

Big Data de Google

Ismael Yuste Varona

Strategic Cloud Engineer, Google

KSchool / Marzo 2020

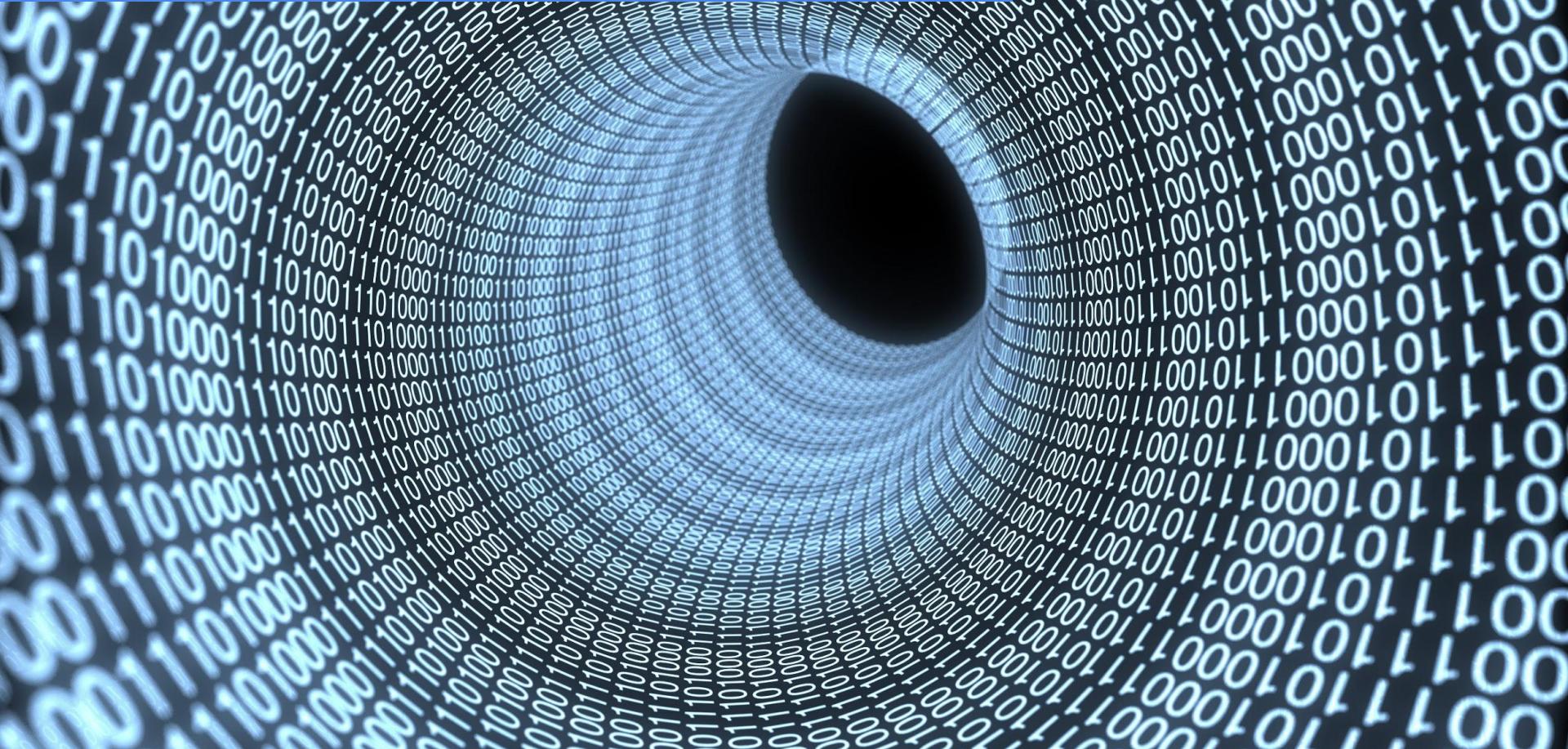
Agenda

- Big Data
- Pub-Sub
- DataProc
- DataFlow
- BigQuery
- DataPrep
- AI Notebooks
- DataStudio
- Composer
- Data Fusion
- Data Catalog

GCP Big Data



A look at Google Scale



... in just 1 minute:



