine (GCE)...	3.1 MB	application/pdf	Jan 29, 2026, 11:54:25 AM	Standard
05. Google Compute Engine (GCE)...	15 MB	application/vnd.openxmlformats-officedocument.presentationml.presentation	Jan 29, 2026, 11:54:26 AM	Standard
06. Container Management_GKE ...	5.4 MB	application/pdf	Jan 29, 2026, 11:54:27 AM	Standard
06. Container Management_GKE ...	22.8 MB	application/vnd.openxmlformats-officedocument.presentationml.presentation	Jan 29, 2026, 11:54:27 AM	Standard
07. Data Architecture on Google Cl...	10.7 MB	application/pdf	Jan 29, 2026, 11:54:28 AM	Standard
07. Data Architecture on Google Cl...	37.5 MB	application/vnd.openxmlformats-officedocument.presentationml.presentation	Jan 29, 2026, 11:54:28 AM	Standard

The screenshot shows the 'Object details' page for the object '01. Introducing Google Cloud.pdf' in the 'coe-lecture-material' bucket. The object is a 'Live object' with a 'Standard' storage class. The 'Overview' section shows the object's type as 'application/pdf', size as '4 MB', and creation/modification dates as 'Jan 29, 2026, 11:54:22 AM'. The 'Public URL' and 'Authenticated URL' are highlighted with a red box. The 'Permissions' section shows 'Public access' set to 'Access granted to public principals'. The 'Protection' section shows 'Version history' as 'None', 'Retention expiration time' as 'None', and 'Object retention retain until time' as 'None'. The 'Bucket retention retain until time' is also 'None'. The 'Hold status' is 'None' and the 'Encryption type' is 'Google-managed'.

Property	Value
Type	application/pdf
Size	4 MB
Created	Jan 29, 2026, 11:54:22 AM
Last modified	Jan 29, 2026, 11:54:22 AM
Storage class	Standard
Custom time	—
Public URL	<a href="https://storage.googleapis.com/coe-lecture-material/01.%20Introducing%20Google%20Cloud.pdf">https://storage.googleapis.com/coe-lecture-material/01.%20Introducing%20Google%20Cloud.pdf</a>
Authenticated URL	<a href="https://storage.cloud.google.com/coe-lecture-material/01.%20Introducing%20Google%20Cloud.pdf">https://storage.cloud.google.com/coe-lecture-material/01.%20Introducing%20Google%20Cloud.pdf</a>
gsutil URI	gs://coe-lecture-material/01. Introducing Google Cloud.pdf
Permissions	Public access: Access granted to public principals
Protection	Version history: —
Retention expiration time	None
Object retention retain until time	None
Bucket retention retain until time	None
Hold status	None
Encryption type	Google-managed

# Agenda

01 The Data Analytics Lifecycle Overview

---

02 Ingest

---

03 Process

---

04 Store

---

## Analyze

Introduction to BigQuery

Querying External Data Sources or Public Datasets **Demo**

05 Control access with Policy Tag **Demo**

Building Pipelines in BigQuery - from low-code to coding method **Demo**

Visualization & Business Intelligence with Looker & Looker Studio **Demo**

Derive insights from your data in BigQuery (Agent, Gemini Enterprise) **Demo**

---

## Activate

06 Introduction AI on Google Cloud

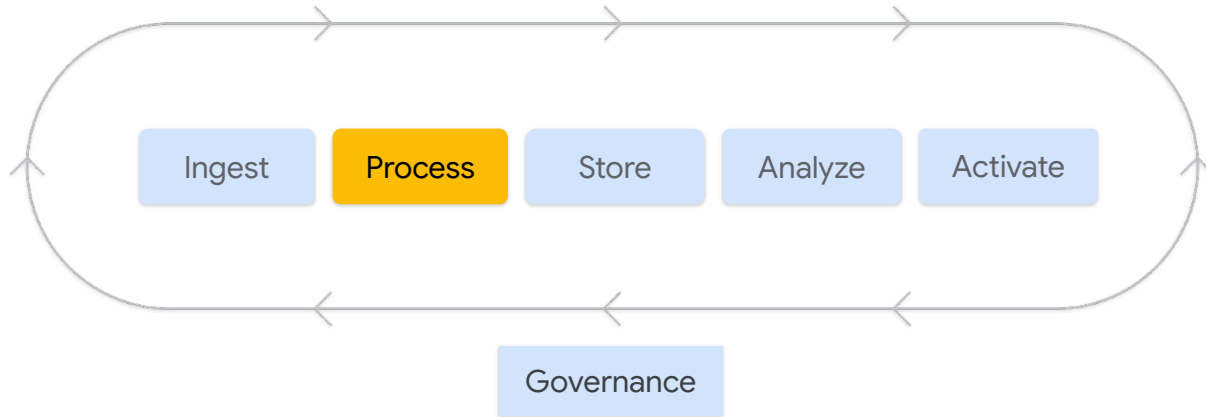
BigQuery ML **Demo**

---

07 Governance

---

# Data analytics lifecycle: Process



**Objective:** Convert raw data into governed, modeled datasets.

**This layer focuses on:**

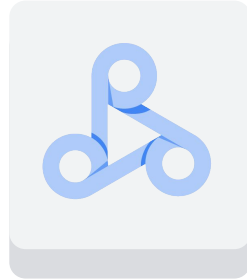
- Data cleaning
- Transformation logic
- Develop & Maintenance of data processing pipelines
- Business logic traceability

**Related Tools:**

- Dataproc
- Dataflow
- Data Fusion
- BigQuery

# Data analytics lifecycle: Process

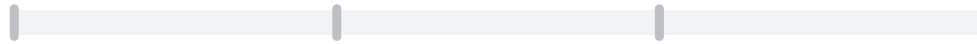
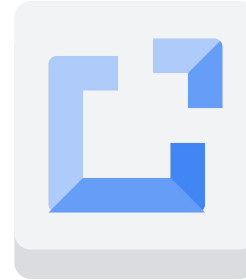
Dataproc



Dataflow

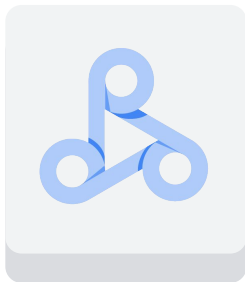


Data Fusion



# Data analytics lifecycle: Process

Dataproc



- ✓ Managed Hadoop (Spark, Hive, Pig, etc)
- ✓ Manager - Worker architecture
- ✓ Process based on MapReduce model
- ✓ Reliable auto scaling
- ✓ Various templates

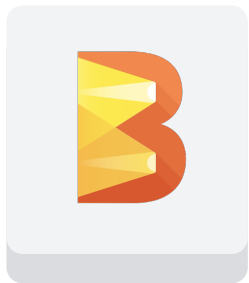
# Data analytics lifecycle: Process

Dataflow



- ✓ Managed Apache Beam
- ✓ Handles both batch and streaming data
- ✓ Reliable auto scaling
- ✓ Serverless
- ✓ Various templates

# Apache Beam is a popular solution for pipeline design



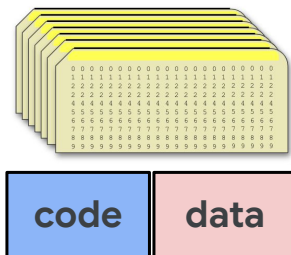
An open source, unified programming model to **define** and **execute** data processing pipelines, including ETL, batch, and stream processing.

Supports **Java**, **Python**, and **Go** SDKs.

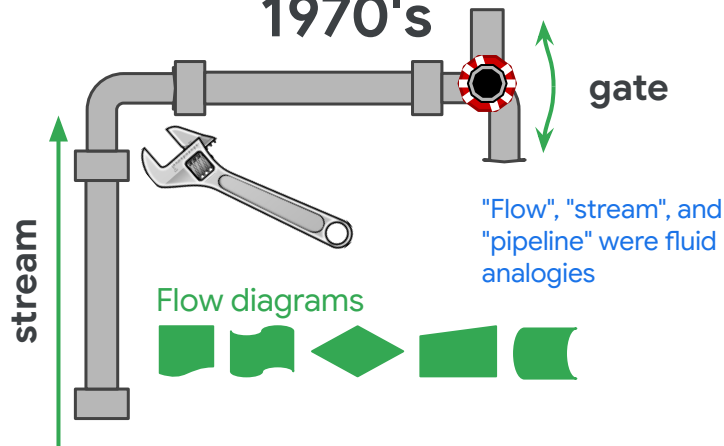
# Batch programming and data processing used to be two very separate and different things

## Batch programming 1940's

"Batch" originally referred to a batch or box of punch-cards

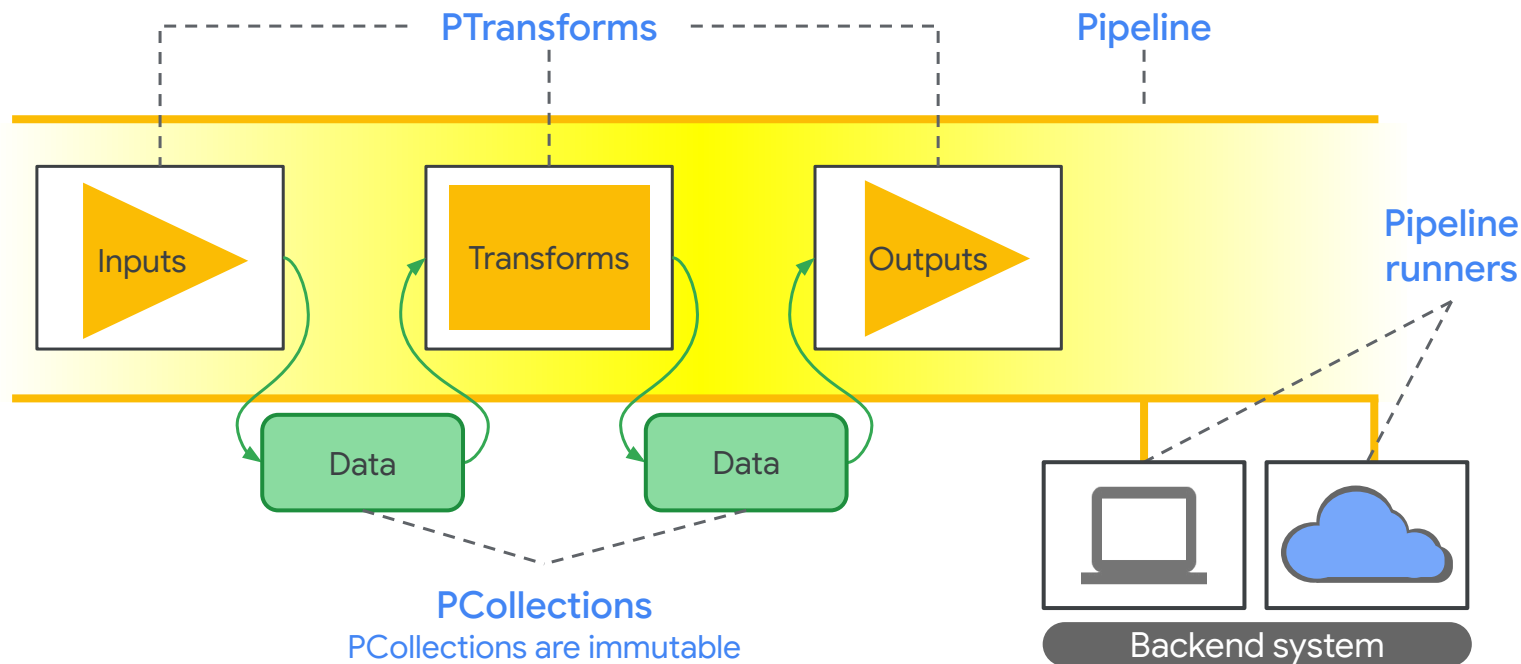


## Data processing 1970's



*Different tools, different platforms, different concepts, different methods.*

# Apache BEAM = Batch + strEAM

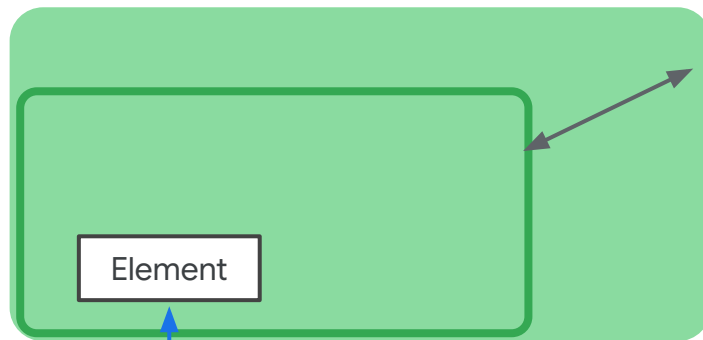


# A PCollection represents batch or stream data

Bounded PCollection



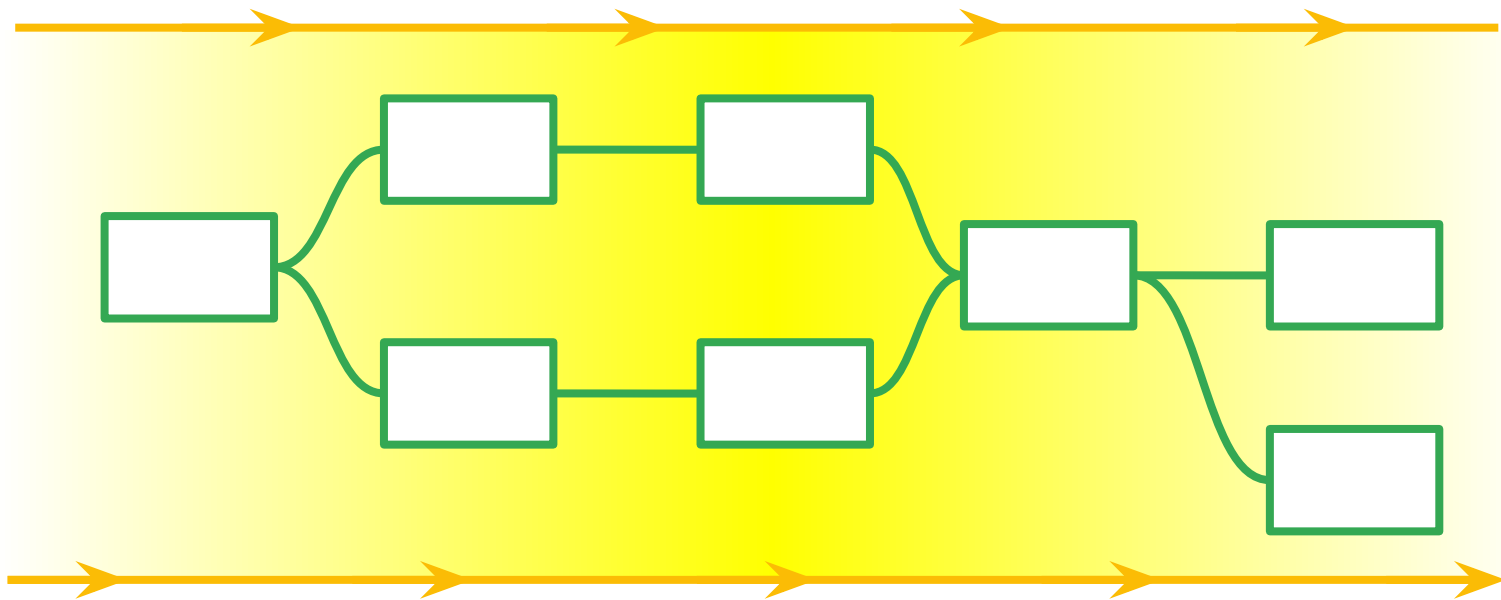
Unbounded PCollection



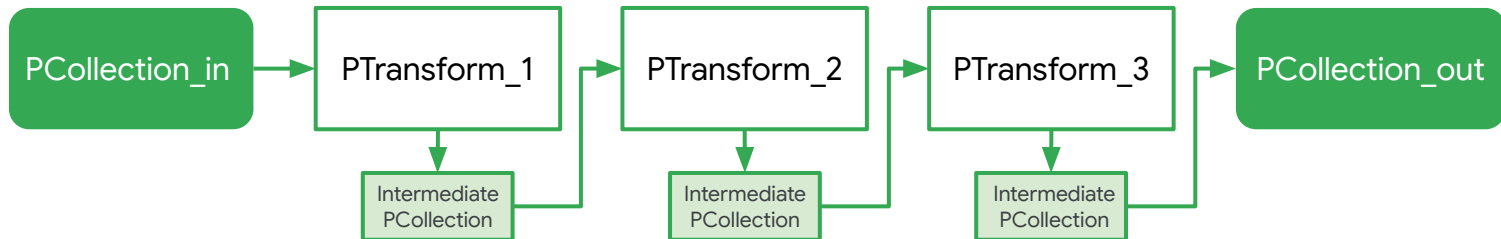
All data types are stored  
as serialized byte strings

**Note:** Bounded means the data has a fixed size not that the PCollection size is limited. A PCollection can be any size and be distributed across many workers.

# A pipeline is a directed graph of steps



# How to construct a simple pipeline



## Python

Python overloads  
the pipe operator

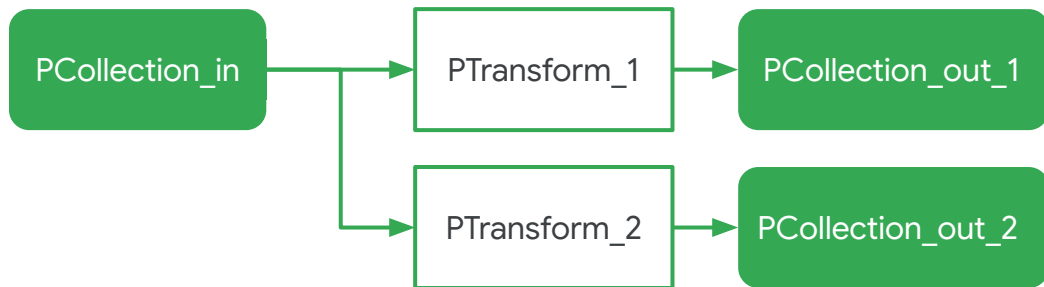
```
PCollection_out = (PCollection_in | PTransform_1  
                  | PTransform_2  
                  | PTransform_3)
```

## Java

Java uses the  
.apply method

```
PCollection_out = PCollection_in.apply(PTransform_1)  
                             .apply(PTransform_2)  
                             .apply(PTransform_3)
```

# How to construct a branching pipeline



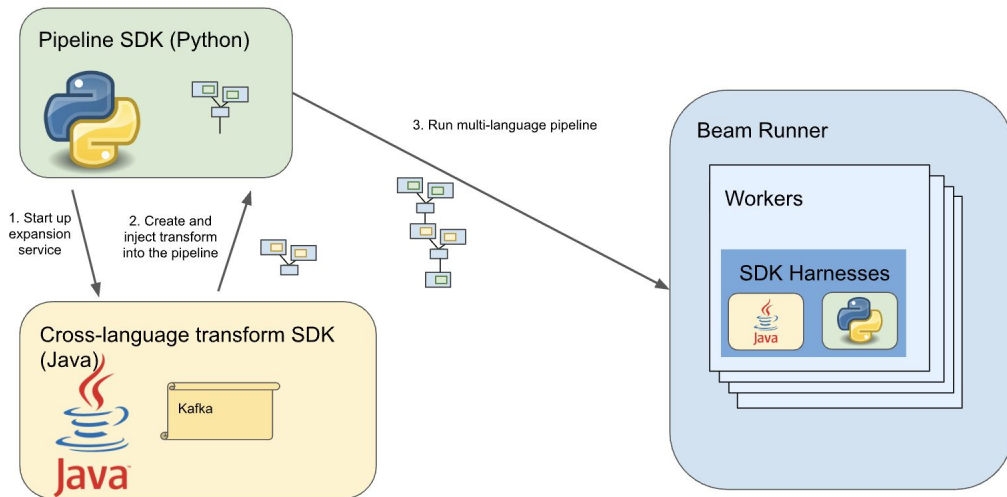
## Python

```
PCollection_out_1 = PCollection_in | PTransform_1  
PCollection_out_2 = PCollection_in | PTransform_2
```

## Java

```
PCollection_out_1 = PCollection_in.apply(PTransform_1)  
PCollection_out_2 = PCollection_in.apply(PTransform_2)
```

# Cross language pipeline



PYTHON

```

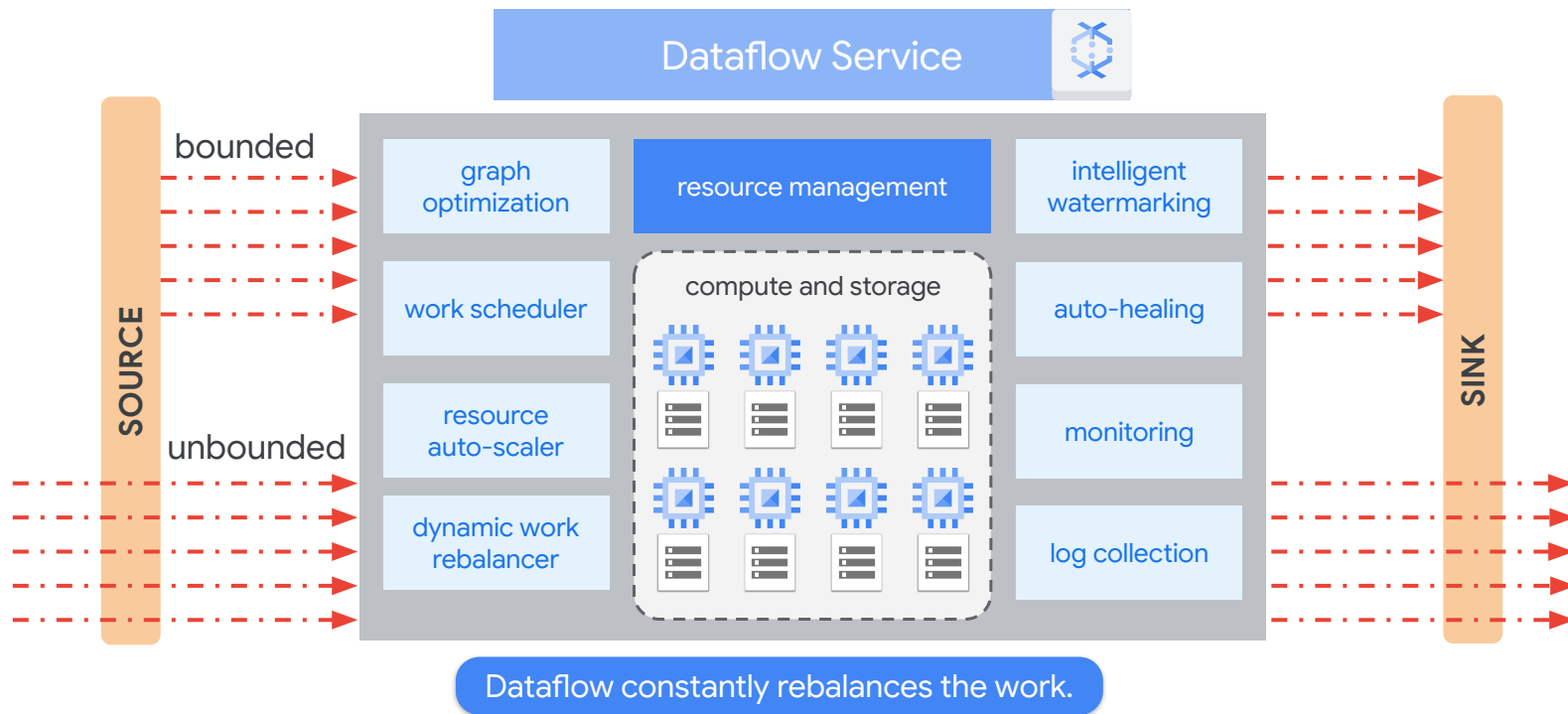
# custom_java_transform.py
# Call a custom Java transform from a Python pipeline
import apache_beam as beam
from apache_beam.transforms.external import JavaExternalTransform

# Define the Java transform URN and parameters
# The URN identifies the transform in the expansion service
java_transform = JavaExternalTransform(
    "beam:transform:org.example:my_custom_transform:v1",
    expansion_service="localhost:8097",
)

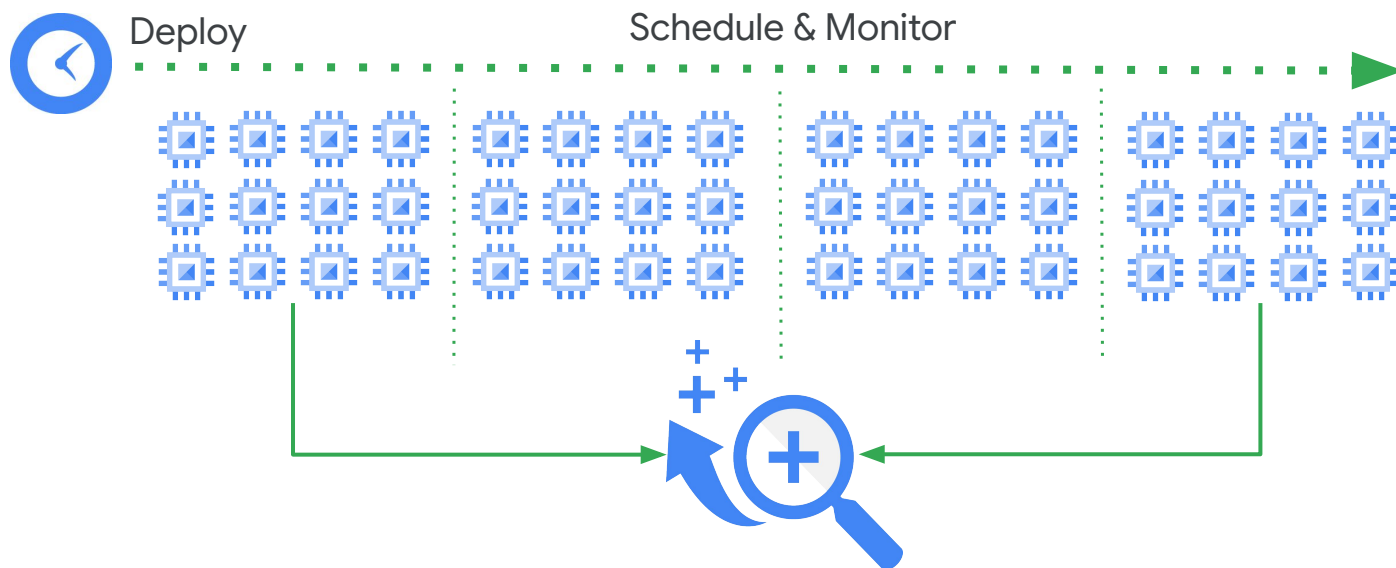
with beam.Pipeline(options=options) as pipeline:
    (
        pipeline
        | "ReadInput" >> beam.io.ReadFromText("gs://my-bucket/input.txt")
        | "ProcessInPython" >> beam.Map(preprocess)
        # Hand off to Java for specialized processing
        | "JavaProcessing" >> java_transform
        # Continue processing in Python
        | "PostProcess" >> beam.Map(postprocess)
        | "WriteOutput" >> beam.io.WriteToText("gs://my-bucket/output/")
    )