3M Searches

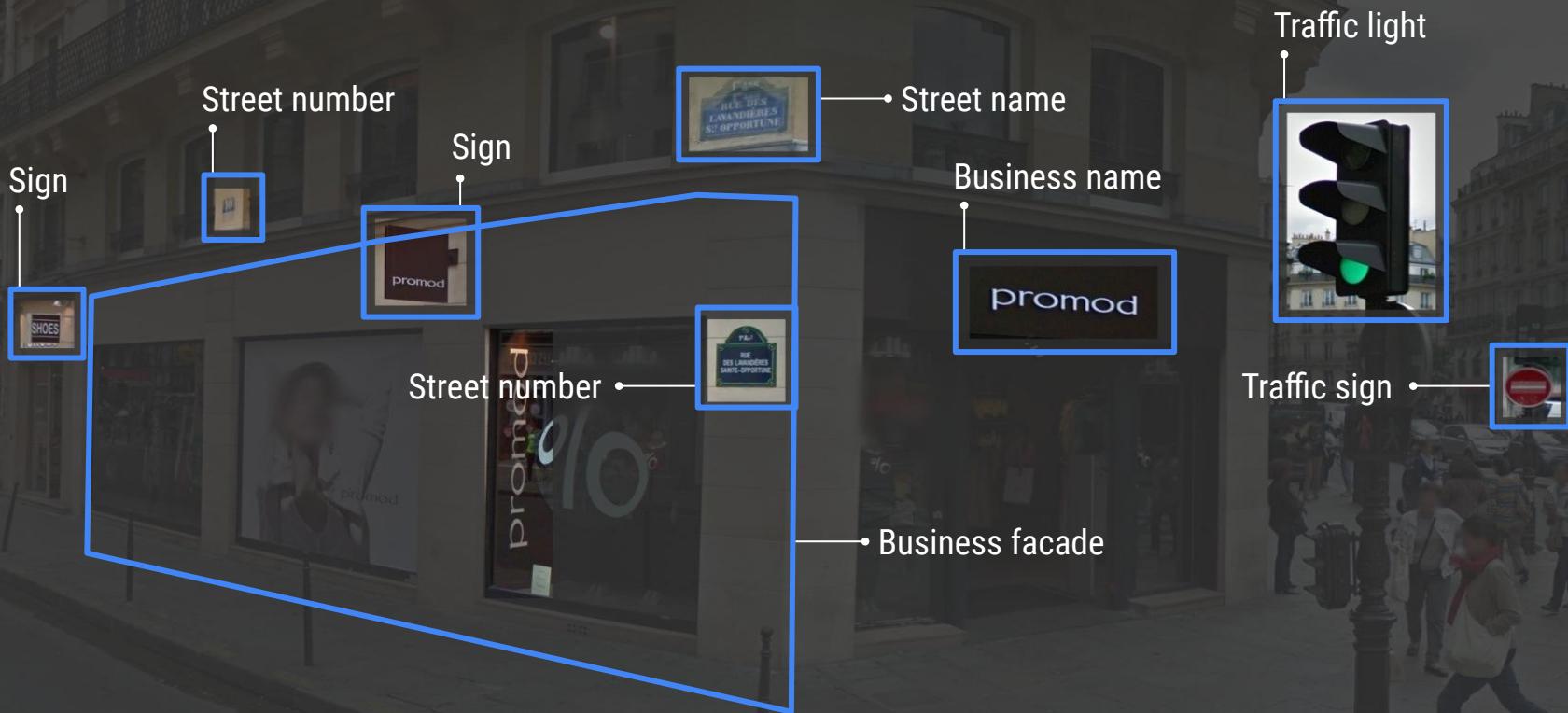


100 Hours



1000 new
devices

Finding new value in data



Data Lifecycle Steps

Ingest

The first stage is to pull in the raw data, such as streaming data from devices, on-premises batch data, application logs, or mobile-app user events and analytics.

Store

After the data has been retrieved, it needs to be stored in a format that is durable and can be easily accessed.

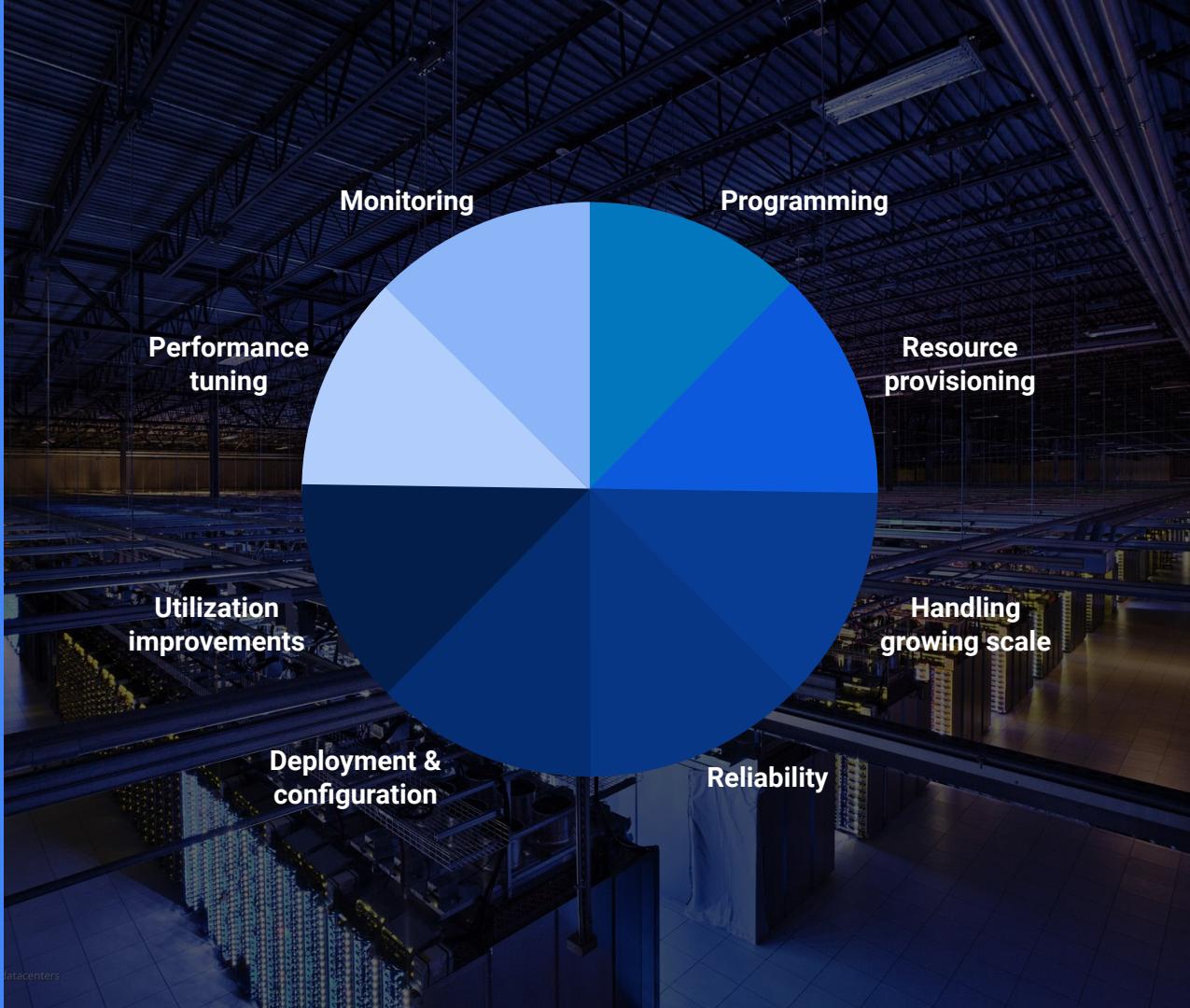
Process & Analyze

In this stage, the data is transformed from raw form into actionable information.

Explore & Visualize

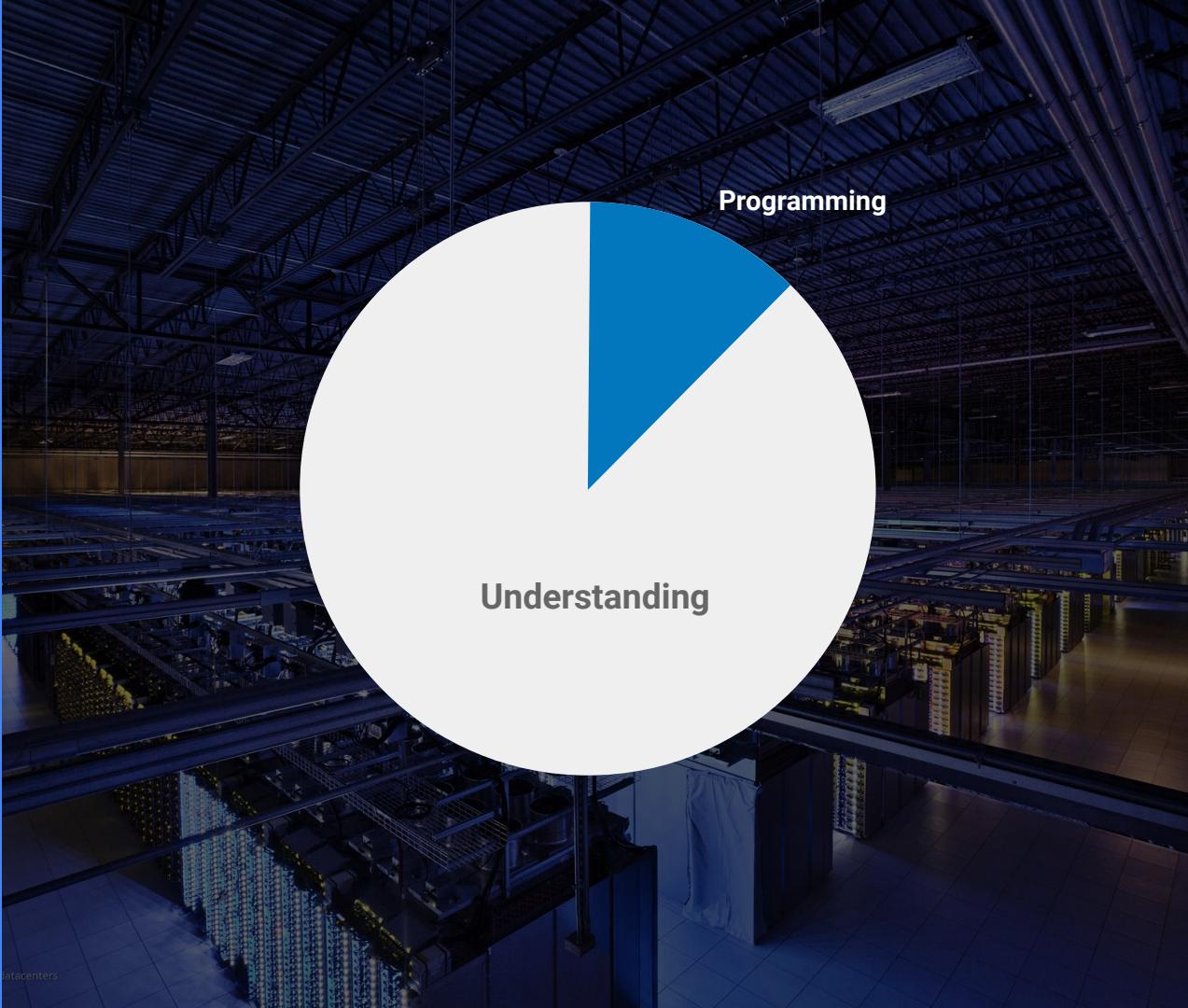
The final stage is to convert the results of the analysis into a format that is easy to draw insights from and to share with colleagues and peers.

Typical Big Data Jobs



Big Data with Google

Focus on insights.
Not infrastructure.
From batch to real-time.



Data & Analytics

Cloud Dataproc

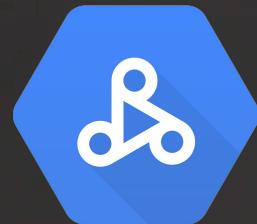
Fully managed Hadoop and Spark with industry-leading performance

BigQuery

Fully managed data warehouse for large-scale analytics

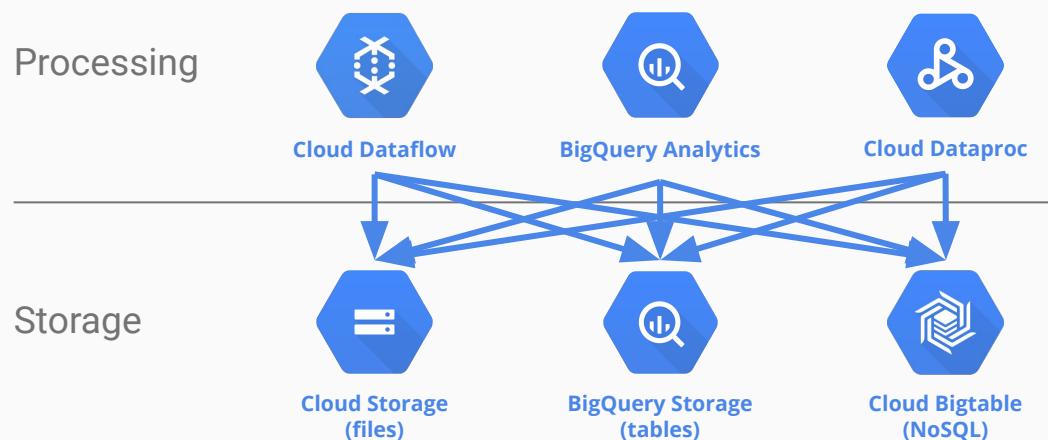
Cloud Dataflow

Real-time data pipelines, with open source SDK via Apache Beam

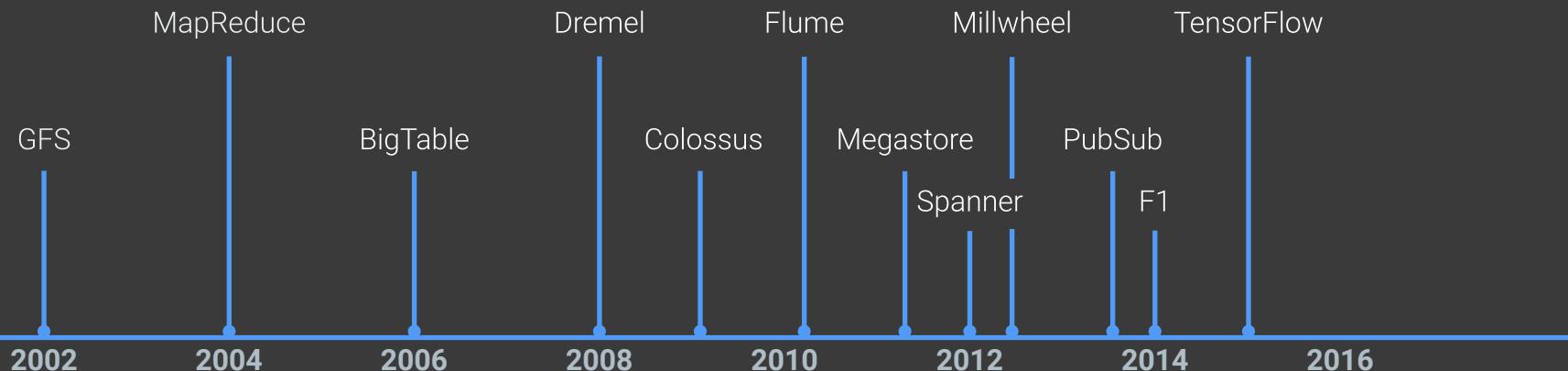


Separation of Storage and Compute

- Access any storage system from any processing tool
- Keep as much data as you want, economically
- Share data in place, no more FTP and copying