```

# How does Dataflow work?

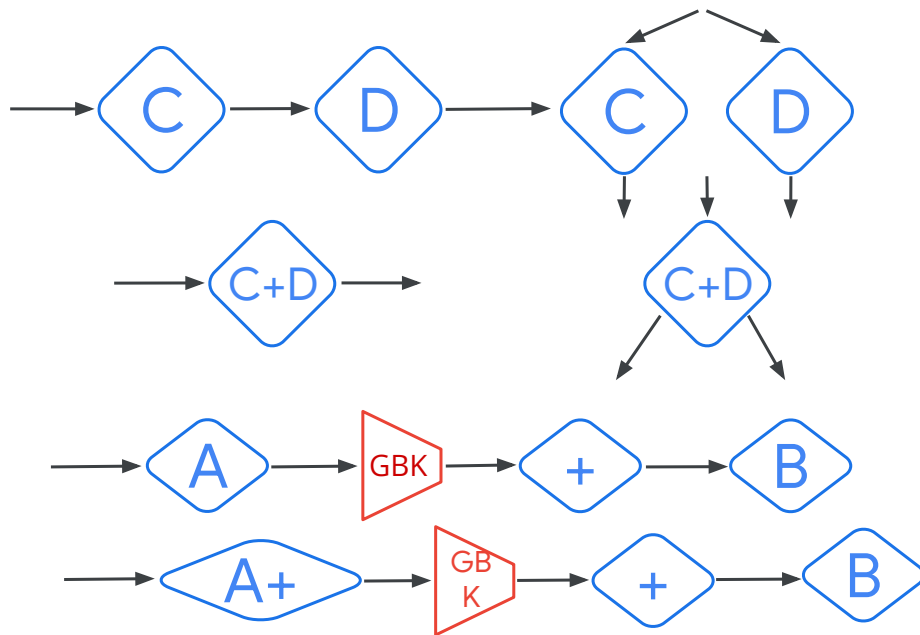


# Why customers value Dataflow: Fully-managed and auto-configured

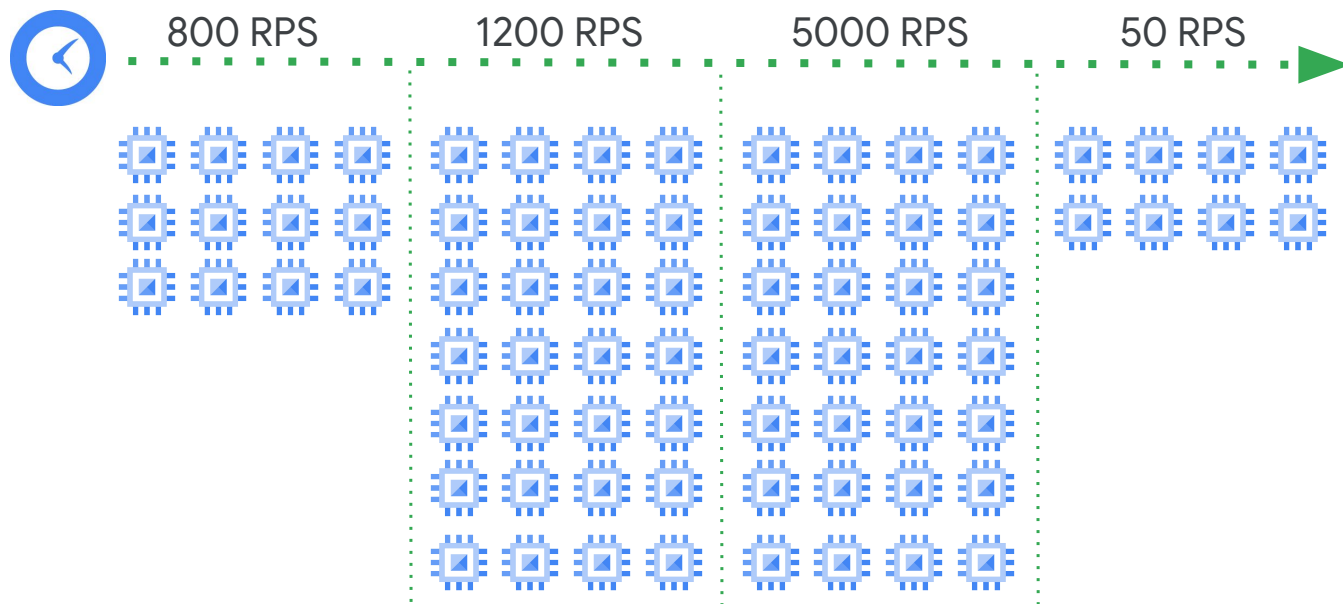


# Why customers value Dataflow: Graph is optimized for best execution path

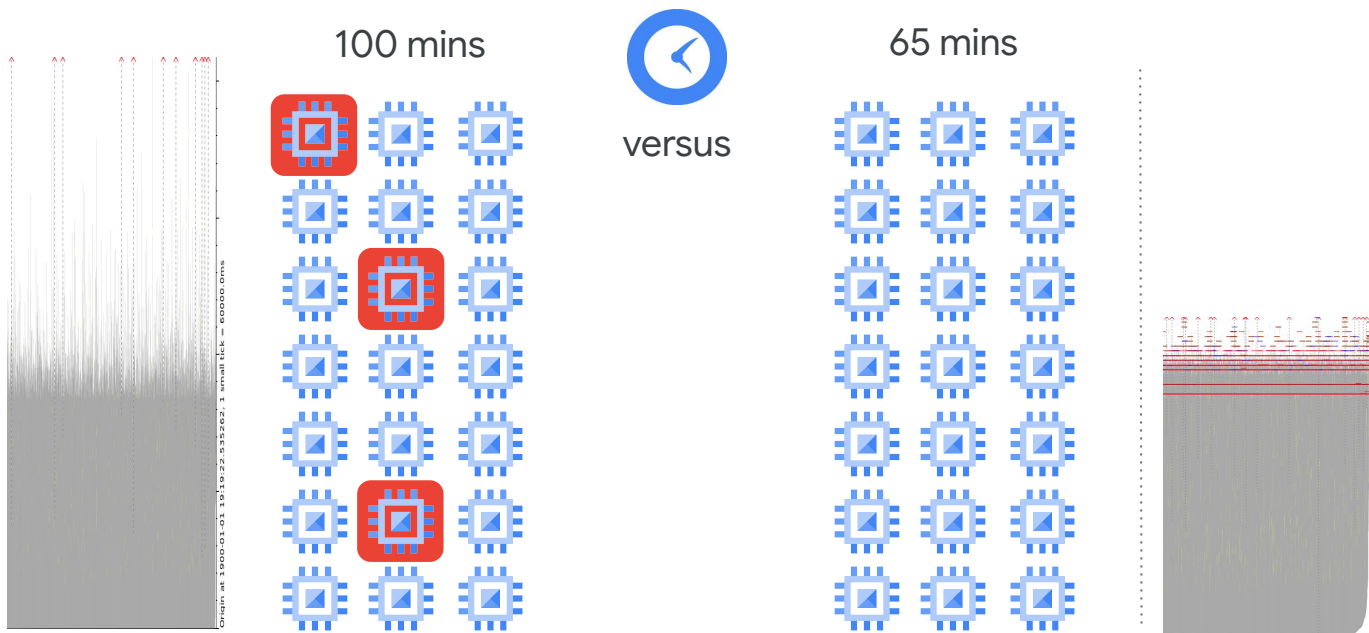
- It optimize the pipeline execution steps
- It doesn't wait for a previous step to finish before starting a new step



# Why customers value Dataflow: Autoscaling mid-job



# Why customers value Dataflow: Dynamic work rebalancing mid-job



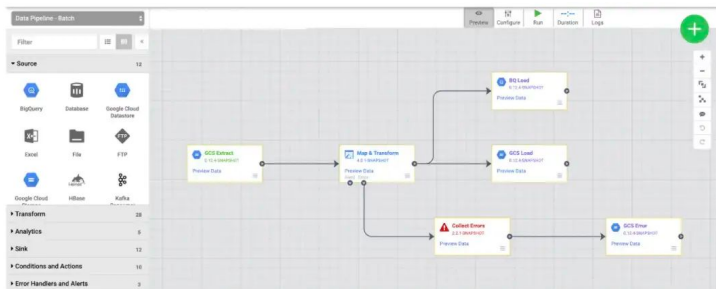
# Demo: Dataflow

The screenshot shows the Google Cloud Dataflow console interface. At the top, the breadcrumb navigation reads 'Dataflow / Jobs / Dataflow job details'. The job name is 'test'. The job status is 'Succeeded' with a green checkmark. The job ID is '2026-01-26\_21\_40\_00-3970752040866938414'. The job type is 'Batch' and the SDK version is 'Apache Beam SDK for Java 2.70.0'. The job region is 'asia-east1'. The job started on 'January 27, 2026, 8:40:02 AM GMT+3' and elapsed for '6 min 27 sec'. The job graph shows four stages: 'ReadLines' (4 sec, 2 of 2 stages succeeded), 'WordCount...ountWords' (1 sec, 2 of 2 stages succeeded), 'MapElements' (0 sec, 1 of 1 stage succeeded), and 'WriteCounts' (1 sec, 5 of 5 stages succeeded). The 'Resource metrics' section shows: Current vCPUs: 1, Total vCPU time: 0.033 vCPU hr, Current memory: 3.75 GB, Total memory time: 0.122 GB hr, Current HDD PD: 25 GB, and Total HDD PD time: 0.816 GB hr.

```
pipeline_template.py
444 def run():
445     # Create the pipeline
446     with beam.Pipeline(options=options) as pipeline:
447         # Step 1: Read input file paths
448         files = pipeline | "Input File Paths" >> beam.Create([known_args.inputFilePath])
449
450         # Step 2: Data transformation with error handling
451         # Read CSV files and add row numbers for error tracking
452         # Note: Null bytes are cleaned in ReadCSVWithRowNumbers, but we add an extra
453         # cleaning step as a safety measure for any remaining null bytes
454         lines_with_row_numbers = (
455             files
456             | "List Files" >> beam.ParDo(ListGCSFiles())
457             | "Reshuffle" >> beam.Reshuffle()
458             | "Read CSV with Row Numbers" >> beam.ParDo(ReadCSVWithRowNumbers())
459             | "Remove Null Bytes" >> beam.Map(lambda x: (x[0], x[1], TransformationContentWith
460         ))
461
462         # Transform with error handling
463         transformation_contents = (
464             lines_with_row_numbers
465             | "Main Transformation with Error Handling"
466             >> beam.ParDo(
467                 TransformationContentWithErrorHandling(
468                     known_args.schemaJSONPath,
469                     known_args.hashSalt,
470                     5000,
471                     known_args.fieldSizeLimitMB
472                 )
473             )
474         )
475
476         # Step 5: Write to BigQuery Final Table (only after staging is complete)
477         write_to_final_result = (
478             transformation_contents
479             | "Write to BigQuery Final Table" >> WriteToBigQuery(
480                 known_args.outputTable,
481                 schema="SCHEMA_AUTODETECT",
482                 # schema=GCSUtils._load_schema(known_args.schemaJSONPath),
483                 write_disposition=(
484                     beam.io.BigQueryDisposition.WRITE_TRUNCATE
485                     if known_args.isTruncate == "true"
486                     else beam.io.BigQueryDisposition.WRITE_APPEND
487                 ),
488                 create_disposition=beam.io.BigQueryDisposition.CREATE_IF_NEEDED,
489                 additional_bq_parameters=additional_bq_parameters,
490                 custom_gcs_temp_location="gs://dataflow-staging-asia-east1-587065697986/tmp",
491                 kms_key=regional_kms_key,
492             )
493         )
494     )
```

# Data analytics lifecycle: Process

## Data Fusion





✓ Managed CDAP for pipeline portability

✓ Designed for data integration

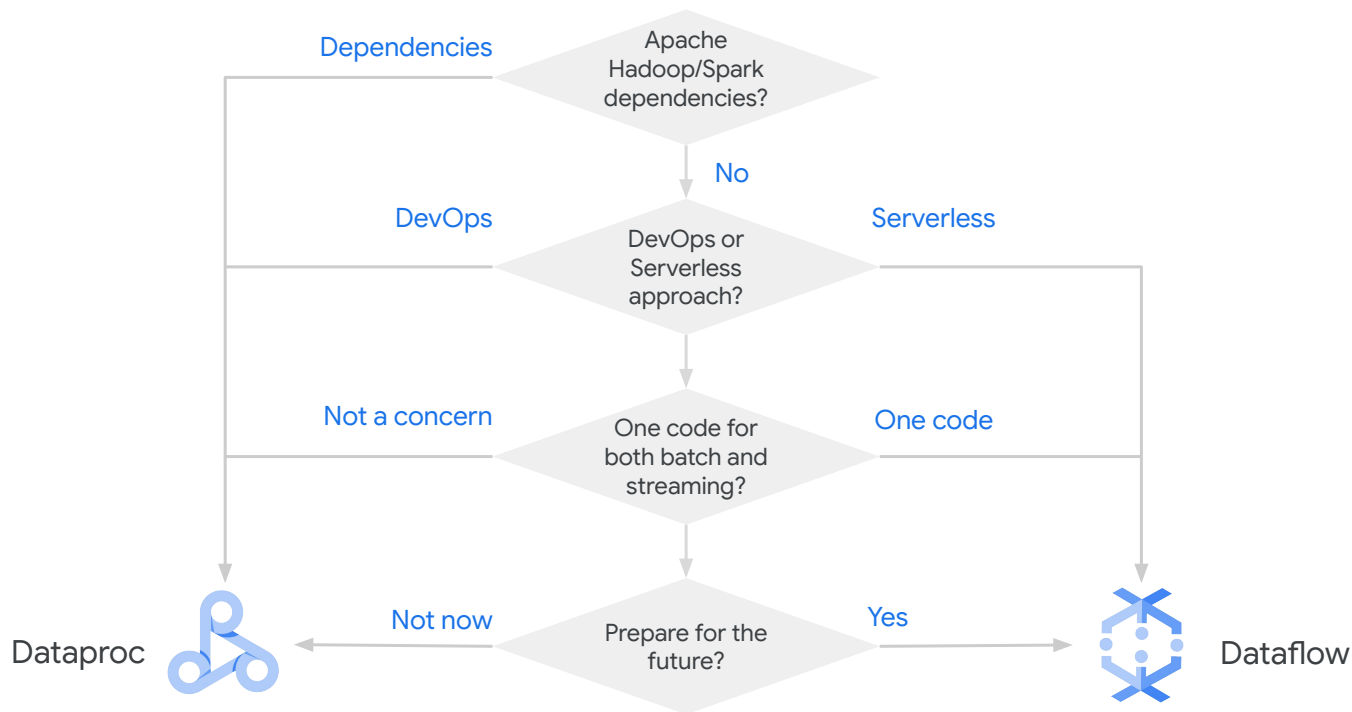
✓ GUI-based, code-free deployment

✓ Broad library of preconfigured connectors and transformations

# Dataflow versus Dataproc

	 Dataflow	 Dataproc
Recommended for:	New data processing pipelines, unified batch and streaming	Existing Hadoop/Spark applications, machine learning/data science ecosystem, large-batch jobs, preemptible VMs
Serverless:	Yes	No
Auto-scaling:	Yes, transform-by-transform (adaptive)	Yes, based on cluster utilization (reactive)
Expertise:	Apache Beam	Hadoop, Hive, Pig, Apache Big Data ecosystem, Spark, Flink, Presto, Druid

# Choosing between Dataflow and Dataproc



# Agenda

01 The Data Analytics Lifecycle Overview

---

02 Ingest

---

03 Process

---

04 Store

---

## Analyze

Introduction to BigQuery

Querying External Data Sources or Public Datasets **Demo**

Control access with Policy Tag **Demo**

Building Pipelines in BigQuery - from low-code to coding method **Demo**

Visualization & Business Intelligence with Looker & Looker Studio **Demo**

Derive insights from your data in BigQuery (Agent, Gemini Enterprise) **Demo**

---

06

## Activate

Introduction AI on Google Cloud

BigQuery ML **Demo**

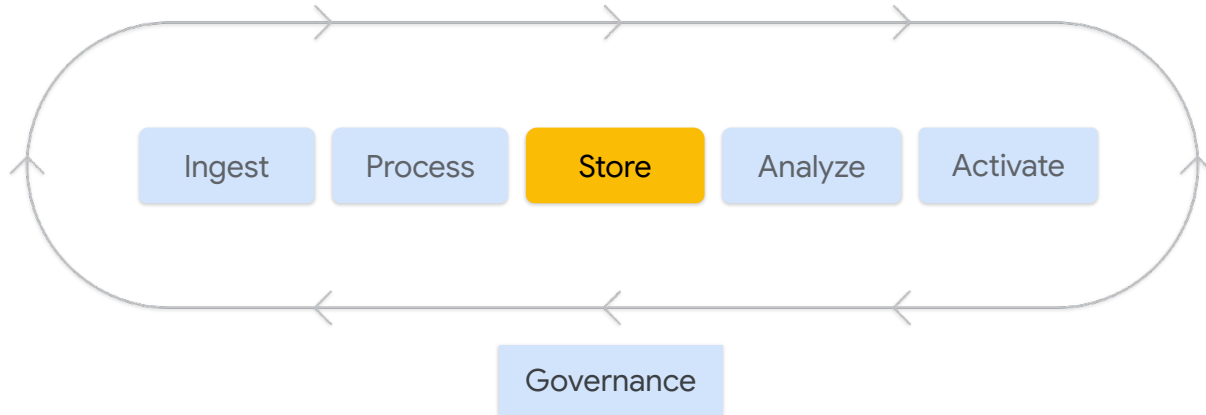
---

07

Governance

---

# Data analytics lifecycle: Store



**Objective:** Data scales with company growth without cost or performance risk.

**This layer focuses on:**

- Scalable storage architecture
- Cost and performance optimization
- Data organization and discoverability
- Layered storage (landing, cleansed, curated layer)

**Related Tools:**

- Cloud Storage
- Cloud SQL
- Spanner
- AlloyDB
- Firestore
- Bigtable
- BigQuery