Google's Data Research



Google's Data Products



Dataproc



BigQuery



Dataflow



Dataflow



ML Engine



Cloud
Storage



Bigtable



Cloud
Storage



Datastore

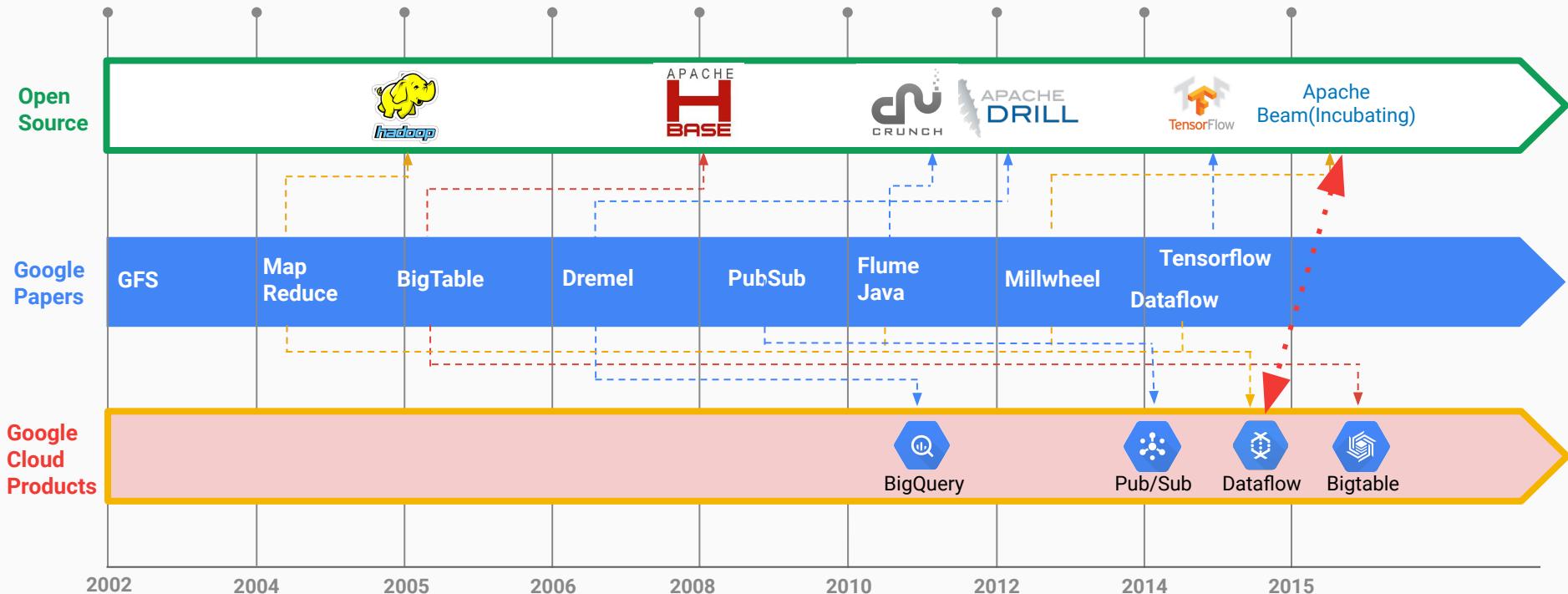


Pub/Sub



Spanner

10+ years of Big Data innovation - Open Source



Product Mapping



Cloud
Pub/Sub



BigQuery



Cloud
Dataflow



Cloud
Dataproc



Cloud
Datalab



Pub Sub



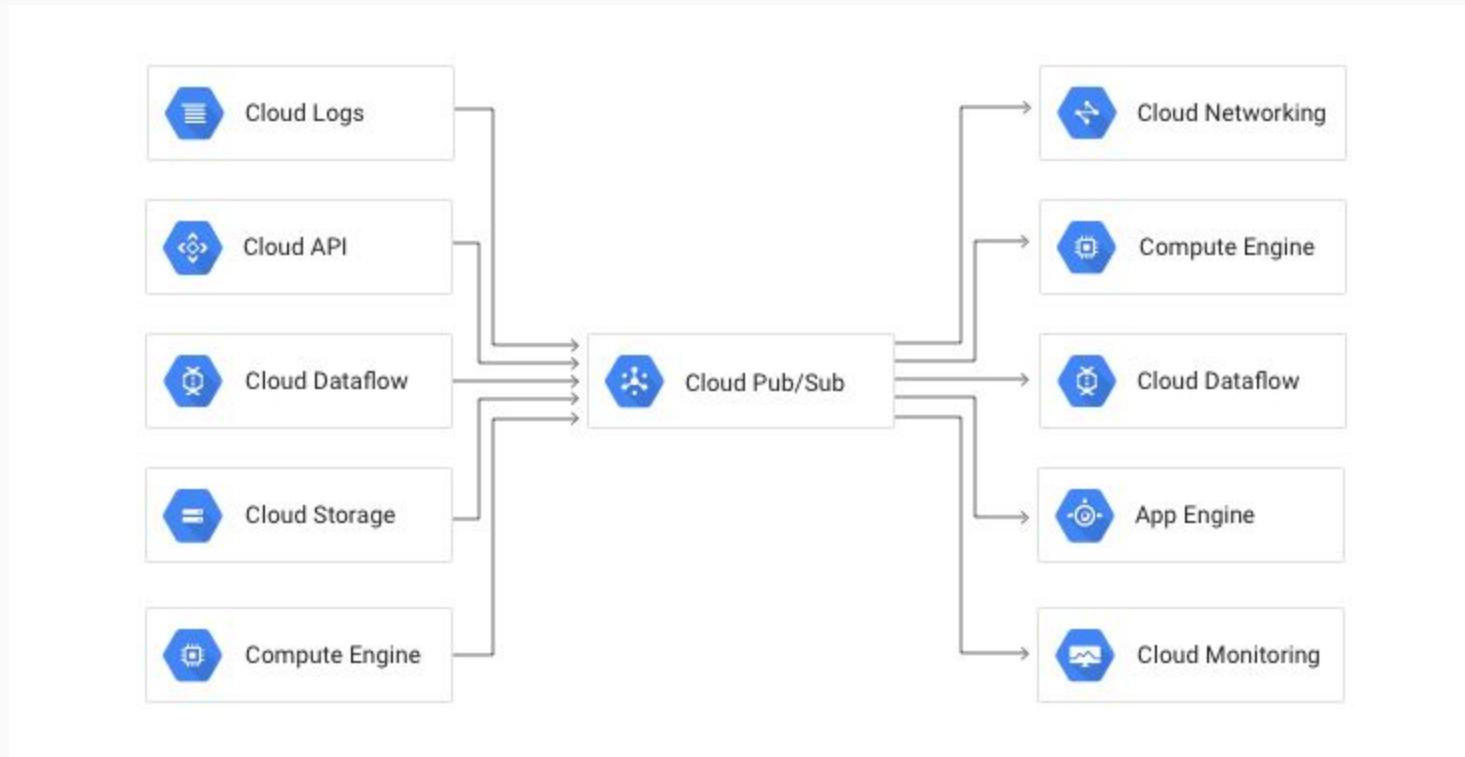
Pub/Sub: 100% serverless event delivery



Google Cloud
Pub/Sub

- ✓ Reliable and real-time messaging
- ✓ Global by design and highly available
- ✓ Uses Google's private fiber network and worldwide points of interconnect
- ✓ Only pay for what you use

Pub/Sub: Google Cloud Pub/Sub passes messages

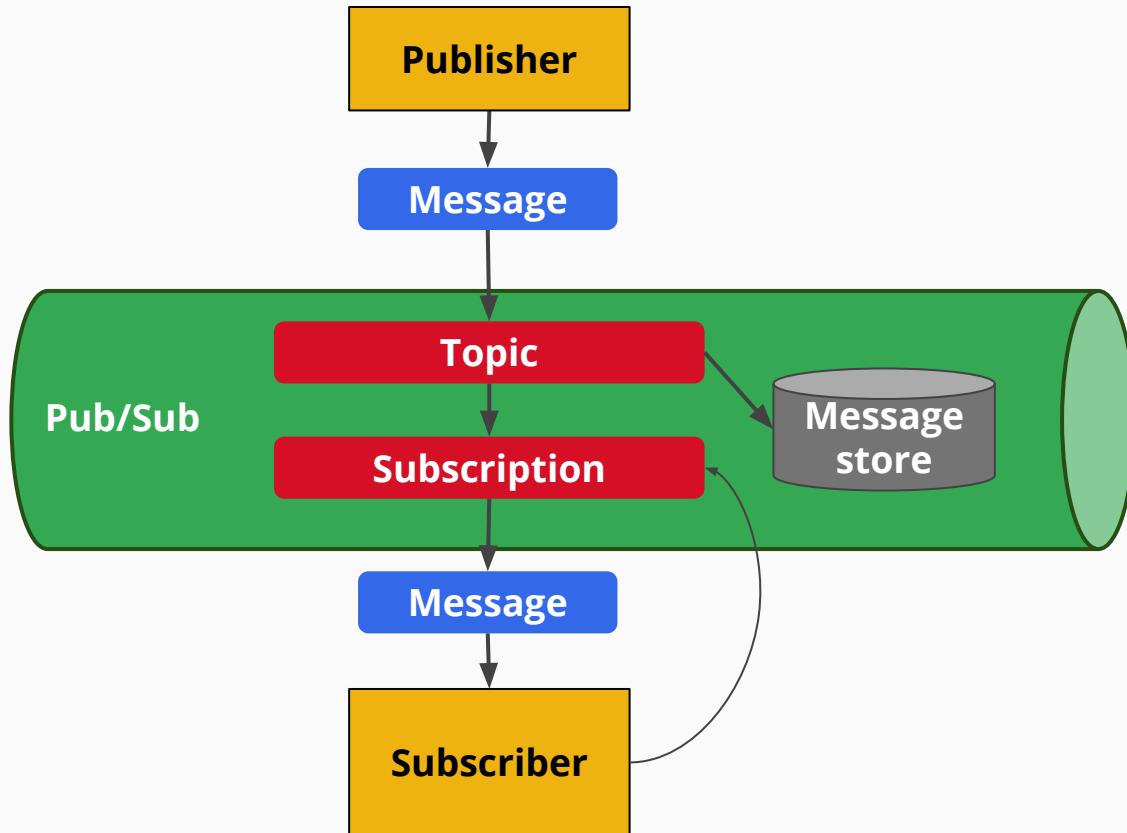


Pub/Sub: What is Google Cloud Pub/Sub

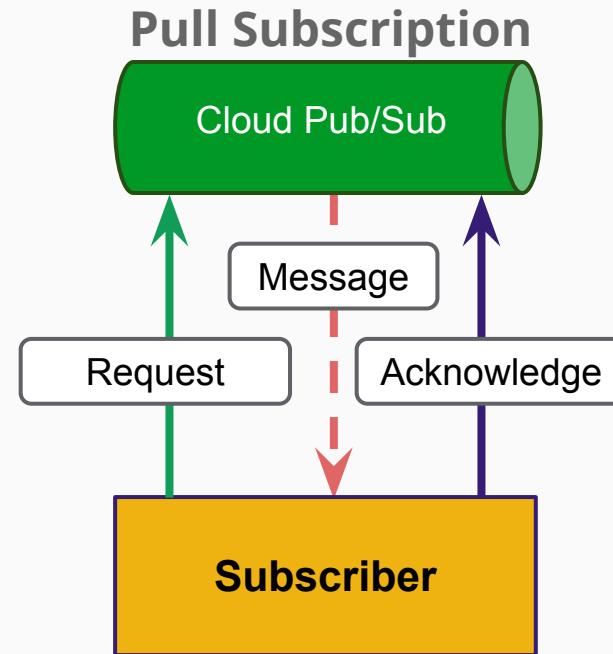
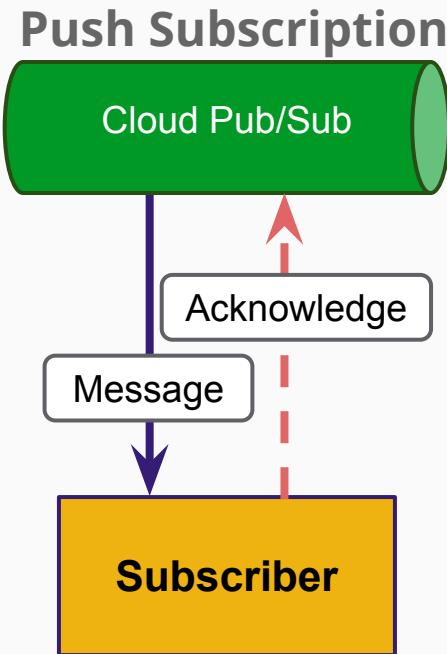
- Cloud Pub/Sub is a service to capture and rapidly pass massive amounts of data (or messages) between software applications with world-class security.
- It uses the **Publish - Subscribe** pattern:
 - **Publisher** applications can send **messages** to a **topic**.
 - **Subscriber** applications can subscribe to that topic to receive the message when the subscriber is ready (asynchronously).
- Pub/Sub acts as a buffer between sending and receiving software applications, making it easier for developers to connect applications.