# Data can have different formats

Storage/process data

✓ Data type

✓ Business need

Unstructured



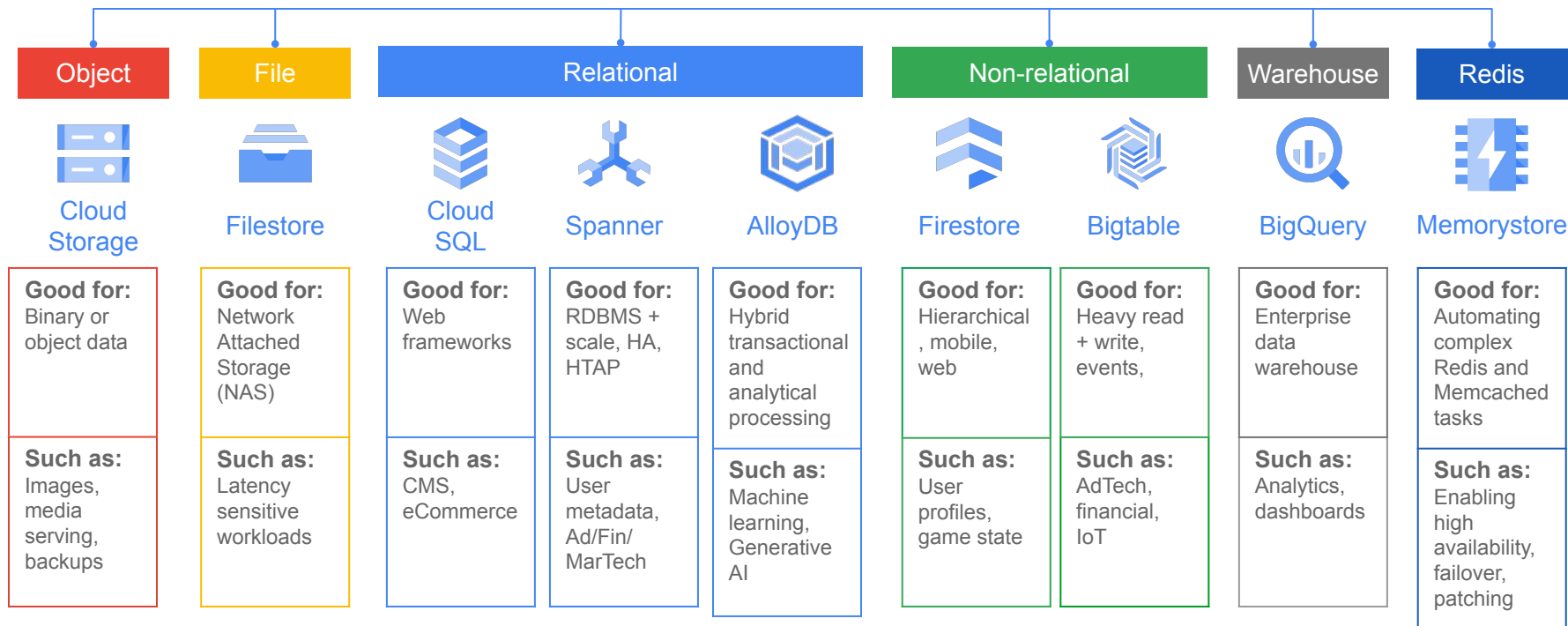
- ◆ Documents
- ▲ Images
- Audio files

Structured

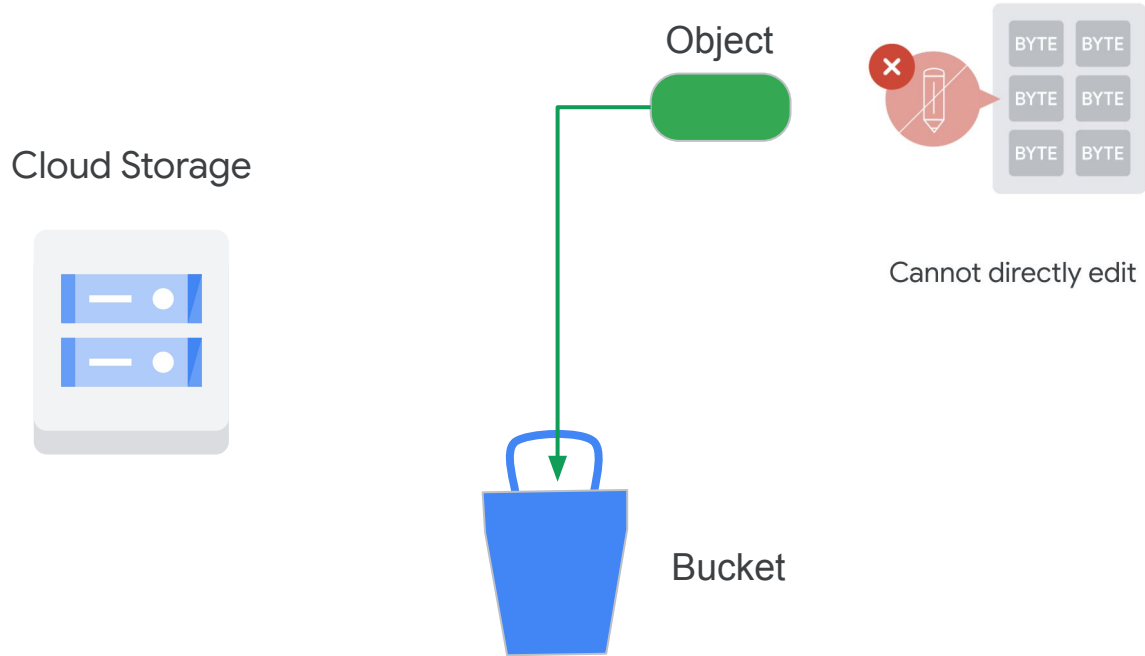


- ▬ Tables
  - Rows
  - Columns

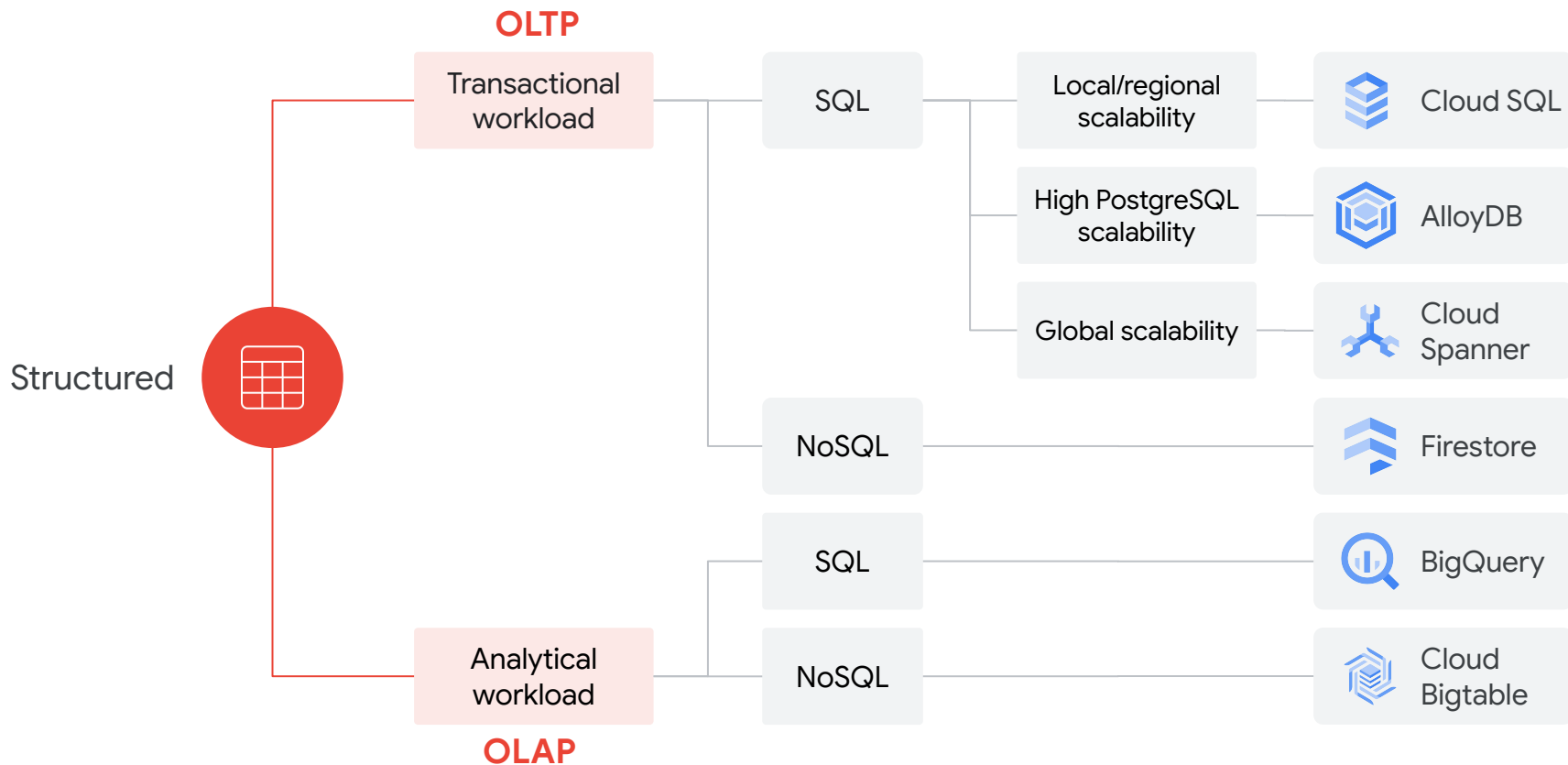
# Storage and database services



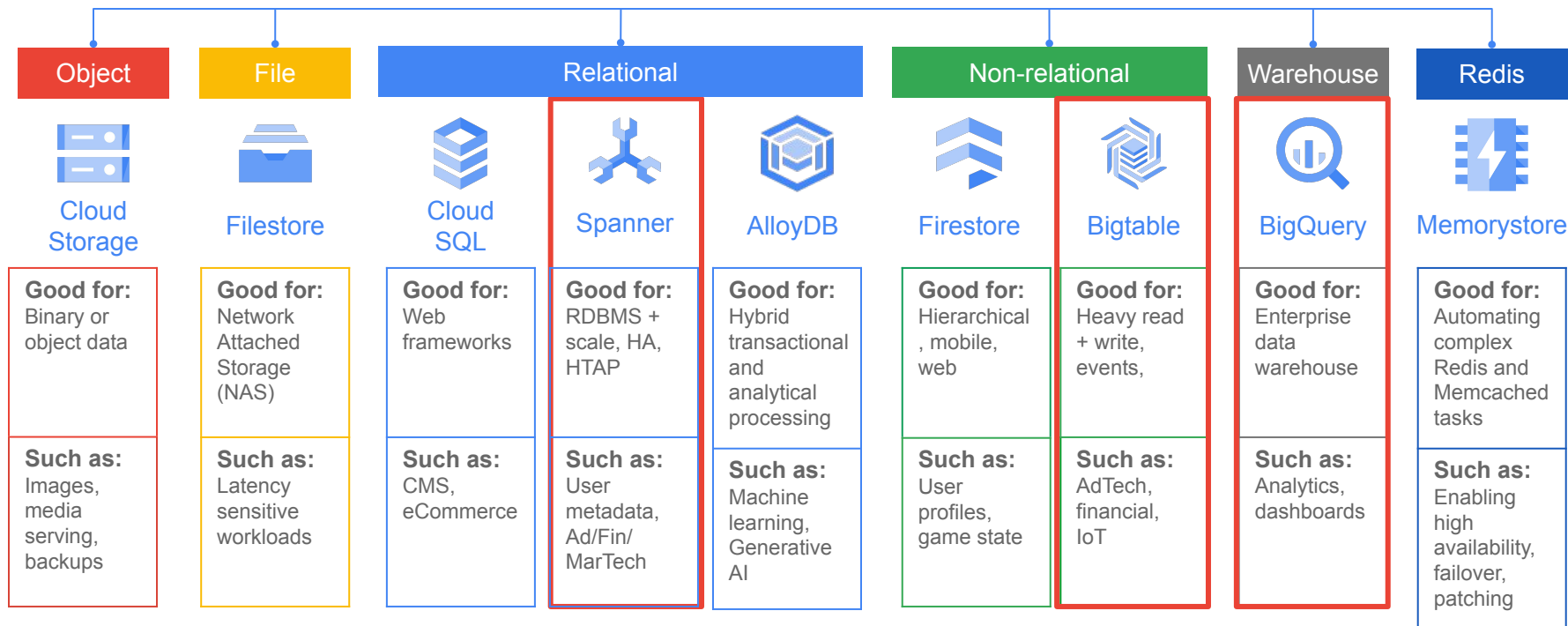
# Options for storing unstructured data



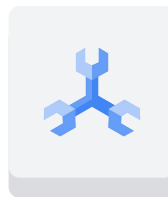
# Options for storing structured data



# Storage and database services





















# Spanner combines the benefits of relational database structure with non-relational horizontal scale



Spanner

- Virtually unlimited database
- powers planet-scale applications like Gmail, YouTube, and Google Ads
- Strong consistency (global)
- High availability
- Used for financial and inventory applications

# Characteristics

	Spanner		Relational DB		Non-Relational DB	
Schema		Yes		Yes		No
SQL		Yes		Yes		No
Consistency		Strong		Strong		Eventual
Availability		High		Failover		High
Scalability		Horizontal		Vertical		Horizontal
Replication		Automatic		Configurable		Configurable

# Spanner

Google Cloud

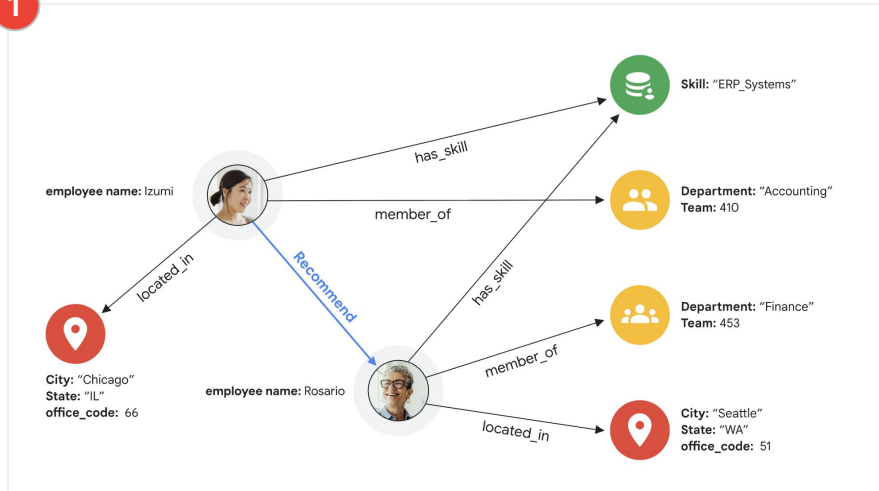
**Spanner:**  
The always-on,  
virtually unlimited  
scale database



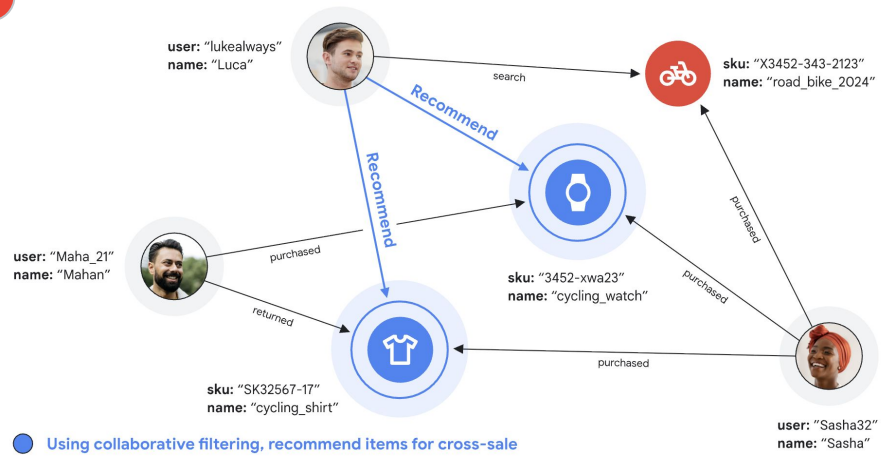
# Use Case

1. Knowledge Graph
2. Recommendation Engine
3. Financial Fraud Detection

1

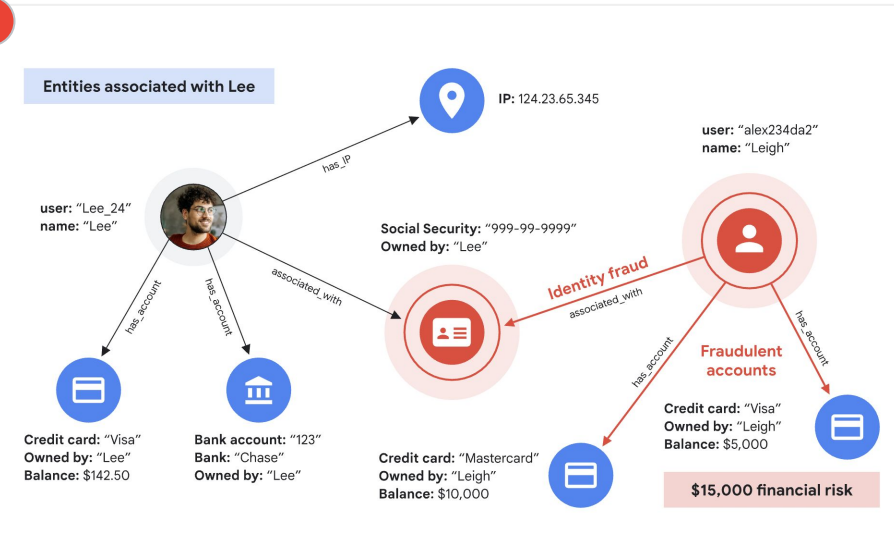


2



Using collaborative filtering, recommend items for cross-sale

3



## Use Case: Palo Alto

**Palo Alto Networks**

**is building**



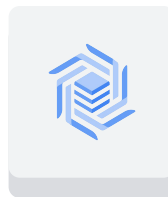
**the future of**



**cybersecurity**



# Bigtable is a NoSQL big data database service



Bigtable

- Petabyte-scale
- **Support high read/write throughput at sub-10ms latency**
- Learns and adjusts to access patterns
- Ideal for Ad Tech, Fintech, and IoT
- Storage engine for ML applications
- Easy integration with open source big data tools



# Intro to BigTable

Cloud Bigtable  
Fundamentals  
**Intro to  
BigTable**



# Bigtable storage model

"follows" column family

	Follows			
Row Key	gwasington	jadams	tjefferson	wmckinley
gwasington		1		
jadams	1		1	
tjefferson	1	1		1
wmckinley			1	

multiple versions

# Agenda

01 The Data Analytics Lifecycle Overview

---

02 Ingest

---

03 Process

---

04 Store

---

## Analyze

Introduction to BigQuery

Querying External Data Sources or Public Datasets **Demo**

Control access with Policy Tag **Demo**

Building Pipelines in BigQuery - from low-code to coding method **Demo**

Visualization & Business Intelligence with Looker & Looker Studio **Demo**

Derive insights from your data in BigQuery (Agent, Gemini Enterprise) **Demo**

---

06

## Activate

Introduction AI on Google Cloud

BigQuery ML **Demo**

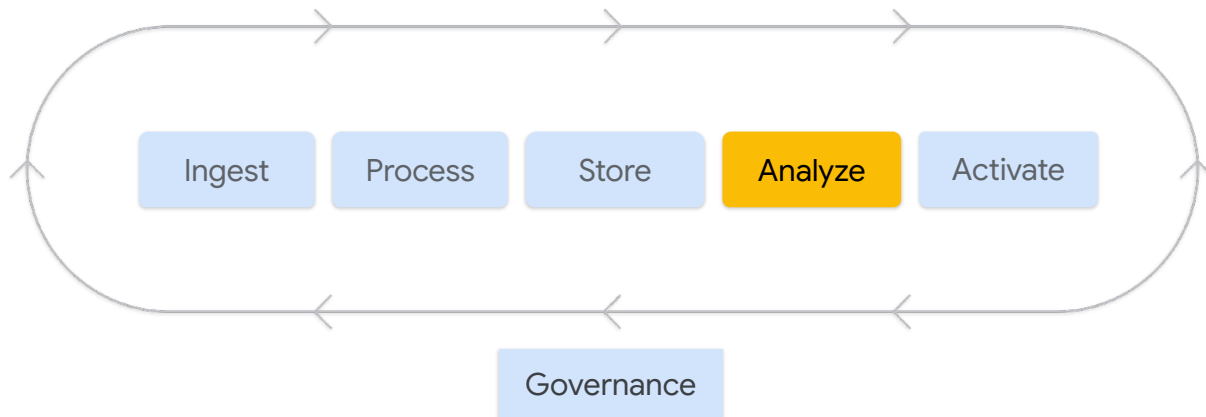
---

07

Governance

---

# Data analytics lifecycle: Analyze



**Objective:** Get insights from data.

**This layer focuses on:**

- Query performance
- Self-service analytics capability
- Insight reliability
- Consistent metric definitions

**Related Tools:**

- BigQuery
- Looker
- Looker Studio
- AI Agents

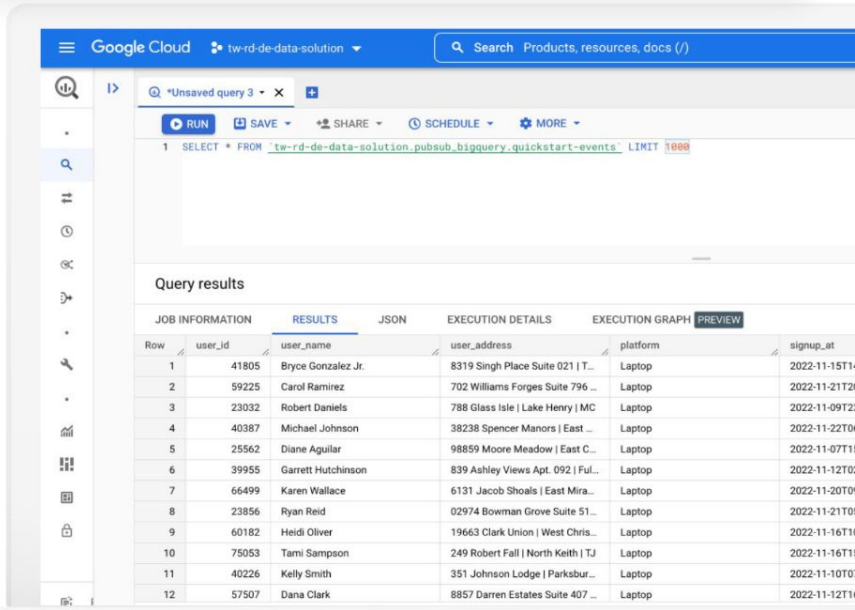
# Introduction to BigQuery



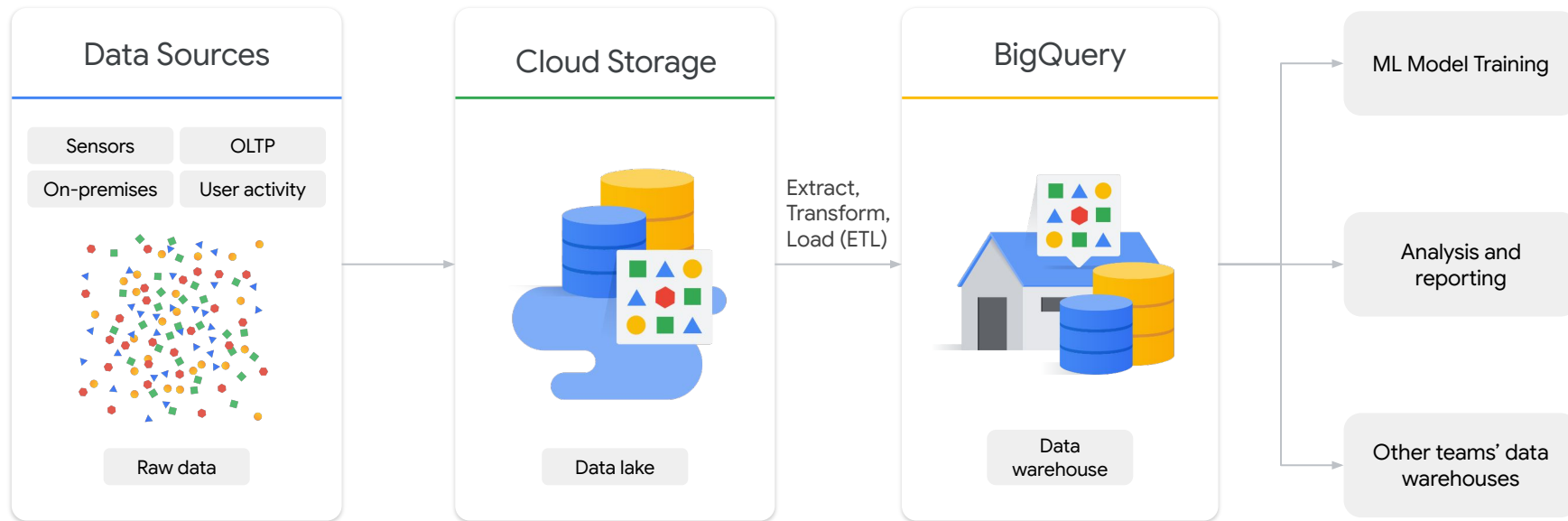
# BigQuery Overview

Serverless multi-cloud enterprise data warehousing services can provide robust support for data-driven innovative operations.

- ✔ Real-time insights from batch or streaming data sources
- ✔ Built-in ML for ready-to-use predictive functionality
- ✔ Encrypted, durable, and highly available
- ✔ Fully managed service for maximum agility
- ✔ Storage and SQL querying from GB to PB scale
- ✔ High-speed, in-memory BI engine for faster reporting and analysis



# BigQuery stores your data for analysis and reporting



# Scales to your needs On-demand

## Process terabytes of data in Seconds

# 1.61 TB

## (~1 billion rows)

data processed in

# 11 secs

translate to higher operational

## Proficiency

Q BQ\_TB\_performance\_demo\_bitcoin\_mom\_trans... RUN

```
1 SELECT
2 block_timestamp_month,
3 count(*) as tranx_count,
4 ANY_VALUE(inputs), ANY_VALUE(outputs)
5 FROM `bigquery-public-data.crypto_bitcoin.transactions` as transactions
6 group by block_timestamp_month
```

### Query results

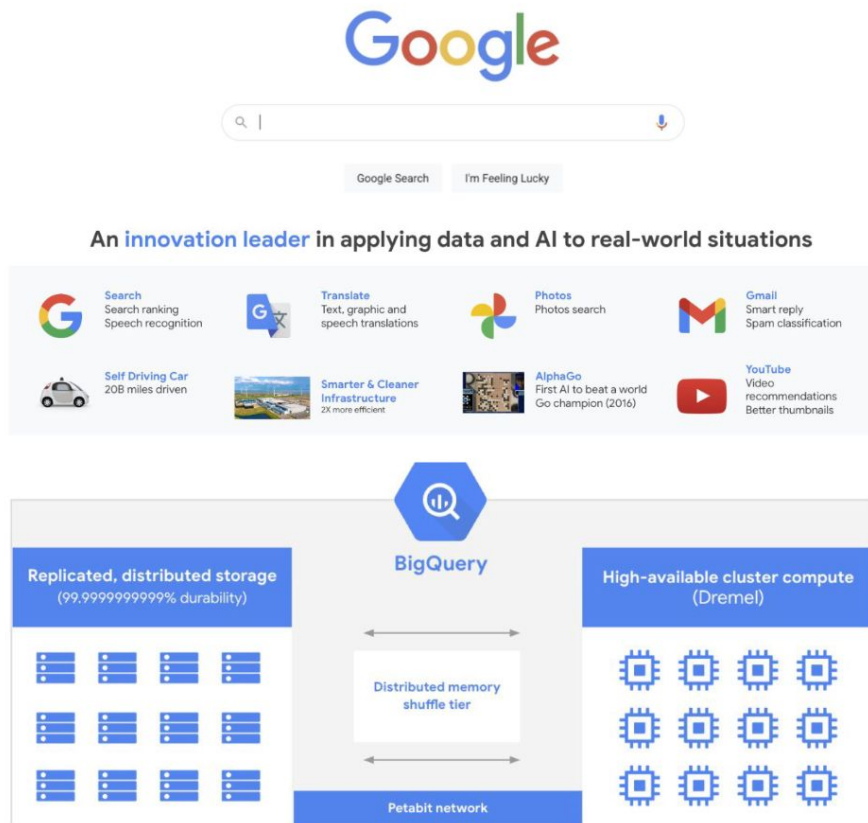
JOB INFORMATION	RESULTS	CHART	JSON	EXECUTION DETAIL
Job ID	sg-rd-sa-chenglong-wu:US.bq:job_186a4a7e_18ea82cc0cc			
User	chenglong.wu@mile.cloud			
Location	US			
Creation time	Apr 4, 2024, 4:14:27 PM UTC+8			
Start time	Apr 4, 2024, 4:14:27 PM UTC+8			
End time	Apr 4, 2024, 4:14:39 PM UTC+8			
Duration	11 sec			
Bytes processed	1.61 TB			
Bytes billed	1.61 TB			
Slot milliseconds	18649776			
Job priority	INTERACTIVE			
Use legacy SQL	false			
Destination table	<a href="#">Temporary table</a>			
Labels				

# Why is BigQuery so fast and scalable?

Same **proprietary** technology that is being used by Google internally, which also used to power Google Search, YouTube, Gmail, etc.