- Cloud Pub/Sub provides:
 - **Scale**—It supports Google's services, such as Gmail and ads.
 - **Reliability**—Dedicated resources in every Google Cloud Platform region enhance availability without increasing latency.
 - **Performance**—Sub-second notifications even when tested at over 1 million messages per second.
 - **Cost-efficiency**—A “pay for what you use” service.
 - **Ease of use and implementation**—Because it's a fully managed service, there's no need to manage your own open source software implementation. Get started in minutes, not days.

Pub/Sub: Architecture



Pub/Sub: Architecture



Legend

- [HTTP POST request](#)
- [HTTP GET request](#)
- [HTTP response](#)

Pub/Sub: Messaging is a shock-absorber

Availability

- Buffer messages and requests during outages
- Prevent message overloads that cause outages
- Redirect requests to recover from outages

Throughput

- Smooth out spikes in new request rate
- Balance load across multiple servers
- Balance arrival rate with service rate
- “Fan-in” from many devices

Latency

- Accept requests closer to the network edge
- Optimize message flow across regions

Pub/Sub: Messaging is a shock-absorber

Sources

- New data sources can plug into old data flows
- New data sources can use new schemas
- Common security policies for all sources

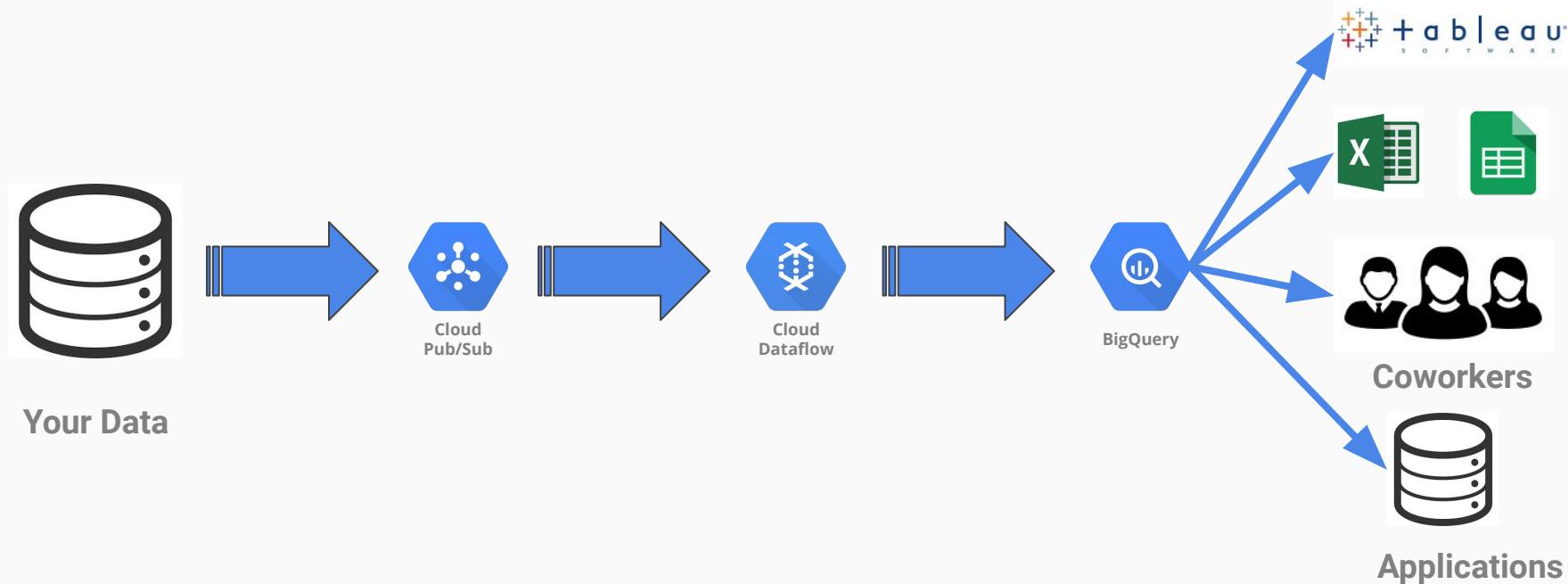
Sinks

- Data can be sent to new destinations
- “Push” and “pull” delivery are both available
- Spans organizational boundaries

Transforms

- Can merge streams into new topics
- Sends messages to Dataproc and Dataflow for transformation

Pub/Sub: Use case: big data stream



Pub/Sub: examples in use at Google

Chat & Mobile

Every time your Gmail displays a new message, it's because of a push notification to your browser or mobile device.

Ads and budgets

One of the most important real-time information streams in the company is advertising revenue—we use Pub/Sub to broadcast budgets to our entire fleet of search engines.

Push notifications

Google Cloud Messaging for Android delivers billions of messages a day, reliably and securely, for Google's own mobile apps and the entire developer community.

Instant search

Updating search results as you type is a feat of real-time indexing that depends on Pub/Sub to update caches with breaking news.

Ejercicio Pub-Sub



Quickstart: using the console

This page shows you how to perform basic tasks in Pub/Sub using the Google Cloud Console.



Note: If you are new to Pub/Sub, we recommend that you start with the [interactive tutorial](#).

Before you begin

1. Set up a Cloud Console project.

[Set up a project](#)

Click to:

- Create or select a project.
- Enable the Pub/Sub API for that project.

You can view and manage these resources at any time in the [Cloud Console](#).

Bonus

Quickstart: Using the gcloud Command-Line Tool

Pub/Sub is a messaging service for exchanging event data among applications and services. A producer of data publishes messages to a Pub/Sub topic. A consumer creates a subscription to that topic. From this point on, Pub/Sub guarantees that the message will be delivered to every consumer of the message at least once. Consumers either pull messages from a subscription or are configured as webhooks for push subscriptions. Every subscriber must acknowledge each message within a configurable time window. Unacknowledged messages are redelivered. Pub/Sub is geographically global and does not require sharding or additional configuration to scale with demand.

This page shows you how to perform basic tasks in Google Cloud Pub/Sub.