**Decoupled** compute and storage cluster, which scales independently.

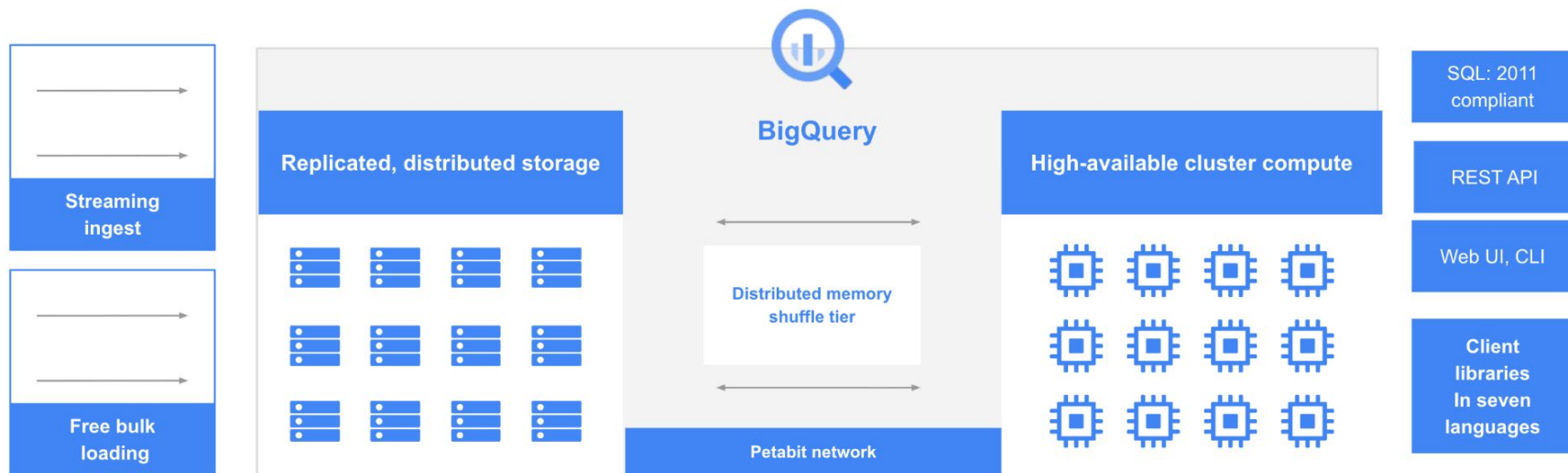
**Remote** shuffle memory cluster to speed up the queries



# BigQuery Architecture

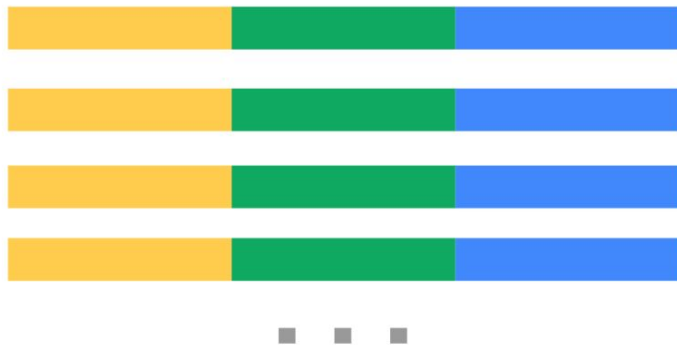
# BigQuery | Architecture

Decoupled storage and compute



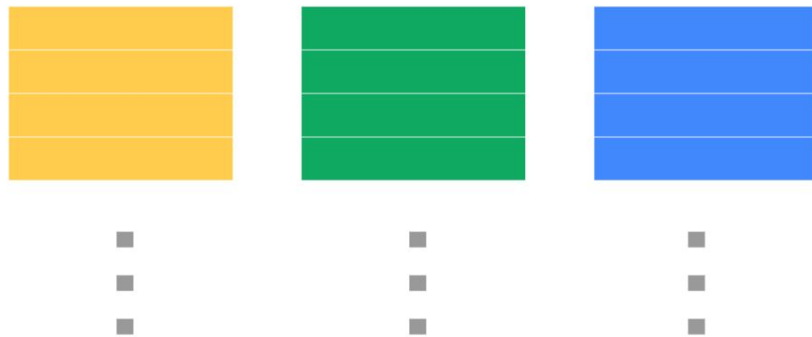
# BigQuery | Columnar Storage

Traditional RDBMS storage



Record-oriented storage  
OLTP

BigQuery storage



Columnar storage  
OLAP

# BigQuery | Managed Storage

Tables are stored in optimized columnar format

---

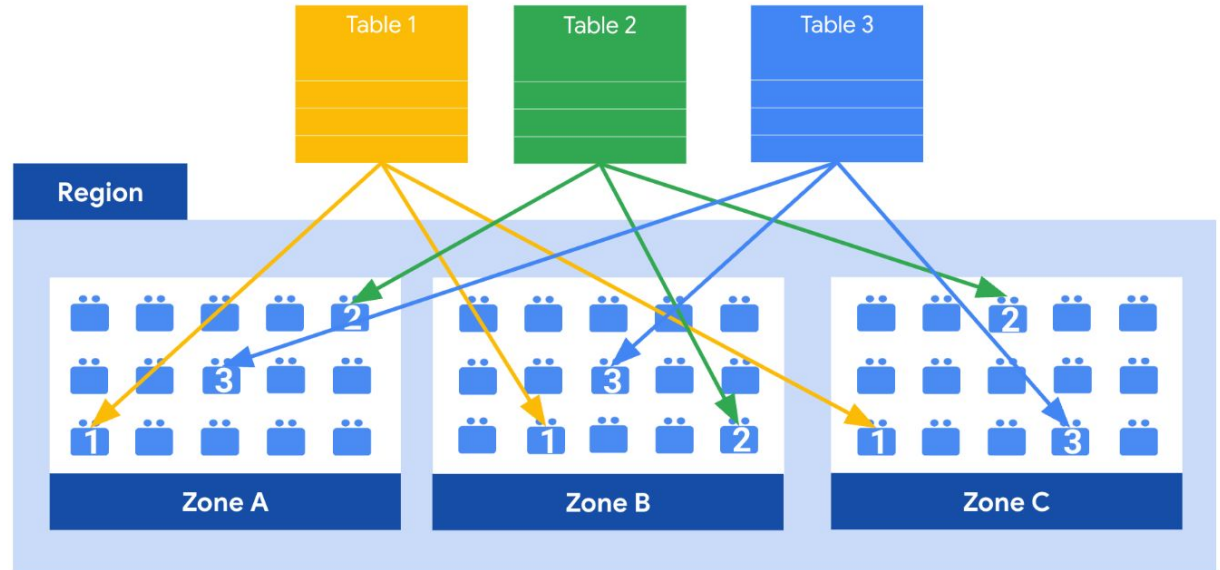
Each table is compressed and encrypted on disk

---

Storage is durable and each table is replicated across data centers

---

Supports streaming ingestion



# BigQuery | Remote Memory Shuffle

Faster performance for complex queries

---

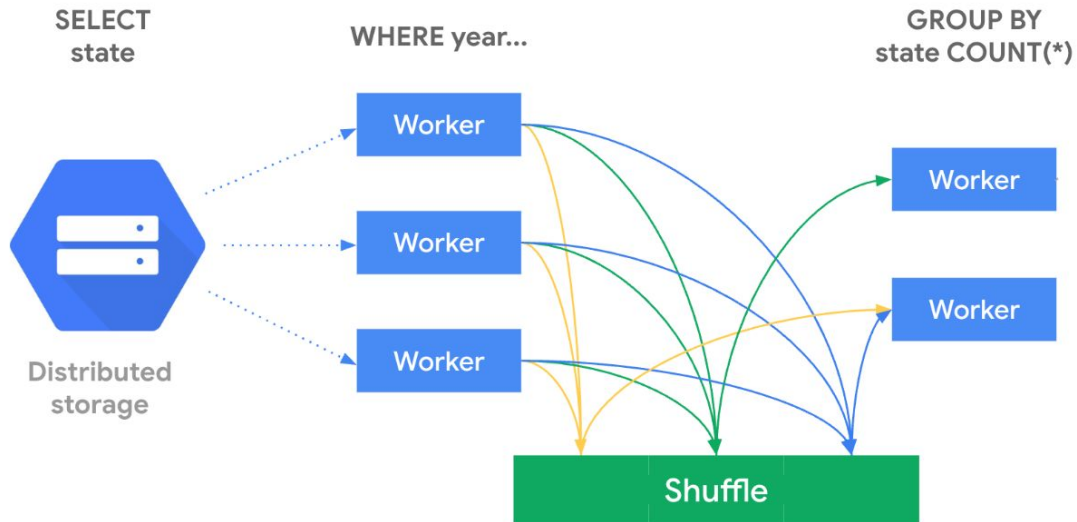
Join and aggregate more data

---

Better scalability

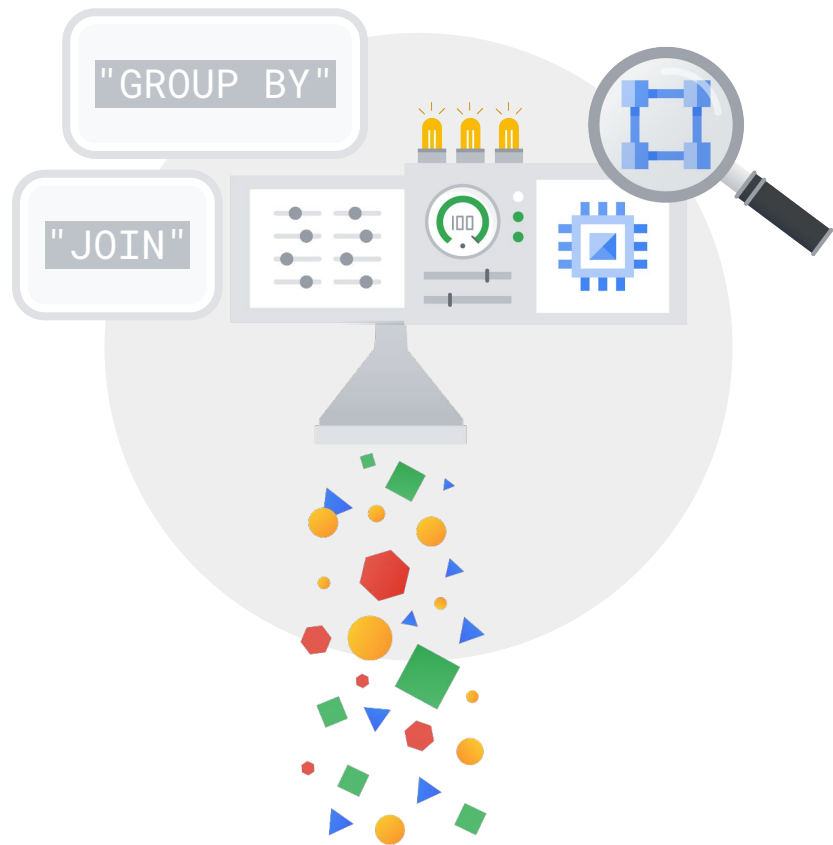
---

Large shuffles may spill to disk



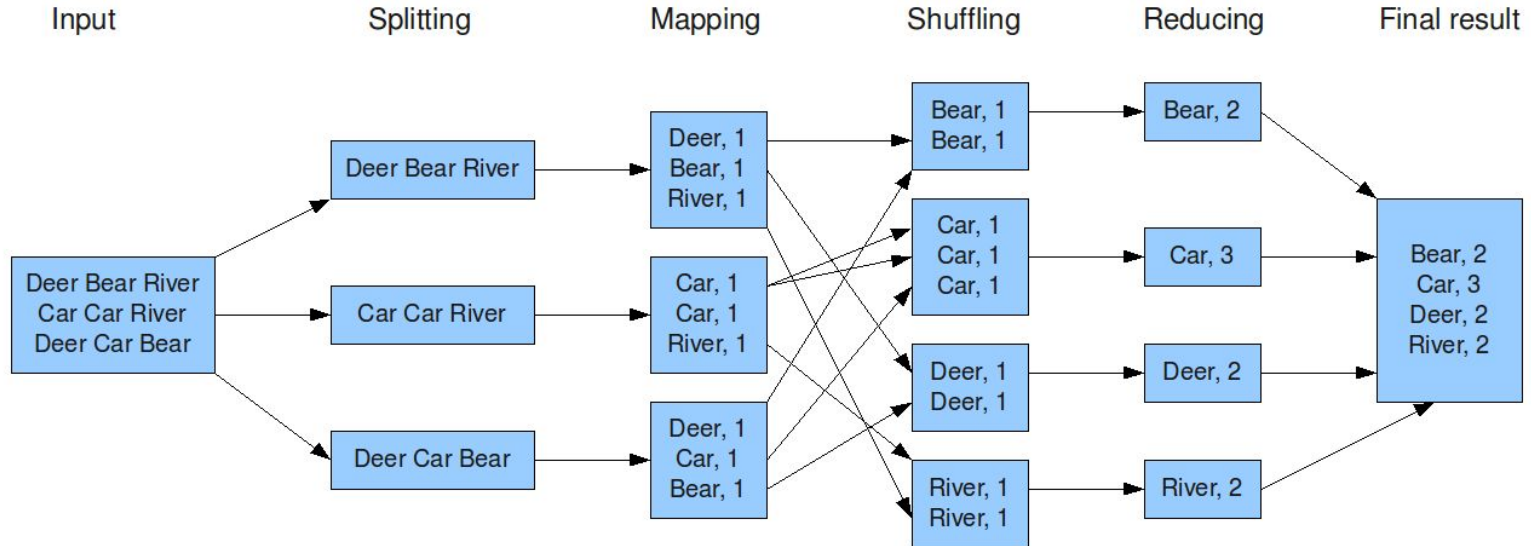
# What is shuffle?

The process of redistributing the intermediate data that the slots have processed.



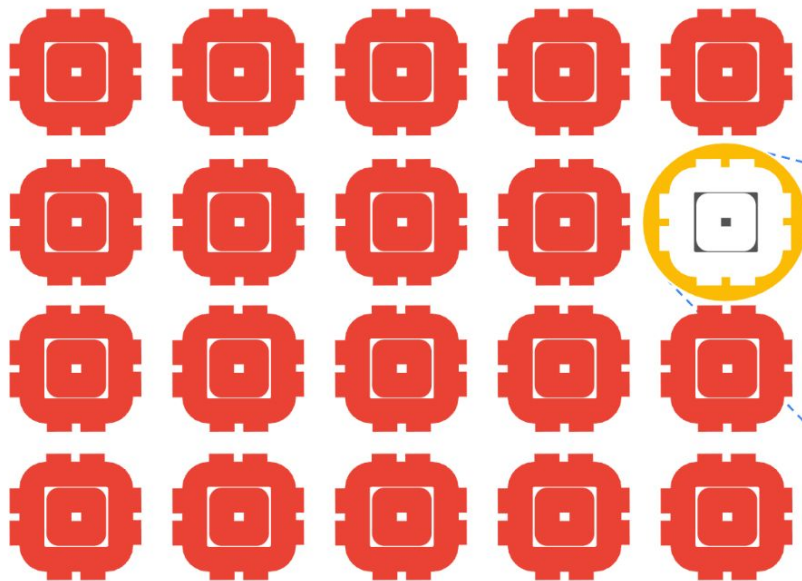
# Hadoop MapReduce

The overall MapReduce word count process



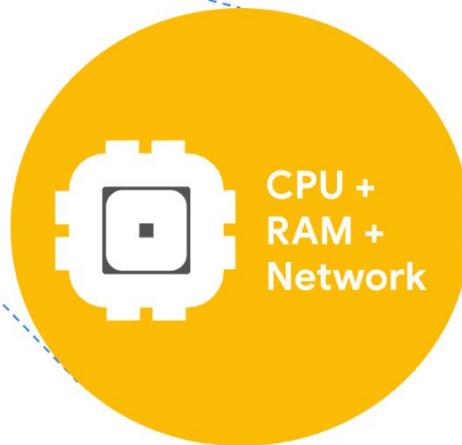
Reduce() will wait until map() is all finished. Pull mode approach.

# BigQuery | Analytics in BigQuery are measured in slots



BigQuery slots

Slots and shuffle ram are independent of each other!

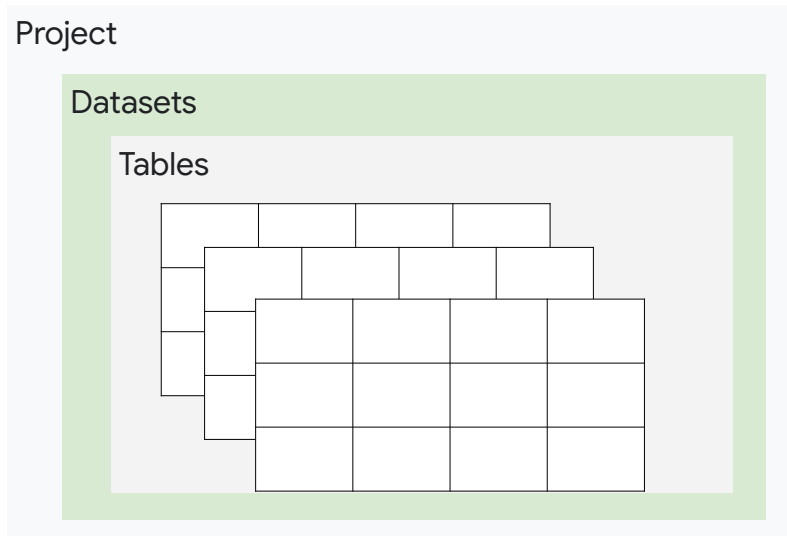


**BigQuery**

**Resource Hierarchy & Data Access Control**

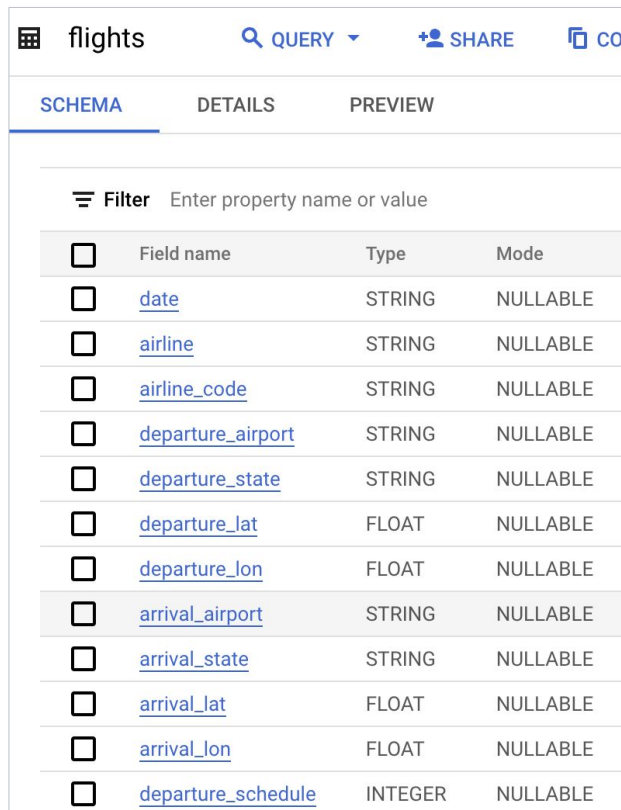
# BigQuery resource hierarchy

- In Google Cloud, you create projects
  - Projects have datasets
    - Datasets have tables (*and some other objects*)
- Users are granted permissions to datasets and/or tables
- The BigQuery service must be enabled within a project to use it
  - There is no need to provision a cluster or create a database



# In BigQuery, tables have a schemas

- The word “schema” is used differently in BigQuery than it is in Snowflake
  - In Snowflake, a schema is a collection of tables that users can query
- In BigQuery, a schema describes the fields and data types within a table

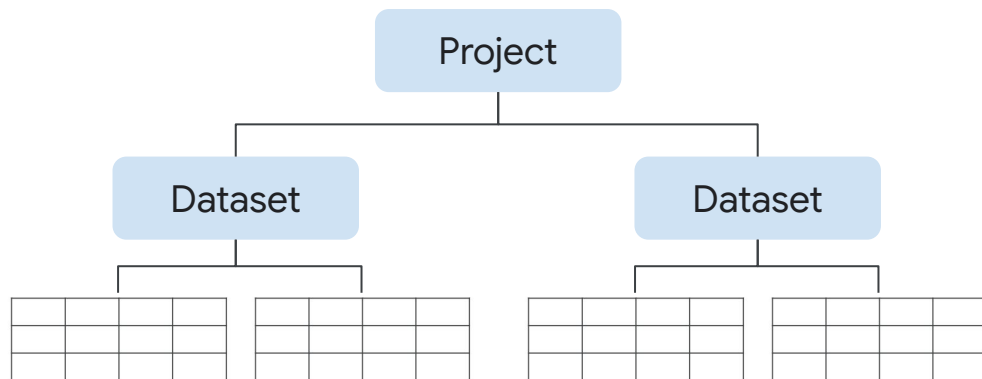


The screenshot shows the BigQuery interface for a table named 'flights'. At the top, there are navigation options: a grid icon, the table name 'flights', a search icon with 'QUERY', a share icon with 'SHARE', and a copy icon with 'CO'. Below this, there are three tabs: 'SCHEMA' (which is selected and underlined), 'DETAILS', and 'PREVIEW'. Under the 'SCHEMA' tab, there is a 'Filter' section with the text 'Filter Enter property name or value'. Below the filter is a table listing the fields of the 'flights' table. Each row has a checkbox on the left, followed by the field name, the data type, and the nullability mode.

<input type="checkbox"/>	Field name	Type	Mode
<input type="checkbox"/>	<a href="#">date</a>	STRING	NULLABLE
<input type="checkbox"/>	<a href="#">airline</a>	STRING	NULLABLE
<input type="checkbox"/>	<a href="#">airline_code</a>	STRING	NULLABLE
<input type="checkbox"/>	<a href="#">departure_airport</a>	STRING	NULLABLE
<input type="checkbox"/>	<a href="#">departure_state</a>	STRING	NULLABLE
<input type="checkbox"/>	<a href="#">departure_lat</a>	FLOAT	NULLABLE
<input type="checkbox"/>	<a href="#">departure_lon</a>	FLOAT	NULLABLE
<input type="checkbox"/>	<a href="#">arrival_airport</a>	STRING	NULLABLE
<input type="checkbox"/>	<a href="#">arrival_state</a>	STRING	NULLABLE
<input type="checkbox"/>	<a href="#">arrival_lat</a>	FLOAT	NULLABLE
<input type="checkbox"/>	<a href="#">arrival_lon</a>	FLOAT	NULLABLE
<input type="checkbox"/>	<a href="#">departure_schedule</a>	INTEGER	NULLABLE

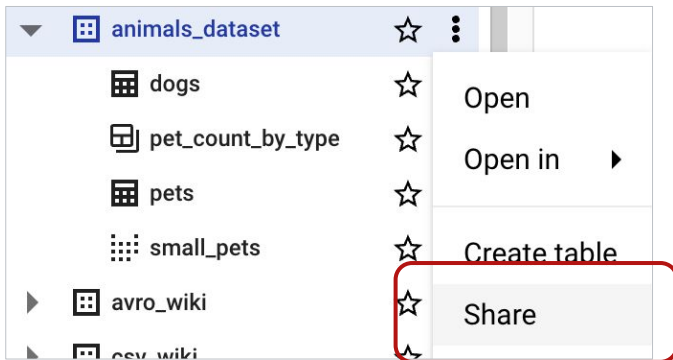
# Use Identity Access Management (IAM) to share datasets, tables, or views

- Can share data at any level of the project hierarchy
- Add one or more principals
  - Emails
  - Groups
  - Service accounts
- Assign a BigQuery data role
  - BigQuery Data Viewer
  - BigQuery Data Editor
  - BigQuery Data Owner



# Sharing datasets

- Click the action menu next to the dataset and select **Share**
- Add principal(s) and role(s)



### Add principals

Principals are users, groups, domains, or service accounts. [Learn more about principals in IAM](#)

New principals

rehnstrom@google.com ✕

### Assign roles

Roles are composed of sets of permissions and are used to control access to this resource. [Learn more](#)

Role \*

BigQuery Data Editor

Access to edit all the contents of datasets

# You can make data public if you like

- Add the principal `allAuthenticatedUsers`
- Assign the BigQuery Data Viewer role

## Add principals

Principals are users, groups, domains, or service accounts. [Learn more in IAM](#)

New principals

allAuthenticatedUsers ✕

## Assign roles

Roles are composed of sets of permissions and defined on a resource. [Learn more](#) with this resource.

Role \*

BigQuery Data Viewer ▼

Access to view datasets and all of their contents

# Row-level security allows access to specific rows in a based on qualifying user conditions

- Added using a DDL statement

```
CREATE ROW ACCESS POLICY
  dog_filter
ON
  animals_dataset.pets
GRANT TO ('user:noir@example.com')
FILTER USING (type = 'Dog');
```

```
CREATE ROW ACCESS POLICY
  cat_filter
ON
  animals_dataset.pets
GRANT TO ('group:cat-people@example.com')
FILTER USING (type = 'Cat');
```

# Column-level security can be done using authorized views

- Suppose you want to make some columns in a table public, but keep other columns private
  1. Create a second dataset and make it public
  2. Create a view in the public dataset that selects the columns you want to be public
  3. Add the view as an **Authorized view** in the private dataset

### Currently Authorized Views

Project ID	Dataset ID	Table ID
No rows to display		

### Authorize view

Authorized View \*

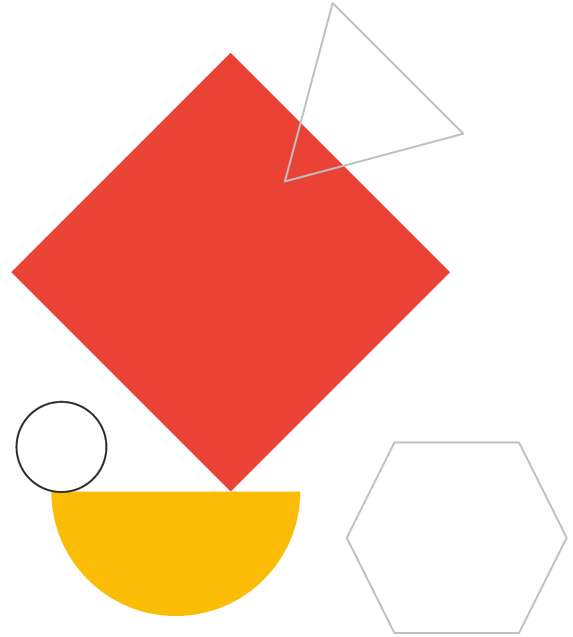
- bigquery-demo-proj.public\_animals.pets

[ADD AUTHORIZATION](#)

Demo

# BigQuery Studio

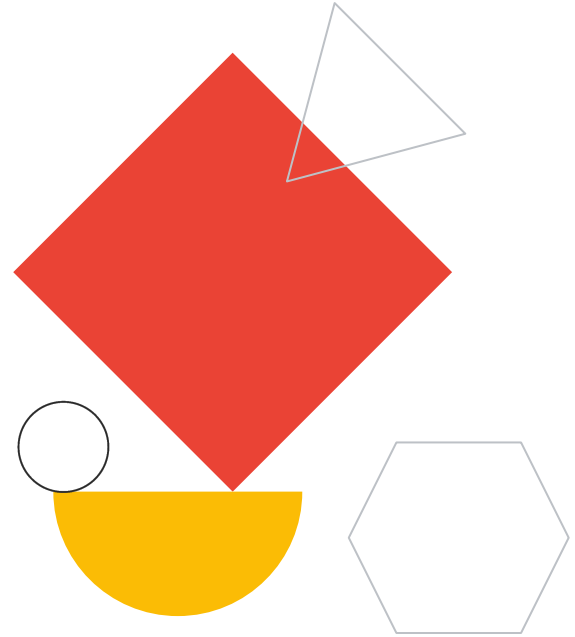
1. Querying External Data Sources or Public Datasets
2. Control access with Policy Tag
3. Building Pipelines in BigQuery - from low-code to coding method
4. Visualization & Business Intelligence with Looker & Looker Studio



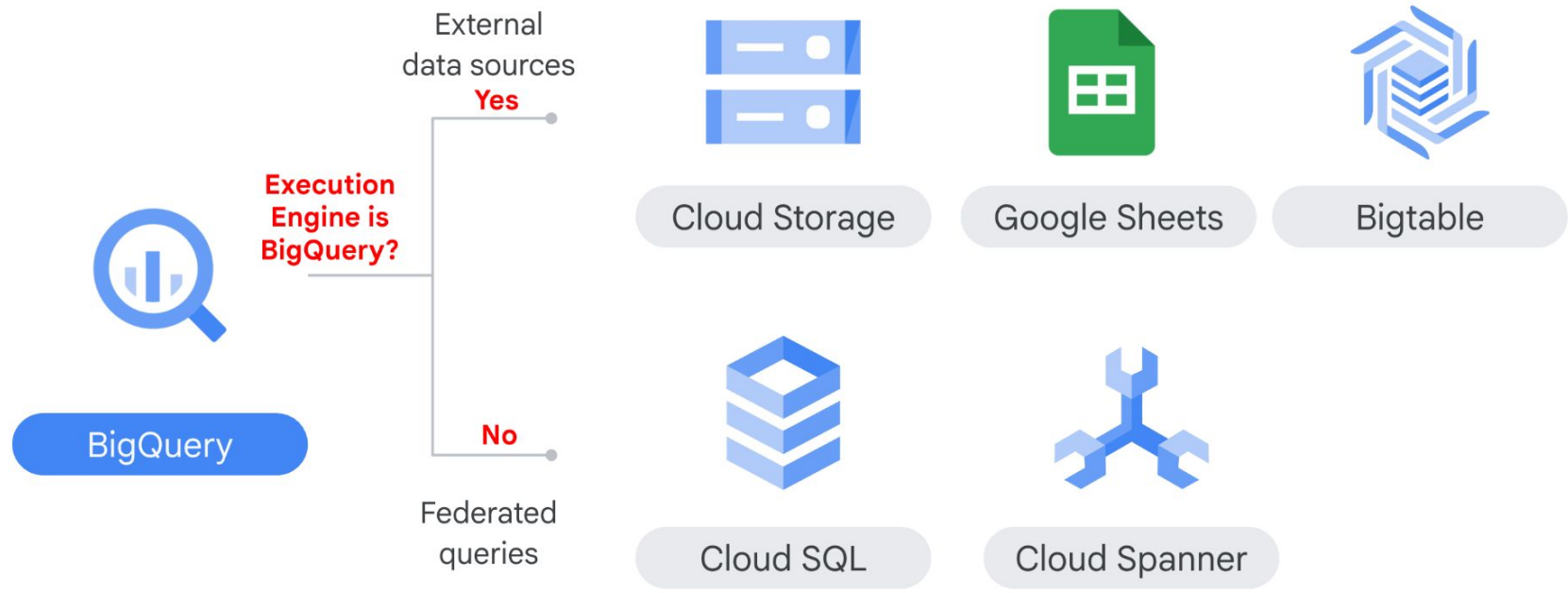
Demo

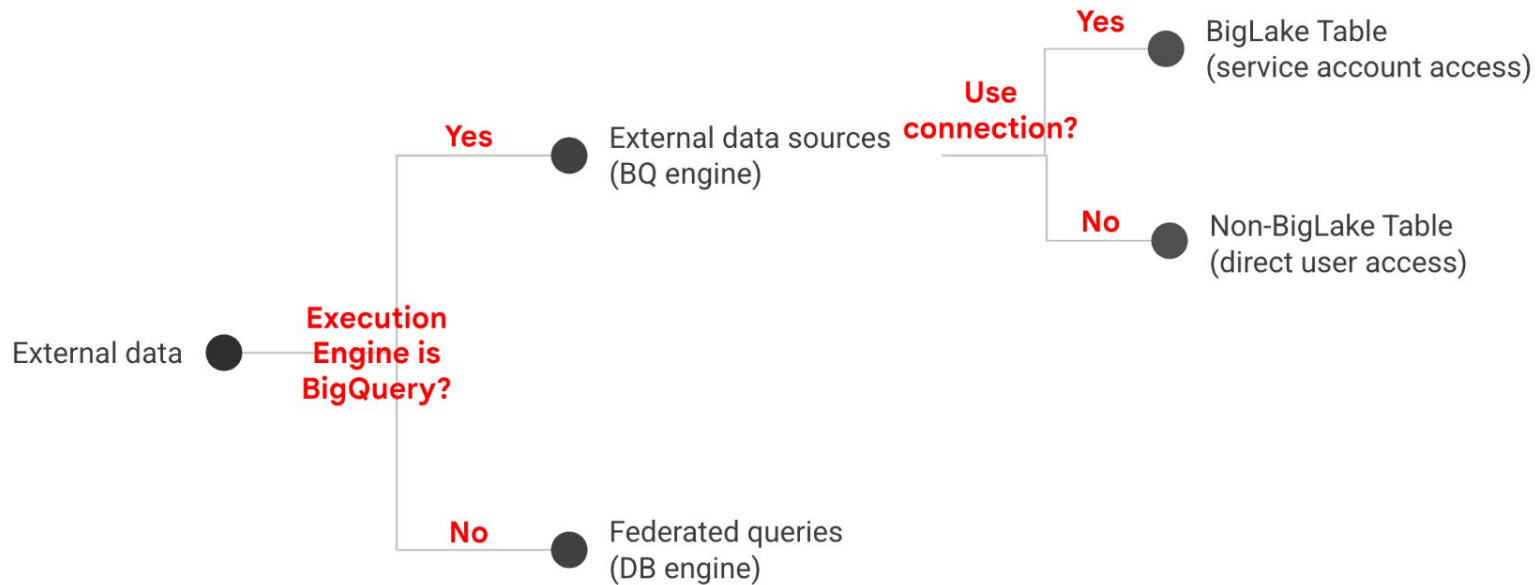
# BigQuery Studio

- 1. Querying External Data Sources or Public Datasets**
2. Control access with Row Policy, Authorized View, & Policy Tag
3. Building Pipelines in BigQuery - from low-code to coding method
4. Visualization & Business Intelligence with Looker & Looker Studio

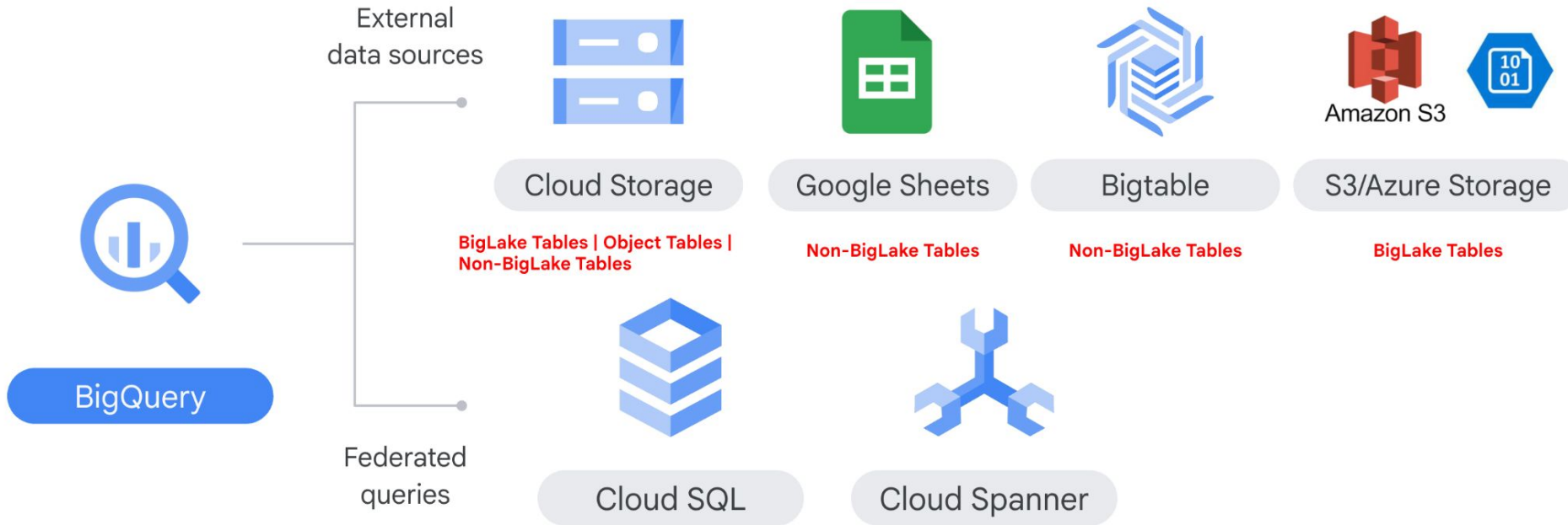


# Your data isn't stored in BigQuery? You can still query it!





# Your data isn't stored in BigQuery? You can still query it!



# Query data directly from Google Sheets in BigQuery

SCHEMA	DETAILS	LINEAGE	...
Table Info			
Table ID	your-project-id.sheets.sheets_table		
...			

## External Data Configuration

Source URI(s)	<a href="https://docs.google.com/spreadsheets/...">https://docs.google.com/spreadsheets/...</a>
Source format	GOOGLE_SHEETS



```
# Query the Google Sheets table
SELECT ...
FROM
`your-project-id.sheets.sheets_table`;
```

- No query cost estimation, table preview, or query caching available.
- Performance is slower than querying a permanent table in BigQuery.

# A BigLake table behaves the same way as a permanent table

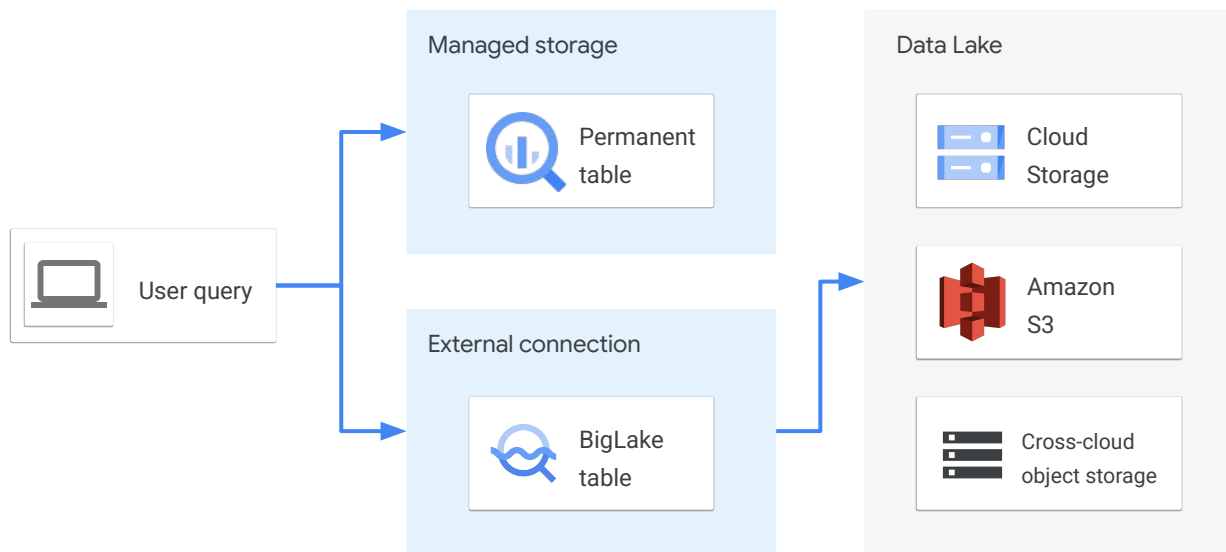
SCHEMA	DETAILS	LINEAGE	...
Table Info			
Table ID	your-project-id.biglake.csv_table		
...			
External Data Configuration			
Source URI(s)	<a href="gs://your-bucket/your_file.csv">gs://your-bucket/your_file.csv</a>		
...			

```
# Query the Biglake table
SELECT ...
FROM
`your-project-id.biglake.csv_table`;
```



- ✓ Use BigLake if loading data into BigQuery isn't an option for your use case.
- ✓ Queries are performant due to metadata caching.
- No query cost estimation or table preview.

# Query your data directly from your data lake using BigLake tables



## Supported formats:

- Avro
- CSV
- Delta Lake
- Iceberg
- JSON
- ORC
- Parquet

# Non-BigLake Table (Creation)

## Create table

### Source

Create table from  
Google Cloud Storage

Select file from GCS bucket or use a [URI pattern](#) \*  
 paggie-bucket/test.csv

File format  
CSV

Source Data Partitioning

### Destination

Project \*  
paggie-svc-project

Dataset \*  
paggie\_ml

Table \*  
demo\_gcs\_external\_table

Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.

Table type  
External table

**i** Regional / dual region GCS buckets are recommended for External table.

Create a BigLake table using a Cloud resource connection

### Schema

Auto detect

# BigLake Table (Creation)

## Create table

### Source

Create table from  
Google Cloud Storage

Select file from GCS bucket or use a [URI pattern](#) \*  
 paggie-bucket/test.csv

File format  
CSV

Source Data Partitioning

### Destination

Project \*  
paggie-svc-project

Dataset \*  
paggie\_ml

Table \*  
demo\_gcs\_biglake\_table

Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.

Table type  
External table

**i** Regional / dual region GCS buckets are recommended for External table.

Create a BigLake table using a Cloud resource connection

Connection ID \*  
asia-east1.demo\_biglake\_connection

# BigLake Table (Creation)

Create a BigLake table using a Cloud resource

Connection ID \*  
asia-east1.demo\_biglake\_connection

## Schema

Create table

Cancel


You are missing permissions and may need to talk to your administrator. Original error message: Failed to create table:  
**Access Denied: BigQuery BigQuery: Permission denied while globbing file pattern. bqcx-304589458397-ba4e@gcp-sa-bigquery-condel.iam.gserviceaccount.com does not have storage.objects.get access to the Google Cloud Storage object. Permission 'storage.objects.get' denied on resource (or it may not exist). Please make sure gs://paggie-bucket/test.csv is accessible via appropriate IAM roles, e.g. Storage Object Viewer or Storage Object Creator.**

Grant BQ Service Account  
access in GCS

Grant access to "paggie-bucket"

Grant principals access to this resource and add roles to specify what actions the principals can take. Optionally, add conditions to grant access to principals only when a specific criteria is met. [Learn more about IAM conditions](#)

Resource

 paggie-bucket

Add principals

Principals are users, groups, domains, or service accounts. [Learn more about principals in IAM](#)

New principals \*

bqcx-304589458397-ba4e@gcp-sa-bigquery-  
condel.iam.gserviceaccount.com

Assign roles

Roles are composed of sets of permissions and determine what the principal can do with this resource. [Learn more](#)

Role \*  
Storage Object Viewer

IAM condition (optional) ⓘ  
IAM conditions disabled

Grants access to view objects and their metadata, excluding ACLs. Can also list the objects in a bucket.

+ Add another role

Save

Cancel

# Non-BigLake Table

paggie-svc-project / Datasets / paggie\_ml / Tables / demo\_gcs\_external\_table

demo\_gcs\_external\_table [Query](#) [Open in](#) [Share](#) [Delete](#)

[Schema](#) [Details](#) [Insights](#) [Lineage](#) [Data Profile](#) [Data Quality](#)

Generate table and column descriptions with Gemini. Ensure a profile scan to ground the

Filter Enter property name or value

<input type="checkbox"/>	Field name	Type	Mode	Description	Key	Collation	Default
<input type="checkbox"/>	first_name	STRING	NULLABLE	-	-	-	-
<input type="checkbox"/>	last_name	STRING	NULLABLE	-	-	-	-
<input type="checkbox"/>	gender	STRING	NULLABLE	-	-	-	-
<input type="checkbox"/>	age	INTEGER	NULLABLE	-	-	-	-

[Edit schema](#)

# BigLake Table

paggie-svc-project / Datasets / paggie\_ml / Tables / demo\_gcs\_biglake\_table

demo\_gcs\_biglake\_table BigLake [Query](#) [Open in](#) [+](#)

[Schema](#) [Details](#) [Insights](#) [Lineage](#) [Data Profile](#) [Data Quality](#)

Generate table and column descriptions with Gemini. Ensure a profile scan to ground the

Filter Enter property name or value

<input type="checkbox"/>	Field name	Type	Mode	Description	Key
<input type="checkbox"/>	first_name	STRING	NULLABLE	-	-
<input type="checkbox"/>	last_name	STRING	NULLABLE	-	-
<input type="checkbox"/>	gender	STRING	NULLABLE	-	-
<input type="checkbox"/>	age	INTEGER	NULLABLE	-	-

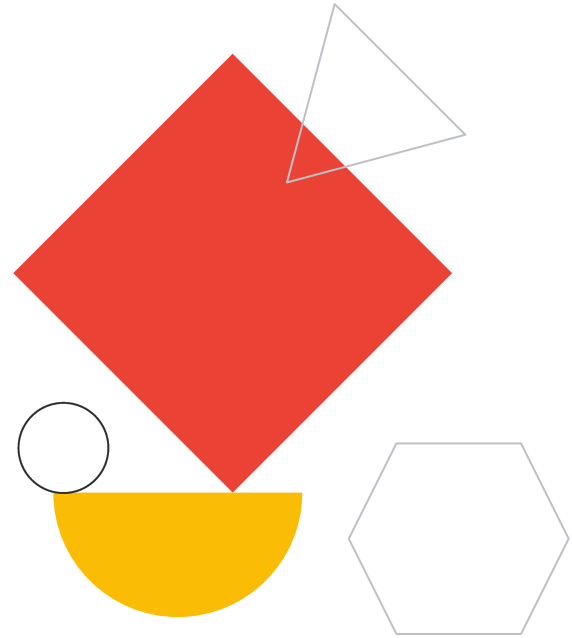
[Edit schema](#) [View row access policies](#)

**Task: Find the difference**

Demo

# BigQuery Studio

1. Querying External Data Sources or Public Datasets
2. **Control access with Row Policy, Authorized View, & Policy Tag**
3. Building Pipelines in BigQuery - from low-code to coding method
4. Visualization & Business Intelligence with Looker & Looker Studio



Google Cloud paggie-svc-project Search (/) for resources, docs

BigQuery / Policy tags / Policy tag taxonomy: 5465440580971000379

Overview **Preview**

Studio

Agents **Preview**

Pipelines & Integration ^

Data transfers

Dataform

Scheduled queries

Scheduling

Governance ^

Sharing (Analytics Hub)

**Policy tags**

Metadata curation

Administration ^

Monitoring

Jobs explorer

Capacity management

BI Engine

Disaster recovery

Recommendations

Partner Center

Settings **Preview**

Release Notes

<|

Policy tag taxonomy **pii** Edit Delete

Policy tags 1

Enforce access control  
Access to BigQuery columns tagged with the policy tags below will be controlled by the Fine-Grained Reader and the Masked Reader roles.

Metadata

Policy tag taxonomy ID	5465440580971000379
Created	Mar 31, 2026, 12:03:52 AM
Modified	Mar 31, 2026, 12:03:52 AM
Project ID	paggie-svc-project
Project display name	paggie-svc-project
Location	asia-east1 (Taiwan)

Policy tags

Policy tags are tags with access control policies that can be applied to BigQuery resources, for example, BigQuery columns.

[Manage Data Policies](#)

<input checked="" type="checkbox"/>	Name ↑	ID
<input checked="" type="checkbox"/>	confidential	2755975979137140608

## Masking Rules

Choose a masking rule, then select principals that the rule would apply to.

Data Policy Name 1  
sha256

**Masking Rule 1** \* **Principals 1**

**Filter** Type to filter

Presets

- Date Year Mask
- Default Masking Value
- Email Mask
- First Four Characters
- Hash (SHA256)**
- Last Four Characters
- Nullify

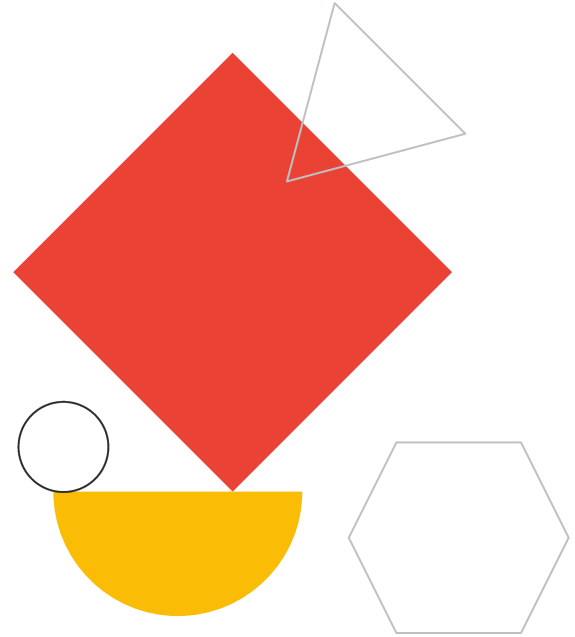
[+ Add Rule](#)

[Submit](#) [Cancel](#)

Demo

# BigQuery Studio

1. Querying External Data Sources or Public Datasets
2. Control access with Row Policy, Authorized View, & Policy Tag
3. **Building Pipelines in BigQuery - from low-code to code method**
4. Visualization & Business Intelligence with Looker & Looker Studio



### Data Canvas Overview

Welcome to Data Canvas!

BigQuery data canvas simplifies complex data analysis through its visual representation using a graph of interconnected nodes. Nodes can be chained together in any sequence, creating a unified workspace that enables iterative querying and multistaged exploration. Gemini is also embedded into the nodes to assist you in generating SQL queries, visualizations, and valuable insights. In this sample canvas, we'll demonstrate this by querying and visualizing retail data to discover what products are expected to have the lowest inventory.

You can get started with a canvas by adding the data you want to explore through a table search or by right-clicking a resource in the resource explorer. For this example, we've added two data nodes to the canvas: order\_items and users. Each node represents a data table, and you can explore its schema, details or generate its data within the nodes itself. By joining these nodes, we've connected them, enabling us to query both simultaneously.

1/5 [Next](#)

#### order\_items

Field name	Type	Mode	Description	Key	Collation	Default Value	Pr
id	INTEGER	NULLABLE					
order_id	INTEGER	NULLABLE					
user_id	INTEGER	NULLABLE					
product_id	INTEGER	NULLABLE					
inventory_item_id	INTEGER	NULLABLE					
status	STRING	NULLABLE					
created_at	TIMESTAMP	NULLABLE					
shipped_at	TIMESTAMP	NULLABLE					
delivered_at	TIMESTAMP	NULLABLE					
returned_at	TIMESTAMP	NULLABLE					
unit_price	FLOAT	NULLABLE					

#### users

Field name	Type	Mode	Description	Key	Collation	Default Value	Pr
id	INTEGER	NULLABLE					
first_name	STRING	NULLABLE					
last_name	STRING	NULLABLE					
email	STRING	NULLABLE					
age	INTEGER	NULLABLE					
gender	STRING	NULLABLE					
state_address	STRING	NULLABLE					
street_address	STRING	NULLABLE					
postal_code	STRING	NULLABLE					
city	STRING	NULLABLE					
country	STRING	NULLABLE					
latitude	FLOAT	NULLABLE					

### SQL node onboarding

#### SQL nodes and querying basics

In data canvas SQL nodes, you can either use Gemini and natural language to generate a query, view prompting best practices, or write queries directly in the editor. For instance, we generated the query shown here by asking, "What are the most commonly sold products in the United States by product ID?"

SQL nodes can also be joined with data nodes, as demonstrated here. We've added the data table products and joined it with our results from the SQL node so that we can get the product names.