Dataproc

Dataproc



Google Cloud
Dataproc

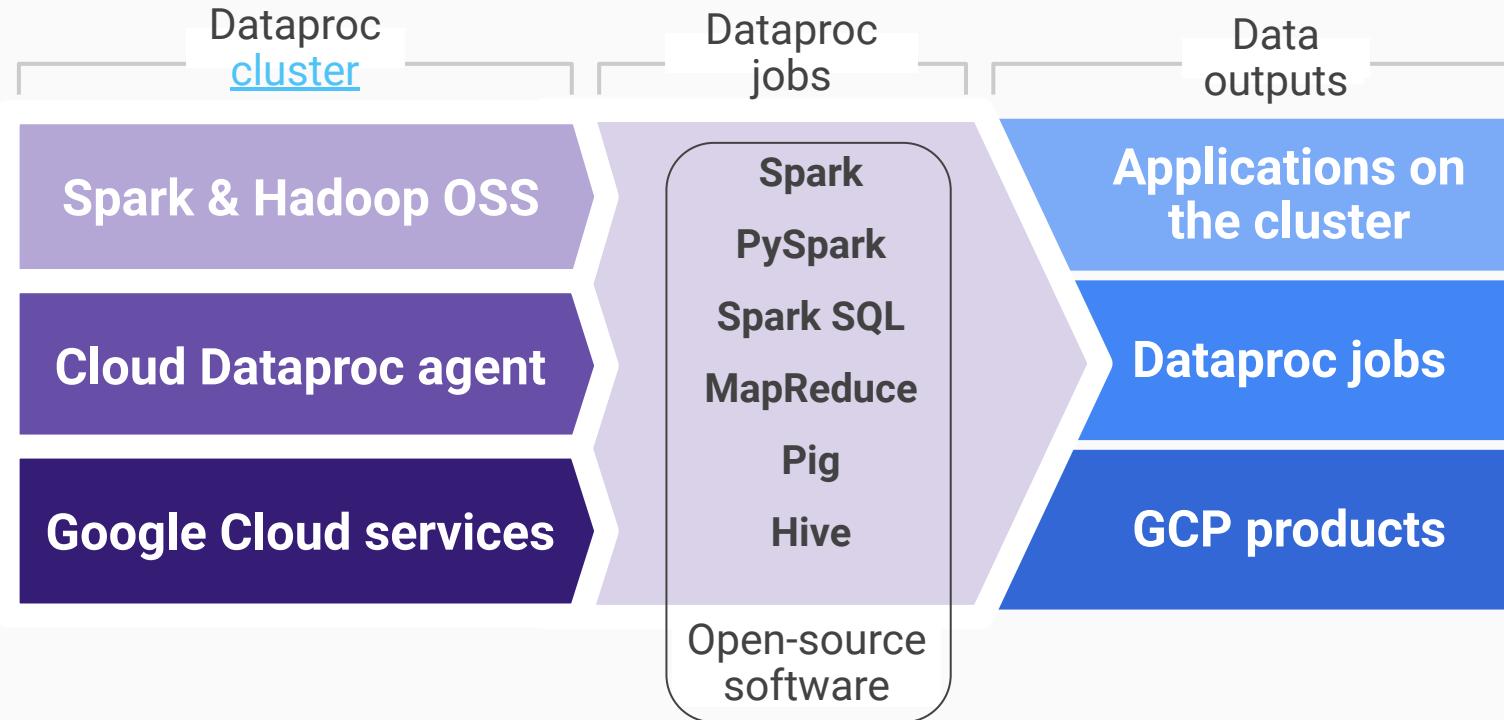
- It is a managed [Hadoop MapReduce](#), [Spark](#), [Pig](#), and [Hive](#) service, to easily process big datasets at low cost
- Is a fast, easy to use, low cost and fully-managed service, powered by Google Cloud Platform



*Servicio
gestionado Spark
y Hadoop*

- Gestión de Cluster integrado.
- Cluster dimensionables.
- Integración.
- Versionado.
- Herramientas de Gestión.
- Acciones de inicialización.
- Gestión manual o automática.
- Máquinas Virtuales flexibles.

Dataproc



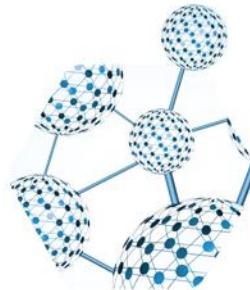
- **Apache Spark™** is a fast and general engine for large-scale data processing.
- The Spark Python API (**PySpark**) exposes the Spark programming model to Python.
- **Spark SQL** is Apache Spark's module for working with structured data.
- **MapReduce** is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.
- **Apache Pig** is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.
- The **Apache Hive** ™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Structure can be projected onto data already in storage. A command line tool and JDBC driver are provided to connect users to Hive.

Dataproc: Features



Native Spark and Hadoop

Run Spark and Hadoop applications out of the box without modification.



Cloud Integrated

Integrated with Cloud Storage, Cloud Logging, Cloud Monitoring, and more.



Minute-by-Minute Billing

While active, Dataproc clusters are billed minute-by-minute.



Preemptible VMs

Dataproc clusters can make use of low-cost preemptible Compute Engine VMs.

Dataproc: Features



Anytime Scaling

Manually scale clusters up or down based on need, even when jobs are running.



Initialization Actions

Execute scripts on cluster creation to quickly customize and configure clusters.



Developer Tools

REST API and Integration with Google Cloud SDK for rapid development.



Easily Configured

Select between multiple Spark and Hadoop versions; configure properties easily.

Dataproc: Benefits



Low-cost

Enjoy lower total cost of ownership due to low prices and minute-based billing.



Superfast

Spend less time waiting for things to happen and more time hands-on with data.



Customizable

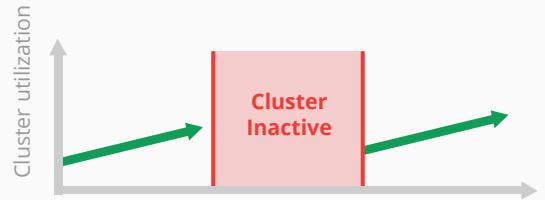
Customizable virtual machine types let you create right-sided clusters.



Easy

Vanilla Spark and Hadoop supported by purpose-built Cloud products.