2/5 [Back](#) [Next](#)

#### SQL

Describe how to refine the SQL

```

1 SELECT
2   _bigquery-public-data.the_ohio_e-commerce.order_items.product_id,
3   FROM
4   _bigquery-public-data.the_ohio_e-commerce.order_items
5 JOIN
6   _bigquery-public-data.the_ohio_e-commerce.users
7 ON
8   _bigquery-public-data.the_ohio_e-commerce.order_items.user_id = _bigquery-public-data.the_ohio_e-commerce.users.id
9 WHERE
10  _bigquery-public-data.the_ohio_e-commerce.country = 'United States'
11 GROUP BY
12  product_id
13 ORDER BY
14  _bigquery-public-data.the_ohio_e-commerce.order_items.product_id
15 
```

Query results

Job information	Results	JSON	Execution details	Execution graph
Row	product_id	total_sales		
1	11871	8		
2	34482	8		
3	12018	8		
4	18782	8		
5	21287	8		
6	22891	8		

Results per page: 50 | 1 - 10 of 10

#### products

Field name	Type	Mode	Description	Key	Collation	Default Value	Pr
id	INTEGER	NULLABLE					
cost	FLOAT	NULLABLE					
category	STRING	NULLABLE					
name	STRING	NULLABLE					
brand	STRING	NULLABLE					
retail_price	FLOAT	NULLABLE					
department	STRING	NULLABLE					
sku	STRING	NULLABLE					
distribution_center_id	INTEGER	NULLABLE					



BigQuery

paggie

Code

Compiled graph

Executions

Start execution

Ask Agent

Compiled

Learn

Overview Preview

Studio

Agents Preview

Pipelines & Integration

Data transfers

Dataform

Scheduled queries

Scheduling

Governance

Sharing (Analytics Hub)

Policy tags

Files

Commit 4 changes

Type to search

\*definitions

\*first\_view.sqlx

\*second\_view.sqlx

includes

\*.gitignore

\*workflow\_settings.yaml

definitions/first\_view.sqlx

Run

Format

```

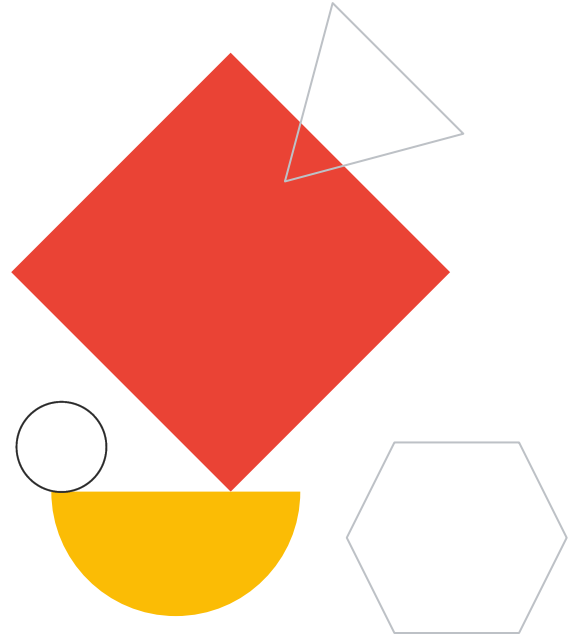
1 config {
2   type: "view",
3   columns: {
4     test: "A description for the test column",
5   }
6 }
7
8 -- The rest of a SQLX file contains your SELECT statement used to create the table.
9
10 SELECT 1 as test
11

```

Demo

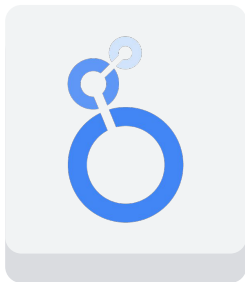
# BigQuery Studio

1. Querying External Data Sources or Public Datasets
2. Control access with Row Policy, Authorized View, & Policy Tag
3. Building Pipelines in BigQuery - from low-code to code method
4. **Visualization & Business Intelligence with Looker & Looker Studio**



# Data analytics lifecycle: Analyze

Looker &  
Looker Studio



✓ Unified Business Intelligence platform

✓ The Looker family (Looker, Looker Studio)

✓ Supports modeling (LookML)

✓ Easy integration and connection

Target Segment Description    Product Matches

is 90s punk rock

is 500



Product Segment

I've identified the top 500 matched products for '90s punk rock'. Let's see the top trending brands globally.

Bullhead DC    Paul Frank  
**Metal Mulisha** Bottoms Out    Wilson  
**Hot Topic**    Wallflower  
True Religion **Rock Revival**



Explore 184 Matched Brands

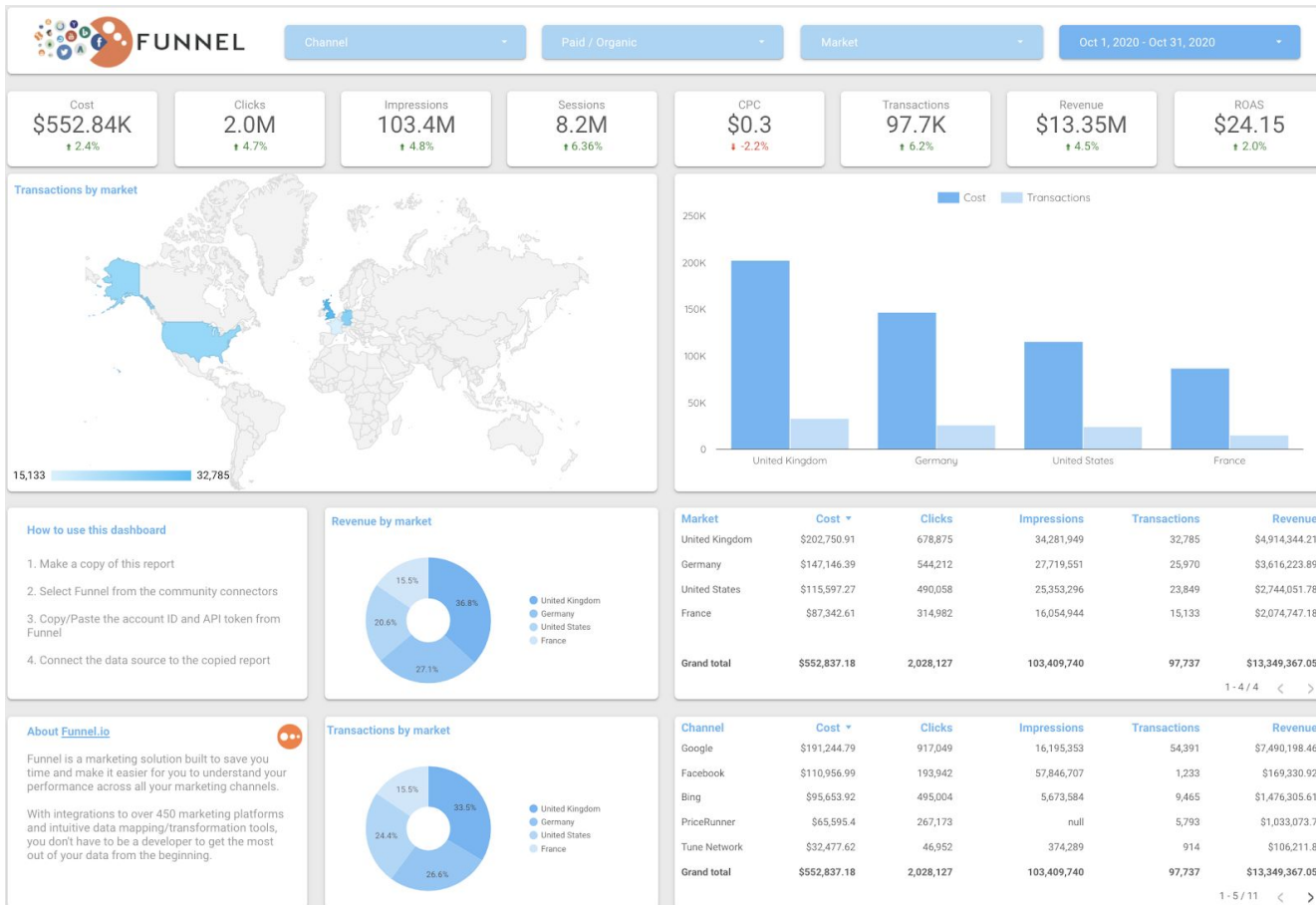
Explore 471 Matched Products

Size of Opportunity

**\$672.2K**

6% of \$10.8M last year's revenue

This product segment is forecasted to earn \$446.3K in the next 6 months.



# Looker Studio (editor mode)

SE Ranking Dashboard for SEO managers and decision-makers [Looker Studio] [...]

File Edit View Insert Page Arrange Resource Help

Reset

Share

View



Page 3 of 12 Add data Add a chart Add a control Theme and layout Pause updates

+ Add quick filter

Reset

Theme and Layout

Data

THEME

LAYOUT

Current Theme  
Custom

Customize

Custom

Text

Custom

Default

Text

Extract theme from image

Search

DEMO GA4 Account - Ecom

SE Ranking Rank Tracker

Competitor

Date

TOP 10 percent

Add a field

Add a parameter

SE Ranking Rank Tracker - Added c...

Domains Total

Backlinks Total

domains new

[demo] Search Visibility Chart (Goo...

Google Algorithm Updates Sheet

SE Ranking Rank CWV

Add Data

Data

Properties

Filter bar

SE Ranking

Data Source on page: GA4

Oct 1, 2024 - Oct 28, 2024

Default Data

New/returning

Country

default data period is last 28 days

Default channel group

Sessions  
233,698

↑ 42.5% from previous 28 days

Engaged sessions  
219,816

↑ 43.1% from previous 28 days

P. Revenue  
\$0

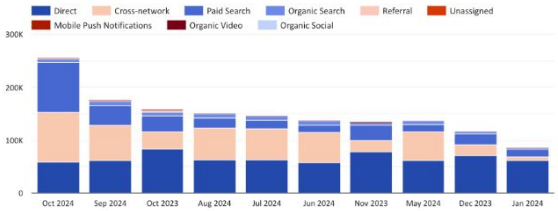
N/A from previous 28 days

Bounce rate  
6%

↓ -6.8% from previous 28 days

Which traffic channel groups are generating the most sessions to the website?

Default channel groups for your visitors over the past 12 months | Default channel group by session for past 12 months



Default channel group

Default channel group	Sessions
Direct	817,187
Cross-network	490,377
Paid Search	312,912
Organic Search	72,527
Referral	9,382
Unassigned	5,245
Mobile Push Notifications	539
Organic Video	130
Organic Social	37

How is your website traffic concentrated - geographically and per section?

Most popular country location of your visitors | Country origin by Sessions, purchase revenue

Country	Sessions	Purchase revenue
Bangladesh	62,688	\$0
Pakistan	33,860	\$0
India	29,924	\$0
United States	24,746	\$0
Philippines	15,048	\$0

Most popular page category | Page category by Sessions, purchase revenue

Event name	Sessions	Purchase revenue
session_start	224,558	\$0
screen_view	217,489	\$0
user_engagement	205,807	\$0
select_content	124,855	\$0
(not set)	92,192	\$0

# LookML - Define the dimensions & measures

e\_thelook    dev-erin-docs-dpzt    personal branch    ✓ No Errors    ⌵    i    Up to Date

**File Browser**    ⌵    🗂    🔍    +

- explores
- models
- views
  - distribution\_centers.view
  - events.view
  - inventory\_items.view
  - order\_facts.view
  - order\_history.view
  - order\_items.view**
  - products.view
  - user\_with\_age\_extension.view
  - users.view
  - users\_extended.view

**order\_items.view**

```
83
84 → dimension: status {
85     type: string
86     sql: ${TABLE}.status ;;
87 }
88
89 → dimension: user_id {
90     type: number
91     # hidden: yes
92     sql: ${TABLE}.user_id ;;
93 }
94
95 → measure: count {
96     type: count
97     drill_fields: [detail*]
98 }
99
100 → measure: wholesale_value {
101     type: number
102     sql: (${TABLE}.sale_price * 0.60) ;;
103 }
104
```

**dimensions**

**measures**

# Agenda

01 The Data Analytics Lifecycle Overview

---

02 Ingest

---

03 Process

---

04 Store

---

## Analyze

Introduction to BigQuery

Querying External Data Sources or Public Datasets **Demo**

Control access with Policy Tag **Demo**

Building Pipelines in BigQuery - from low-code to coding method **Demo**

Visualization & Business Intelligence with Looker & Looker Studio **Demo**

Derive insights from your data in BigQuery (Agent, Gemini Enterprise) **Demo**

---

06

## Activate

Introduction AI on Google Cloud

BigQuery ML **Demo**

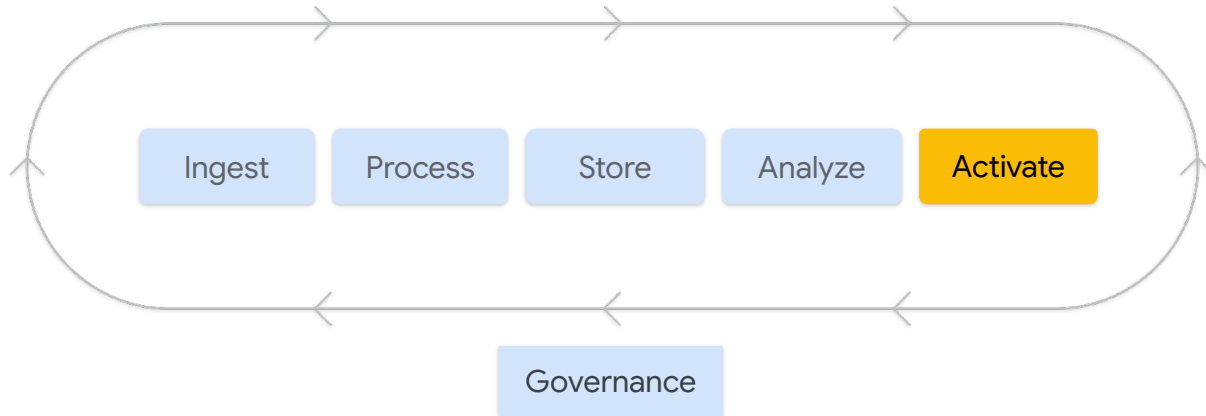
---

07

Governance

---

# Data analytics lifecycle



**Objective:** Leverage AI to make data-driven decision

**This layer focuses on:**

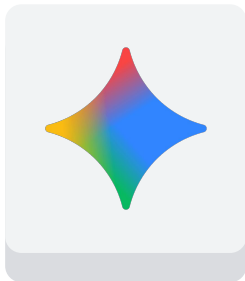
- ML and prediction serving
- Automation and decision triggering
- Feedback loops for continuous improvement

**Related Tools:**

- Vertex AI custom model or AutoML
- BigQuery ML

# Data analytics lifecycle: Analyze

Agent



Conversational Analytics



BigQuery Agent



Gemini App / Gemini Enterprise



Google Workspace with Gemini  
(Google Sheets)

# BigQuery Agent

Google Cloud paggie-svc-project Search (/) for resources, docs, products, and more Search

Conversations Agent Catalog

What questions can I ask about this data?

Here's the query result for Top Brands Men Revenue.

	brand	total_revenue
8	voicom	80,838
9	Columbia	74,895
10	The North Face	72,037

Top 10 Brands by Total Revenue (Men's Department)

Brand	Total Revenue
Diesel	175,000
Carhartt	150,000
Calvin Klein	135,000
True Religion	110,000
7 For All Mankind	105,000
Tommy Hilfinger	95,000
Quiksilver	90,000
Voicom	80,838
Columbia	74,895
The North Face	72,037

Ask a question Thinking

[View best practices](#) for better answers. The responses from Conversational Analytics in BigQuery may not be complete or accurate. [Learn more](#)

Details

Agent

The Look Ecommerce

Description

You are a sample agent that has insights to the looks ecommerce dataset. Advanced analytics is not enabled for you.

[+ New conversation](#) [View config](#)

Knowledge sources

- events
- order\_items
- products
- users

Labels

-

# Gemini Enterprise



Barbie Plus



summarize the marketing data using the marketing datastore

Show thinking ▾

Hello Paggie,

Of course, I can help you with that. Here is a summary of the marketing data from the marketing datastore.

## Social Media Campaigns 📊

Platform	Campaign	Conversions	Spend
Instagram	Reels Ad - Cookie Baking	110	\$112.90
Instagram	Stories Flash Sale - Sugar Cookies	70	\$62.15
Facebook	Spring Cookie Promotion	65	\$385.50
Facebook	Website Visitor Retargeting	45	\$180.90
Twitter	Oatmeal Raisin Cookies Ad	25	\$88.40

📄 Export to Sheets



## Email Marketing 📧

Campaign	Conversions	Spend
March Newsletter - New Brownies	90	\$45.20

Ask a follow-up

+ 🗨️ 2 of 9 Auto ▾ ▶

Generative AI may display inaccurate information, including about people, so double-check its responses.

# Google Workspace with Gemini

[Demo] Gemini in Google Sheet

File Edit View Insert Format Data Tools Gemini Extensions Help

100% \$ % .0 .00 123 Defaul... - 10 +

	A	B	C	D	E	F	G	H
1	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	me
2	-122.23	37.88	41	880	129	322	126	
3	-122.22	37.86	21	7099	1106	2401	1138	
4	-122.24	37.85	52	1467	190	496	177	
5	-122.25	37.85	52	1274	235	558	219	
6	-122.25	37.85	52	1627	280	565	259	
7	-122.25	37.85	52	919	213	413	193	
8	-122.25	37.84	52	2535	489	1094	514	
9	-122.25	37.84	52	3104	687	1157	647	
10	-122.26	37.84	42	2555	665	1206	595	
11	-122.25	37.84	52	3549	707	1551	714	
12	-122.26	37.85	52	2202	434	910	402	
13	-122.26	37.85	52	3503	752	1504	734	
14	-122.26	37.85	52	2491	474	1098	468	
15	-122.26	37.84	52	696	191	345	174	
16	-122.26	37.85	52	2643	626	1212	620	
17	-122.26	37.85	50	1120	283	697	264	
18	-122.27	37.85	52	1966	347	793	331	
19	-122.27	37.85	52	1228	293	648	303	
20	-122.26	37.84	50	2239	455	990	419	
21	-122.27	37.84	52	1503	298	690	275	
22	-122.27	37.85	40	751	184	409	166	
23	-122.27	37.85	42	1639	367	929	366	
24	-122.27	37.84	52	2436	541	1015	478	
25	-122.27	37.84	52	1688	337	853	325	
26	-122.27	37.84	52	2224	437	1006	422	
27	-122.28	37.85	41	535	123	317	119	
28	-122.28	37.85	49	1130	244	607	239	
29	-122.28	37.85	52	1898	421	1102	397	
30	-122.28	37.84	50	2082	492	1131	473	
31	-122.28	37.84	52	729	160	395	155	
32	-122.28	37.84	49	1916	447	863	378	
33	-122.28	37.84	52	2153	481	1168	441	

Generate insights or trends for this data

Analysis steps

Median House Value vs. Median Income

Median House Value (Y-axis, 0 to 550,000)

Median Income (in \$10,000s) (X-axis, 0 to 16)

Ask Gemini

Gemini in Workspace can make mistakes. [Learn more](#)

Create a column chart of the median house value grouped by ocean proximity to compare location-based pricing.

Analysis steps

Median House Value by Ocean Proximity

Median House Value (\$) (Y-axis, 0 to 500,000)

Ocean Proximity (X-axis: ISLAND, NEAR BAY, NEAR OCEAN, <1H OCEAN, INLAND)

# Options for ML models on Google Cloud



Cloud TPUs



Compute Engine



Dataproc



Google Kubernetes Engine

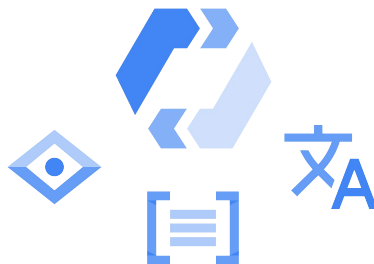


Vertex AI



BigQuery ML

## AutoML



Cloud Translation API



Vision API



Speech-to-Text API



Video Intelligence API



Data Loss Prevention API



Text-to-Speech API



Cloud Natural Language API



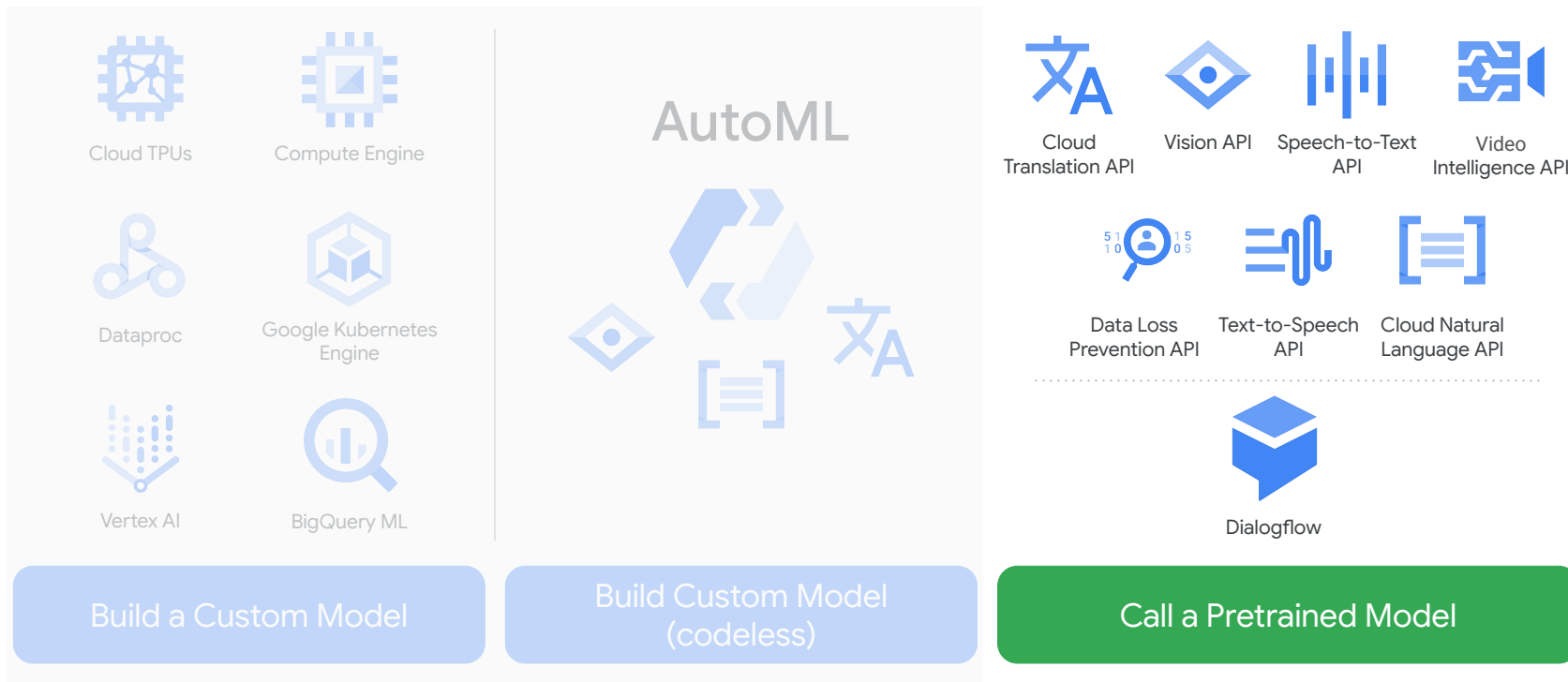
Dialogflow

Build a Custom Model

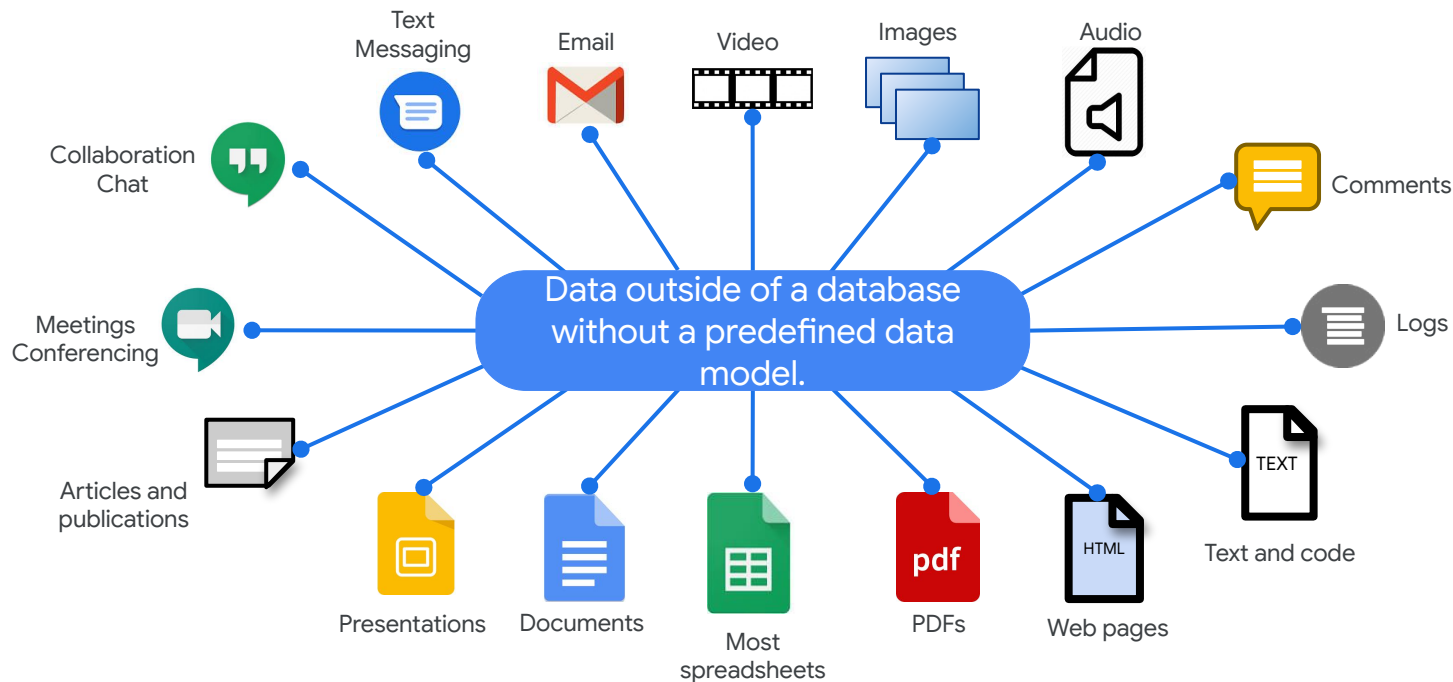
Build Custom Model  
(codeless)

Call a Pretrained Model

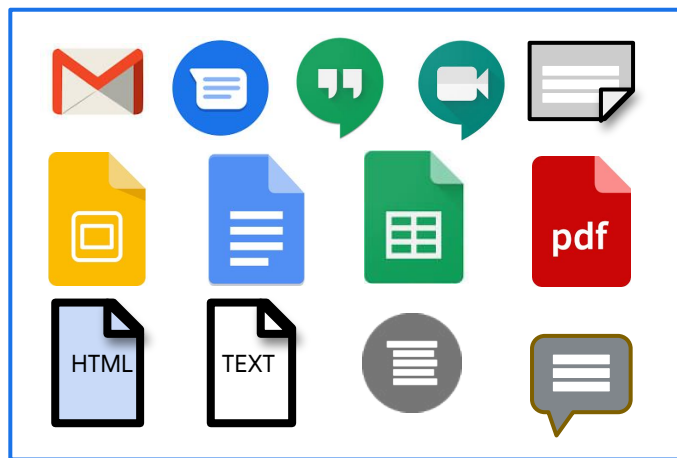
# For common ML tasks, consider pretrained APIs



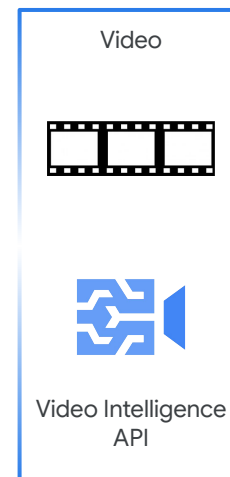
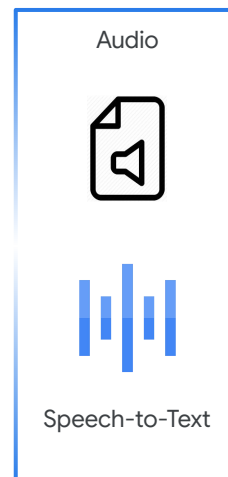
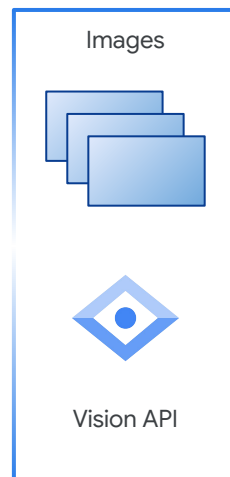
# Most business data is unstructured data, and mainly text



# You process unstructured data by labeling it with AI



Cloud Natural  
Language API

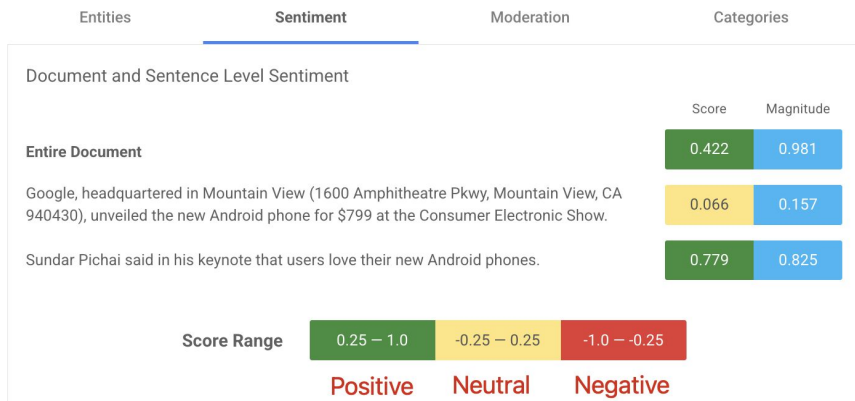


# Demo: Cloud Natural Language API

## Use cases:

- Entity Analysis
- Sentiment Analysis
- Moderation
- Content Classification

## Basic Example: [Natural Language AI](#)



Google, headquartered in Mountain View unveiled the new Android phone for \$799 at the Consumer Electronic Show. Sundar Pichai said in his keynote that users love their new Android phones. ↻ RESET

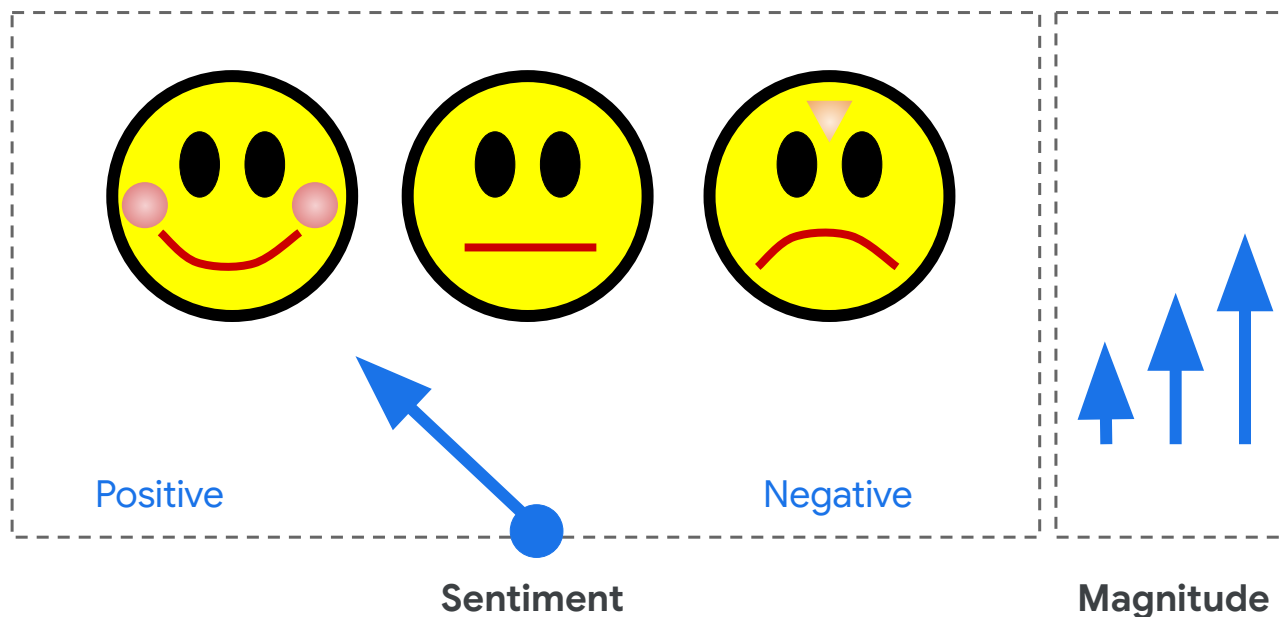
[See supported languages](#)

Entities Sentiment Syntax Categories

<Google><sub>1</sub> , headquartered in <Mountain View><sub>6</sub> unveiled the new <Android><sub>4</sub> <phone><sub>3</sub> for <\$799><sub>10</sub> <799><sub>11</sub> at the <Consumer Electronic Show><sub>7</sub> . <Sundar Pichai><sub>5</sub> said in his <keynote><sub>9</sub> that <users><sub>2</sub> love their new <Android><sub>4</sub> <phones><sub>8</sub> .

1. Google <a href="#">Wikipedia Article</a> Salience: 0.26	ORGANIZATION	2. users Salience: 0.15	PERSON
3. phone Salience: 0.13	CONSUMER GOOD	4. Android <a href="#">Wikipedia Article</a> Salience: 0.12	CONSUMER GOOD
5. Sundar Pichai <a href="#">Wikipedia Article</a> Salience: 0.11	PERSON	6. Mountain View <a href="#">Wikipedia Article</a> Salience: 0.10	LOCATION

# Sentiment analysis labels feelings in text



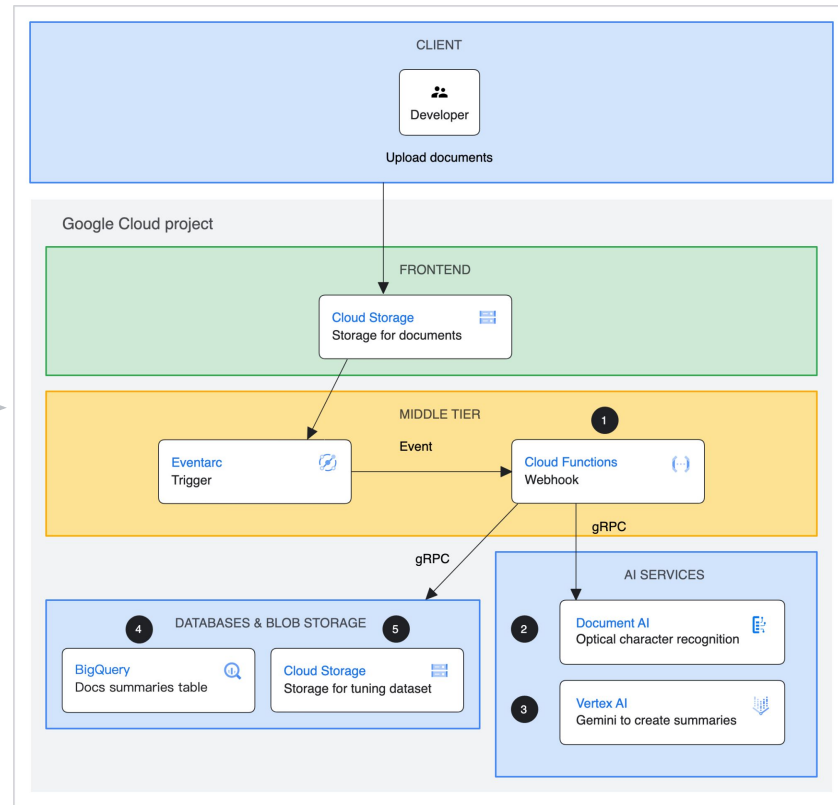
# Demo: Document API

Use cases:

- Extract data from documents (digitize)
- Summarize documents
- Classify documents

Basic Example: [Document AI | Google Cloud](#)

Example Architecture: [Generative AI Document Summarization](#).



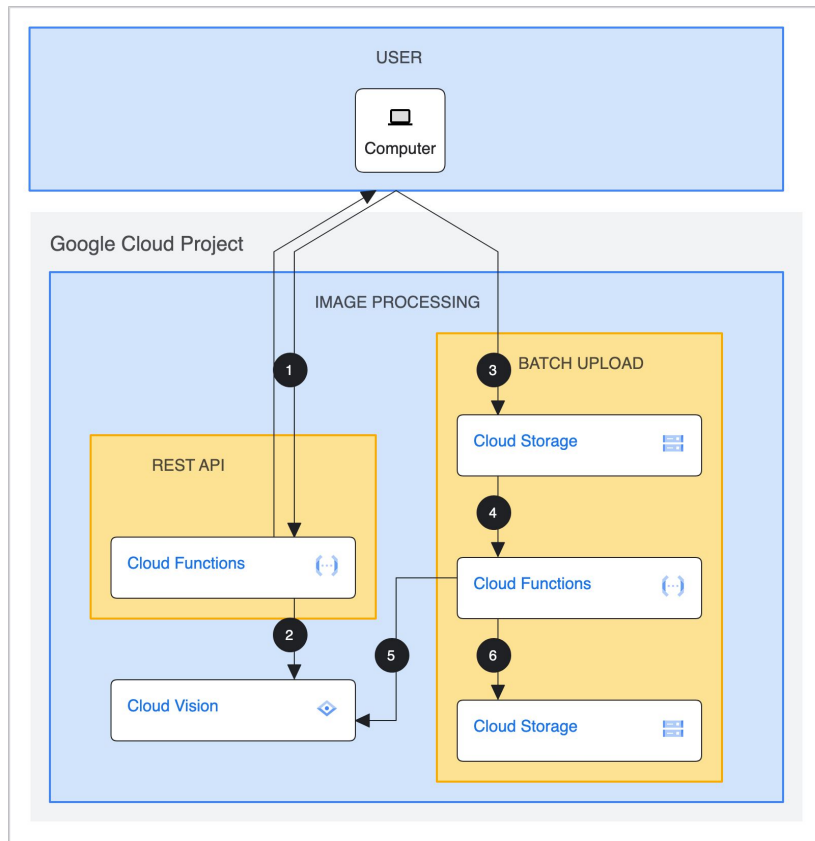
# Demo: Vision API

Use cases:

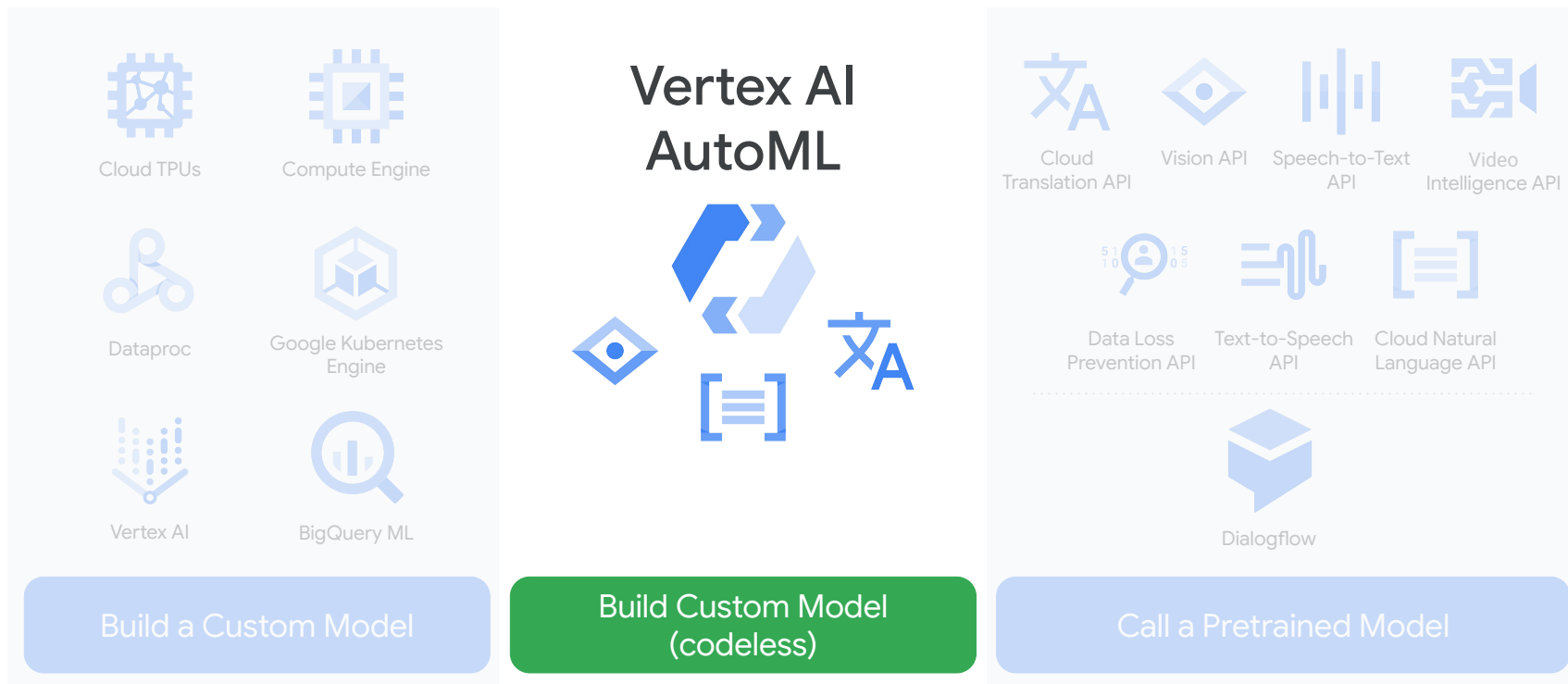
- image recognition and classification

Basic Example: [Vision AI: Image and visual AI tools | Google Cloud](#)

Example Architecture: [AI/ML Image Processing on Cloud Functions.](#)

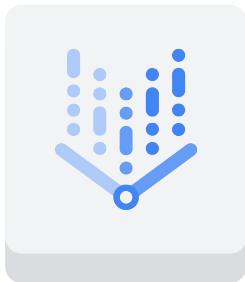


# Create and deploy custom models with Vertex AI AutoML



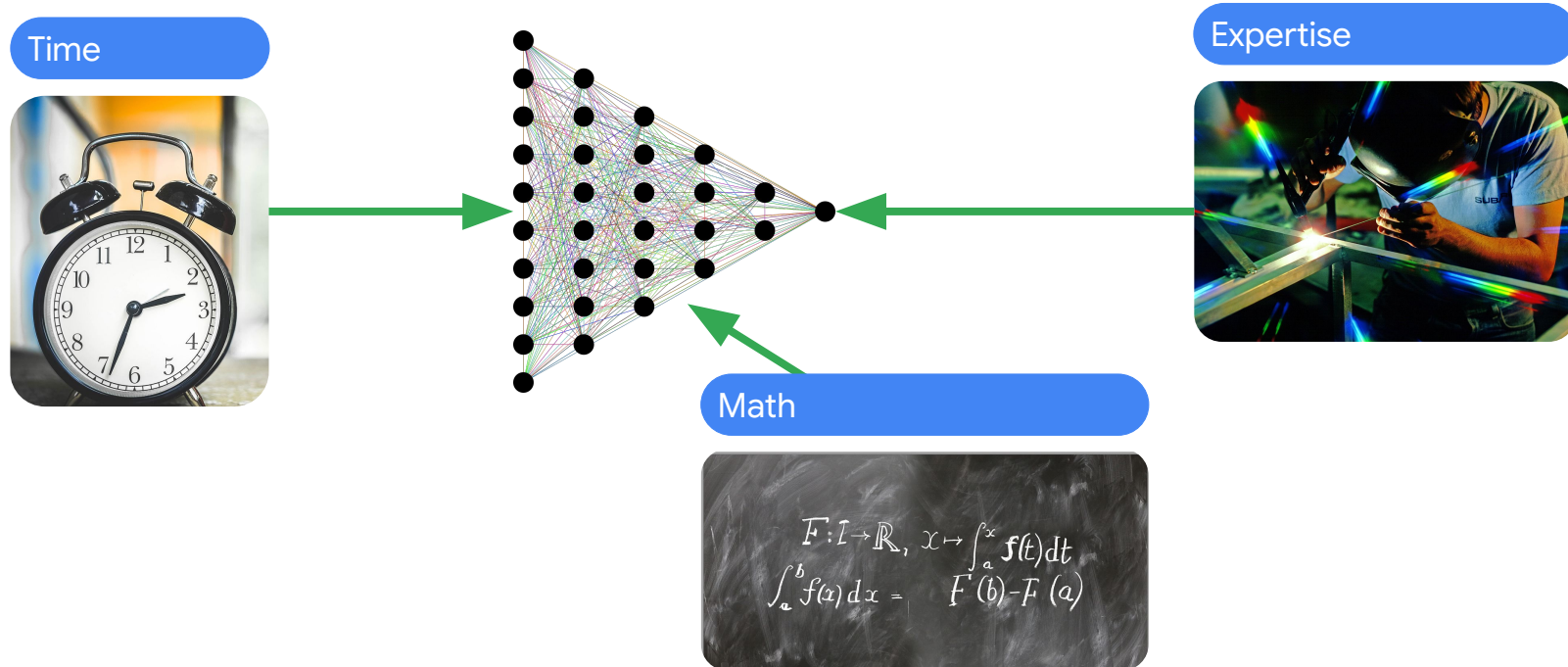
# Data analytics lifecycle: Activate

Vertex AI



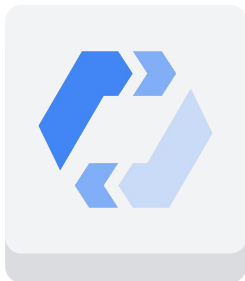
- ✓ Unified AI development platform
- ✓ Vertex AI Studio for Generative AI
- ✓ Supports end-to-end ML lifecycle
- ✓ Dataset, Model development, Deployment
- ✓ Two training mode: Custom, AutoML

# Training high-quality, custom ML models requires a lot of effort and expertise



# Data analytics lifecycle: Activate

Auto ML



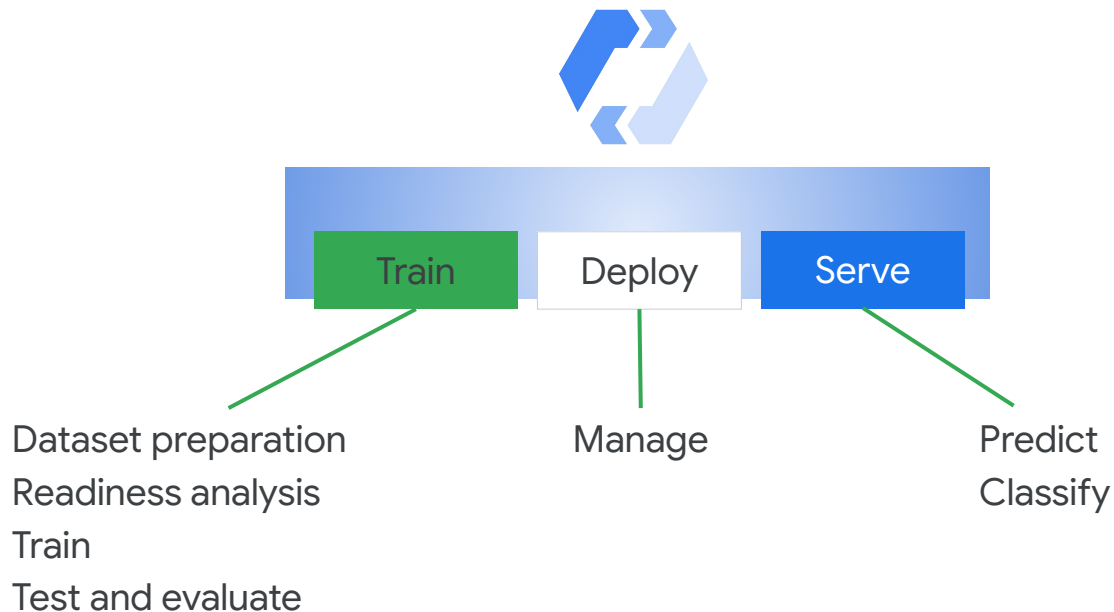
High-quality custom model with minimal effort



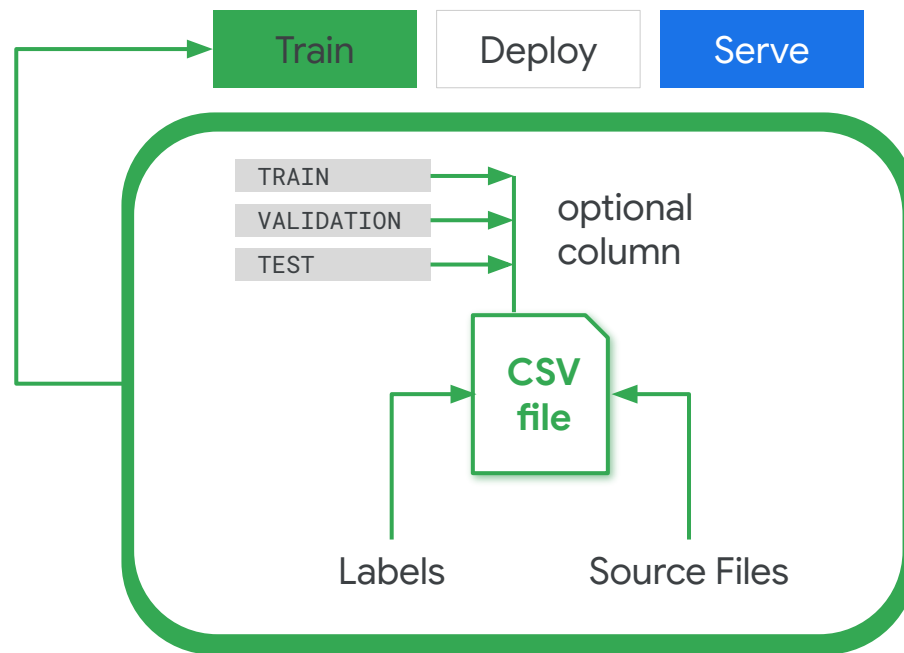
Supports various types of models

Data type	Supported Objectives
Image data	Classification, Object detection
Video data	Action recognition, classification, object tracking
Text data	Classification, Entity extraction, Sentiment analysis

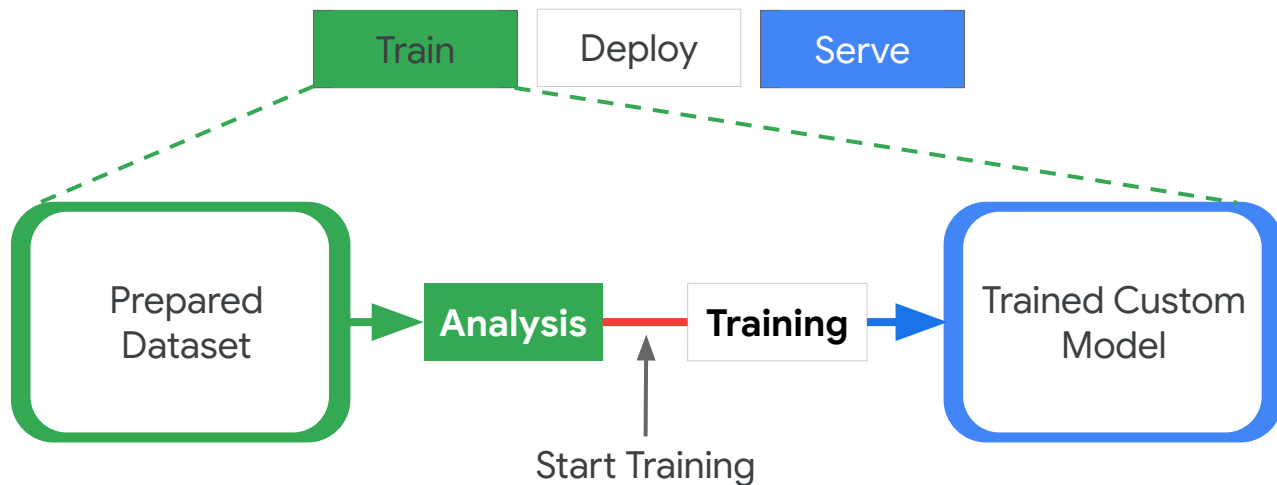
# AutoML follows a standard procedure that is divided into train, deploy, and serve phases



# AutoML uses a Prepared Dataset to train a Custom Model



AutoML performs basic checks and a preliminary analysis of the Prepared Dataset to determine if there is enough information and if it is properly organized



# Deploy your model

The screenshot shows the Google Cloud AI Platform console. At the top, there's a search bar and a dropdown menu for 'cloud-training-demos'. Below that, the breadcrumb navigation shows 'clouds > Version 1'. There are buttons for 'VIEW DATASET' and 'EXPORT'. The main navigation tabs are 'EVALUATE', 'DEPLOY & TEST' (which is active), 'BATCH PREDICT', and 'VERSION DETAILS'.

## Deploy your model

Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.

[DEPLOY TO ENDPOINT](#)

Name	ID	Status	Models	Deployment resource pool
clouds-endpoint	7938221064868331520	Deploying model	0	—

## Test your model PREVIEW

**i** Your model must be successfully deployed to an endpoint before you can test it.

[UPLOAD IMAGE](#)

# Make predictions

## Batch prediction

---

- Allows you to make many prediction requests at once.
- Is asynchronous.

## Online prediction

---

- Deploy your model to make it available for prediction requests using a REST API.
- Is synchronous (the model will quickly return a prediction, but only accepts one prediction request per API call).
- This is useful if parts of your system are dependent on a quick prediction turnaround.

# Demo: AutoML

## 1. Prepare the dataset and assign labels

Google Cloud paggie-svc-project Search (/) for resources, docs, products, and more Search

animal\_dataset animal\_dataset\_icn Train new model

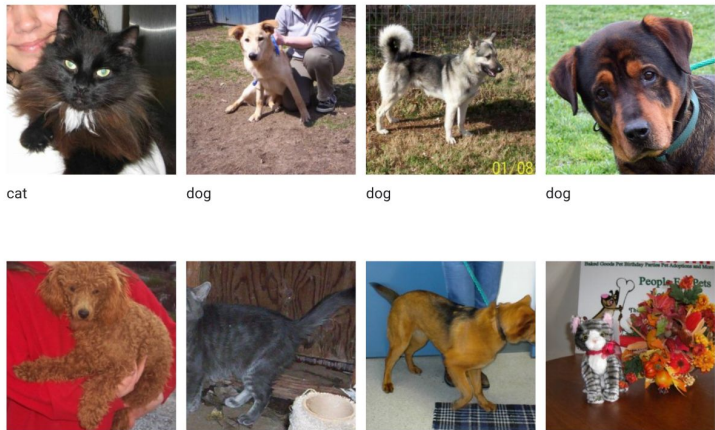
Import Browse Analyze Lineage

Labels 1< Images

Labels	Count
All	202
Labeled	202
Unlabeled	0
Filter	+
cat	101
dog	101
<a href="#">Add new label</a>	

Filter Enter label or property name

Select all



cat dog dog dog

dog cat dog cat

<https://console.cloud.google.com/vertex-ai/locations/us-central1/datasets/4169571843142844416/browse?inv=1&inv=Ab0JDw&project=paggie-svc-project&supportedpurview=project,organizationid,folder>

# Demo: AutoML

## 2. Train the model

The image shows two screenshots from the Google Cloud AutoML interface. The top screenshot is the 'Train new model' configuration screen, and the bottom screenshot is the 'Training' details page.

**Train new model configuration:**

- Dataset:** animal\_dataset
- Annotation set:** animal\_dataset\_icn
- Objective:** Image classification (Single-label)
- Model training method:**
  - AutoML**  
Train high-quality models with minimal effort and machine learning expertise. Just specify how long you want to train. [Learn more](#)
  - Custom training (advanced)  
Run your TensorFlow, scikit-learn, and XGBoost training applications in the cloud. Train with one of Google Cloud's pre-built containers or use your own. [Learn more](#)
- Choose where to use the model:**
  - Cloud**  
Deploy to an endpoint for online inferences or use your own infrastructure.
  - Edge  
Export for on-prem and on-device use. Typical use cases include image classification, object detection, and speech recognition.

**Training details page:**

Google Cloud | paggie-svc-project | Search (/) for resources, docs, products, and more

Vertex AI | animal\_model | View training dataset

Model development

- Feature Store
- Datasets
- Training**
- Experiments
- Metadata
- Ray on Vertex AI

Deploy and use

- Model Registry
- Endpoints
- Batch Inference
- Monitoring

Get API key

Provisioned Throughput

Pipelines

Tutorials

Training began at Feb 15, 2026, 2:55:01 PM and is still in progress.

Status	Training
Training pipeline ID	7531335156238385152
Created	Feb 15, 2026, 2:54:33 PM
Start time	Feb 15, 2026, 2:55:01 PM
Budget (original)	8 node hours
Elapsed time	1 min 19 sec
Region	us-central1
Encryption type	Google-managed
Dataset	animal_dataset
Annotation set	animal_dataset_icn
Data split	Randomly assigned (80/10/10)
Objective	Image classification (Single-label)

# Demo: AutoML

## 3. Deploy and test the model

← animal\_model > Version 1 ▾ [View dataset](#) [Export](#)


Evaluate **Deploy & test** Batch infer Version details Lineage

[Deploy to endpoint](#)

Name	ID	Status	Models	Deployment resource pool	Region	Monitoring	Most recent monitoring job
<a href="#">animal_model</a>	732326421145321472	Active	0	—	us-central1	Disabled	—

Test your model [Preview](#)

🔍 🔍 📏 ⚙️ ▾ Filter Filter labels



cat 0.003  
dog 0.997

[Upload image](#)

# BigQuery ML is a way to build custom models



Cloud TPUs



Compute Engine



Dataproc



Google Kubernetes Engine



Vertex AI



BigQuery ML

Build a Custom Model

## AutoML



Build Custom Model  
(codeless)



Cloud Translation API



Vision API



Speech-to-Text API



Video Intelligence API



Data Loss Prevention API



Text-to-Speech API



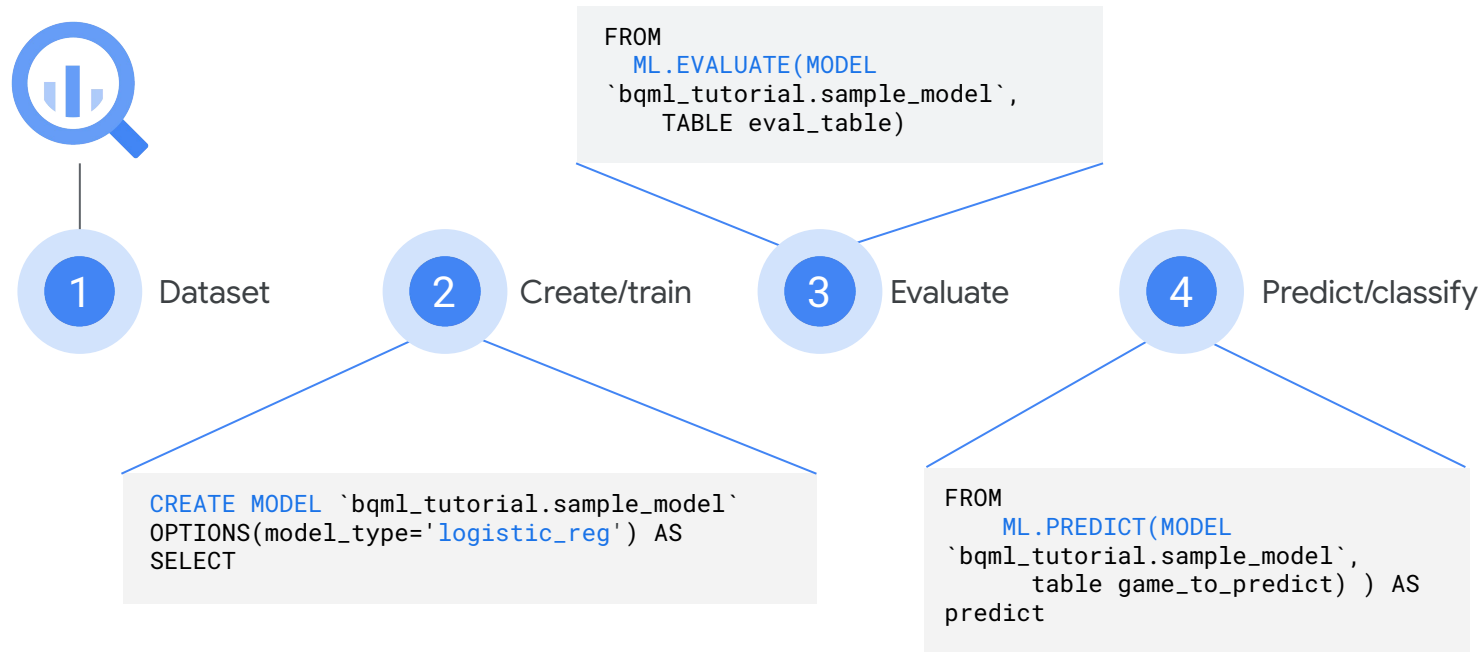
Cloud Natural Language API



Dialogflow

Call a Pretrained Model

# Working with BigQuery ML - Train your own model



# Demo BQML–Predict NYC taxi fares using BQML

In this demo, we'll build a model to predict taxi fares in NYC using one of BigQuery's public datasets `nyc-tlc.yellow.trips`.

Category	Key Columns	Description
<b>Time</b>	<code>pickup_datetime</code> , <code>dropoff_datetime</code>	The precise date and time when the meter was engaged and disengaged.
<b>Location</b>	<code>pickup_longitude</code> , <code>pickup_latitude</code> / <code>PULocationID</code>	The longitude/latitude (older data) or TLC Taxicab Boundary Zone (newer data) where the trip began/ended.
<b>Fare</b>	<code>fare_amount</code> , <code>tip_amount</code> , <code>tolls_amount</code>	The cost components: base fare, credit card tips, and bridge/tunnel tolls.
<b>Trip Details</b>	<code>passenger_count</code> , <code>trip_distance</code>	The number of passengers (driver-entered) and the elapsed trip distance reported by the taximeter.
<b>Payment</b>	<code>payment_type</code>	A numeric code signifying how the passenger paid (e.g., 1=Credit Card, 2=Cash).

# Demo BQML

## 1. Create dataset

```
#standardSQL
create or replace table `taxi.taxi300k` as
WITH taxi_preproc AS (
SELECT
  ABS(MOD(FARM_FINGERPRINT(STRING(pickup_datetime)), 10000)) AS dataset,
  (tolls_amount + fare_amount) AS fare_amount,
  pickup_datetime,
  EXTRACT(DAYOFWEEK FROM pickup_datetime) AS dayofweek,
  EXTRACT(HOUR FROM pickup_datetime) AS hourofday,
  pickup_longitude AS pickuplon,
  pickup_latitude AS pickuplat,
  dropoff_longitude AS dropofflon,
  dropoff_latitude AS dropofflat,
  passenger_count
FROM
  `nyc-tlc.yellow.trips`
WHERE
  trip_distance > 0
  AND fare_amount >= 2.5
  AND fare_amount < 200
  AND pickup_longitude > -78
  AND pickup_longitude < -70
  AND dropoff_longitude > -78
  AND dropoff_longitude < -70
  AND pickup_latitude > 37
  AND pickup_latitude < 45
  AND dropoff_latitude > 37
  AND dropoff_latitude < 45
  AND passenger_count > 0
  AND ABS(MOD(FARM_FINGERPRINT(STRING(pickup_datetime)), 10000)) < 3
)
SELECT
  dataset,
  fare_amount,
  pickup_datetime,
  hourofday,
  dayofweek,
  CAST(dayofweek * 24 + hourofday AS STRING) AS dayhour,
  pickuplon,
  pickuplat,
  dropofflon,
  dropofflat,
  SQRT(POW((pickuplon - dropofflon),2) + POW(( pickuplat - dropofflat), 2)) AS dist,
  #Euclidean distance between pickup and drop off
  pickuplon - dropofflon AS londiff,
  pickuplat - dropofflat AS latdiff,
  passenger_count
FROM taxi_preproc
WHERE dataset < 3
```

# Demo BQML

## 2. Create/train model

```
CREATE OR REPLACE MODEL
  taxi.taxifare_dnn OPTIONS (model_type='dnn_regressor',
    hidden_units=[144, 89, 55],
    labels=['fare_amount']) AS
SELECT
  fare_amount,
  hourofday,
  dayofweek,
  pickuplon,
  pickuplat,
  dropofflon,
  dropofflat,
  passenger_count
FROM
  `taxi.taxi300k`
WHERE
  dataset = 0;
```

# Demo BQML

## 3. Evaluation

paggie-svc-project / Datasets / demo\_bq

☆ taxifare\_dnn [Refresh](#)

Details Training metrics Evaluation

Mean absolute error	2.5286
Mean squared error	14.7226
Mean squared log error	0.0813
Median absolute error	2.0137
R squared	0.8204

Metric	Definition & Meaning
<b>Mean absolute error (MAE)</b>	The average of the absolute differences between the predicted fare and the actual fare. On average, your predictions are off by <b>\$2.53</b> .
<b>Mean squared error (MSE)</b>	The average of the squared differences. Because it squares the error, it penalizes large outliers more heavily than MAE.
<b>Mean squared log error</b>	Measures the relative error between the log of the predicted and actual values. It is useful when the target values have a wide range.
<b>Median absolute error</b>	The median of all error magnitudes. 50% of your predictions have an error of less than <b>\$2.01</b> . It is more robust to outliers than MAE.
<b>R squared</b>	The "Coefficient of Determination". It indicates that <b>82.04%</b> of the variance in the taxi fare is explained by your model. Closer to 1.0 is better.

# Demo BQML

## 4. Predict

1 SELECT  
2 fare\_amount,  
3 predicted\_fare\_amount,  
✓ This query will process 23.22 MB when run.

Using on-demand processing quota Processing location: US ✕

Query results [+ Create conversation](#) [Save results](#) [Open in](#)

Job information **Results** Visualization JSON Execution details Execution graph

Row	fare_amount	predicted_fare_a...	hourofday	dayofweek	pickuplon	pickuplat
1	4.6	7.020443916320...	4	4	-73.958026	40.77446
2	12.2	14.52977752685...	4	4	-74.006158	40.74366
3	9.8	12.53529262542...	4	4	-73.999302	40.72825
4	6.1	8.04678726196289	4	4	-74.005826	40.7402
5	4.1	7.401981353759...	4	4	-73.962267	40.76062
6	6.9	10.7291841506958	4	4	-74.006501	40.73972
7	5.7	10.26519107818...	4	4	-73.949382	40.80225
8	4.9	9.40158748626709	4	4	-73.978439	40.7413
9	4.1	9.242655754089...	4	4	-73.987827	40.74882
10	4.5	7.433818340301...	4	4	-73.960693	40.77812

```

SELECT
  fare_amount,
  predicted_fare_amount,
  hourofday,
  dayofweek,
  pickuplon,
  pickuplat,
  dropofflon,
  dropofflat,
  passenger_count
FROM
  ML.PREDICT(MODEL taxi.taxifare_dnn, (SELECT
    fare_amount,
    hourofday,
    dayofweek,
    pickuplon,
    pickuplat,
    dropofflon,
    dropofflat,
    passenger_count
  FROM
    `taxi.taxi300k`
  WHERE
    dataset = 2
  ))
LIMIT 10

```

# Working with BigQuery ML - Inference function

```

1 SELECT
2   title,
3   AI.GENERATE(CONCAT("Summarize in one sentence: ", body)).result AS article_summary
4 FROM `bigquery-public-data.bbc_news.fulltext`
5 LIMIT 3;

```

✓ Query completed

Query results

[+ Create conversation](#)

[Link](#)

Job information	Results	Visualization	JSON	Execution details	Execution graph
Row	title	article_summary			
1	Global digital divide 'narrowing'	A World Bank report claims the "digital divide" between rich and poor nations is rapidly narrowing due to widespread access to technology, while the UN's World Summit on the Information Society maintains the divide is still very real and has established a voluntary Digital Solidarity Fund to address it.			
2	Sporting rivals go to extra time	EA Sports' exclusive acquisition of ESPN branding and talent rights, following an earlier NFL deal, has eliminated a competing sports game series and sparked significant fan concern over market monopolization and reduced consumer choice.			
3	Global blogger action day called	The Committee to Protect Bloggers is mobilizing the global web blog community to dedicate their sites on February 22nd to "Free Mojtaba and Arash Day" in support of two imprisoned Iranian bloggers, Mojtaba Saminejad and Arash Sigarchi.			

# Agenda

01 The Data Analytics Lifecycle Overview

---

02 Ingest

---

03 Process

---

04 Store

---

## Analyze

Introduction to BigQuery

Querying External Data Sources or Public Datasets **Demo**

Control access with Policy Tag **Demo**

Building Pipelines in BigQuery - from low-code to coding method **Demo**

Visualization & Business Intelligence with Looker & Looker Studio **Demo**

Derive insights from your data in BigQuery (Agent, Gemini Enterprise) **Demo**

---

06

## Activate

Introduction AI on Google Cloud

BigQuery ML **Demo**

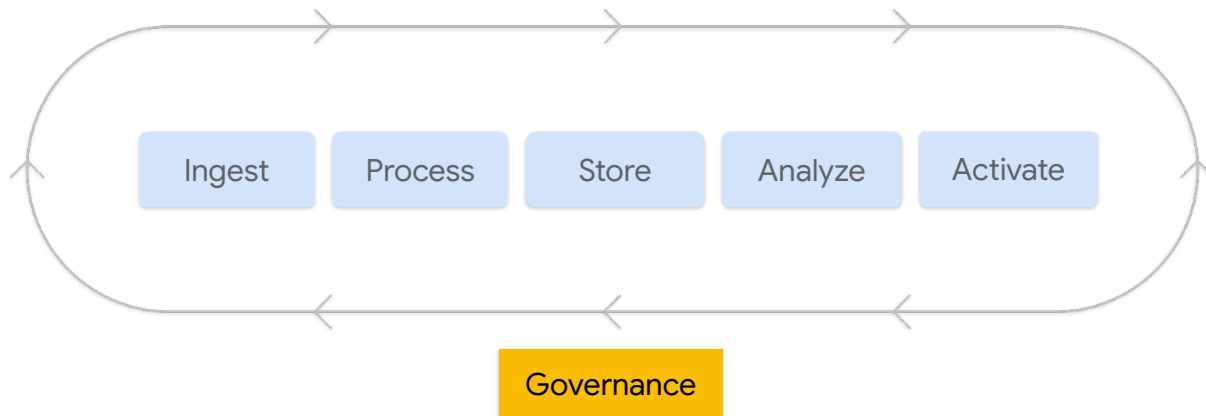
---

07

Governance

---

# Data analytics lifecycle



**Objective:** Data is safe, compliant, and reliable across the enterprise.

**This layer focuses on:**

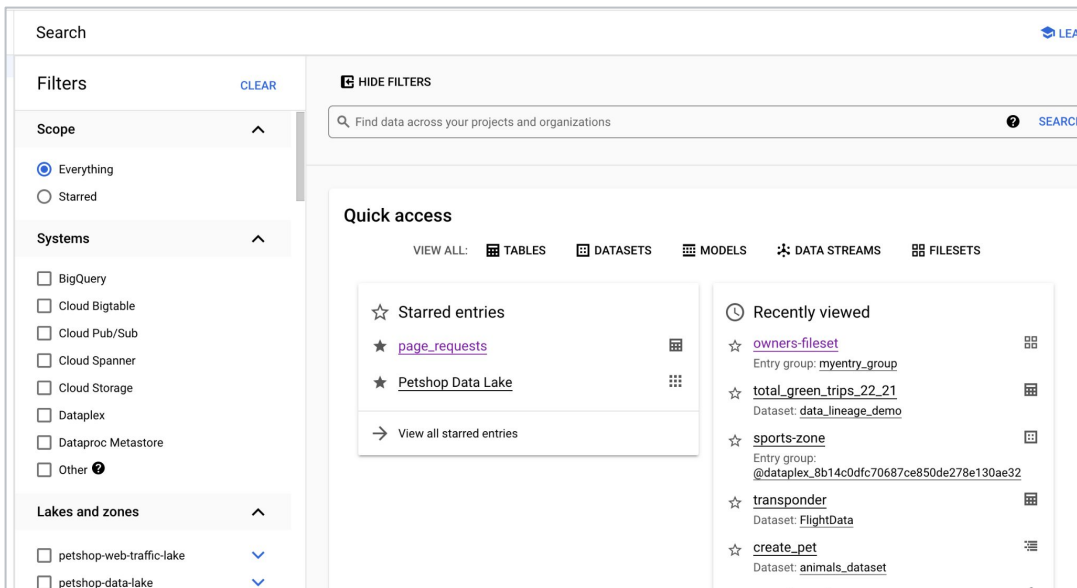
- Data access control and security
- Data Catalog
- Data quality monitoring
- Policy enforcement and compliance

**Related Tools:**

- BigQuery
- Dataplex

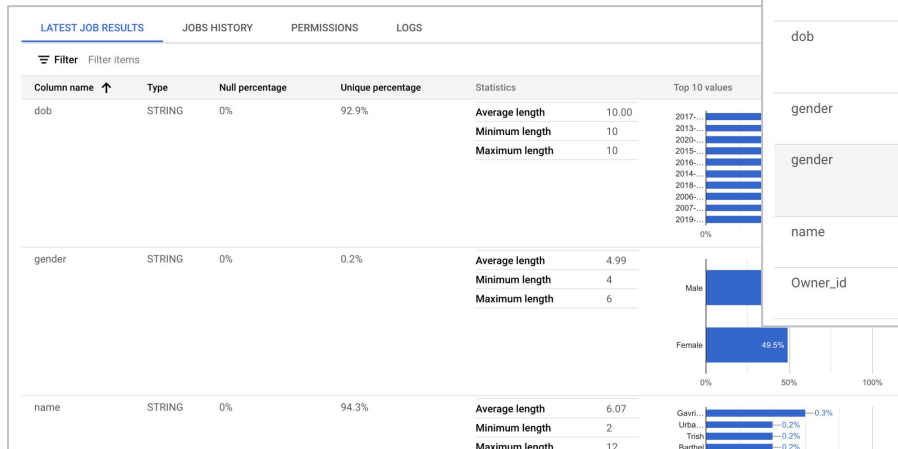
# Dataplex Universal Catalog allows you to discover data across your organization

- Search and filter by:
  - Projects
  - Data sources
  - Lakes and zones
  - Tags
  - Metadata
  - Etc.
- Favorites allow for quick access to known datasets



# Intelligent scanning allows schedule scans to ensure data quality over time

- Automatically profile data assets
- Scan data for quality issues and PII
- Track lineage of data assets over time



Column name ↑	Rule type	Status	Evaluation	Dimension	Parameters	Failed rows
dob	Null Check	✓ Passed	Per row	Completeness		0%
dob	Row Condition Check	✓ Passed	Per row	Validity	(LENGTH('dob') >= 10 A...	0%
gender	Value Set Check	✓ Passed	Per row	Validity	set of: Male, Female	0%
gender	Row Condition Check	✓ Passed	Per row	Validity	(LENGTH('gender') >= 4 ...	0%
name	Null Check	✓ Passed				
Owner_id	Null Check	✓ Passed				

**Dimensions succeeded**

- Completeness** ✓ 0 Errors
- Validity** ✓ 0 Errors
- Uniqueness** ✓ 0 Errors



Database, Data lake, Data Warehouse, Data Lakehouse



Data analytics lifecycle, including which Google Cloud products are best suited for each step

Google Cloud