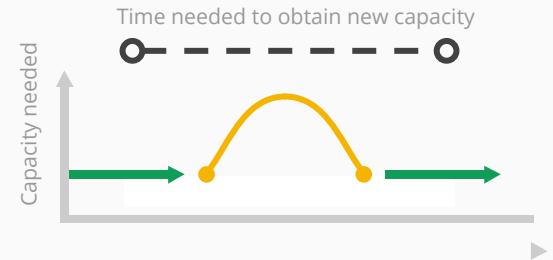
Idle clusters



Most traditional clusters are utilized only a portion of the time they're online

Idle cluster capacity (time, money, computing capacity) is wasted

Scaling inflexibility

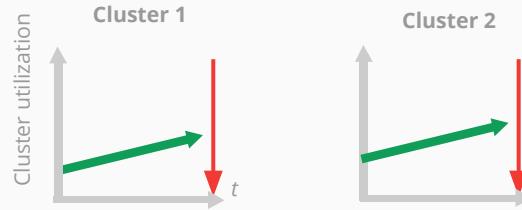


Job demands can be hard to predict, and scaling can take considerable time

Clusters may be constrained at the time when you need them most

Dataproc: Scaling

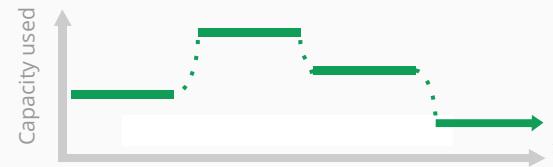
Anytime clusters



Clusters aren't idle. When work is done, simply turn off the cluster and create a new one when required.

Run clusters only when you need them

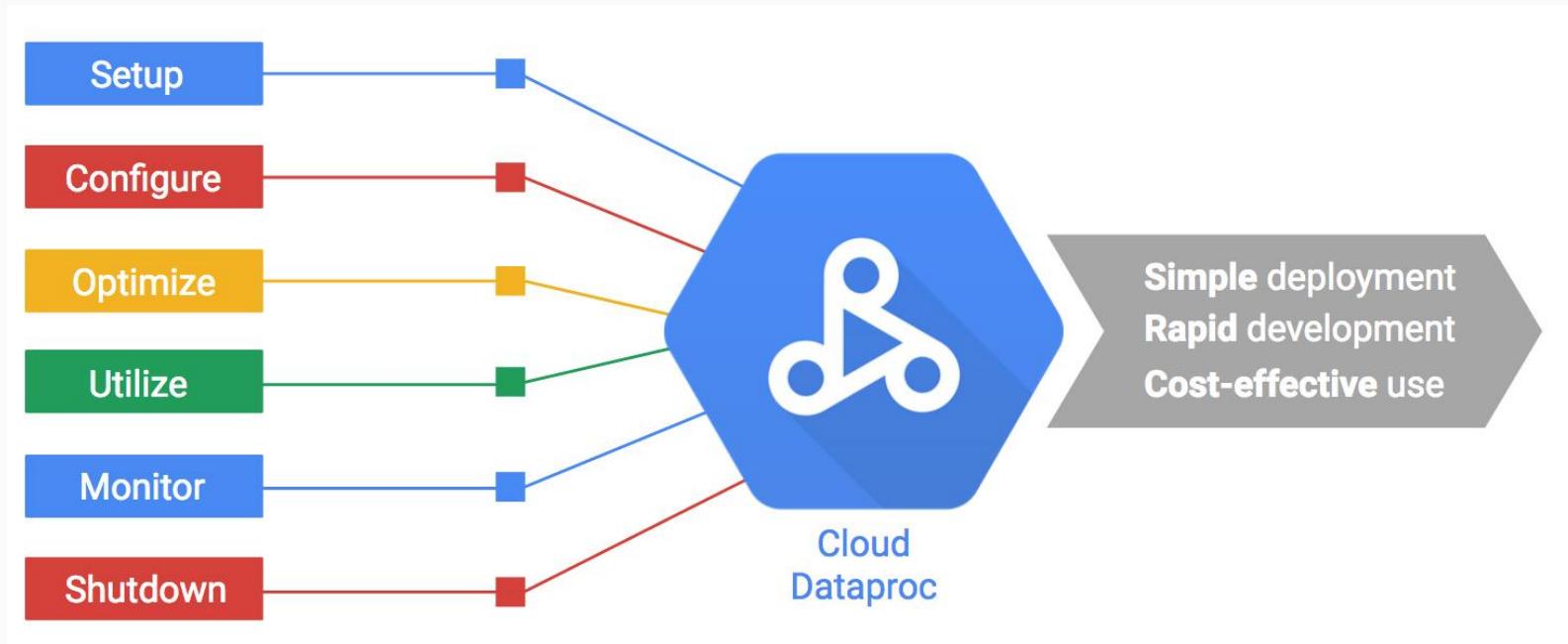
Flexible scaling



By scaling clusters at anytime, your jobs can get exactly the resources they need when required.

Clusters are the right size, anytime

Dataproc: Managed Service



Dataproc: Use Case #1: Log Processing

Need

A customer processes 50 gigabytes of (text) log data per day to produce aggregated metrics. They have used a persistent on-premise cluster to store and process the logs with MapReduce.

How Dataproc addresses this need

Google Cloud Storage can act as a landing zone for the log data for low-cost and high-durability storage. A Dataproc cluster can be created in less than 2 minutes to process this data with their existing MapReduce. Once finished, the Dataproc cluster can then be removed immediately.

Dataproc value

Instead of running all the time and incurring costs even when not used, Dataproc only runs to process the logs which saves money and reduces complexity.

Dataproc: Use Case #2: Ad-hoc Data Mining & Analysis

Need

A customer uses Spark standalone on one computer to perform data mining and analysis. The data is stored locally and they are using the Spark shell to examine the data along with Spark SQL.

How Dataproc addresses this need

Dataproc can create clusters that scale for speed and mitigate any single point of failure. Since Dataproc supports Spark, Spark SQL, and PySpark, they could use the web interface, Cloud SDK, or the native spark shell via SSH to perform their analysis safe from a single machine failure.

Dataproc value

Dataproc quickly unlocks the power of the cloud for anyone without added technical complexity. Running complex computations now take seconds instead of minutes or hours.

Dataproc: Use Case #3: Machine Learning

Need

A customer uses the Spark Machine Learning Libraries (MLlib) to run classification algorithms on very large datasets. They rely on cloud-based machines where they install, and customize Spark.

How Dataproc addresses this need

Since Spark and the MLlib installed on any Dataproc cluster, the customer can save time by quickly creating Dataproc clusters. Any additional customizations can be applied easily to the entire cluster via initialization actions. To keep an eye on workflows, they can use the built-in Cloud Logging and Monitoring.

Dataproc value

With Dataproc, resources spent on cluster creation and management can now be focused on the data. Integrations with new Google Cloud products unlock new features for Spark clusters.

Ejercicio DataProc

Introduction to Cloud Dataproc: Hadoop and Spark on Google Cloud Platform

14 min

Updated Oct 31, 2019



Start

Bonus

<https://codelabs.developers.google.com/>

Dataproc

Preprocessing BigQuery Data with
PySpark on Cloud Dataproc

40 min

Updated Sep 27, 2019



Start

Provisioning and Using a Managed
Hadoop/Spark Cluster with Cloud
Dataproc (Command Line)

20 min

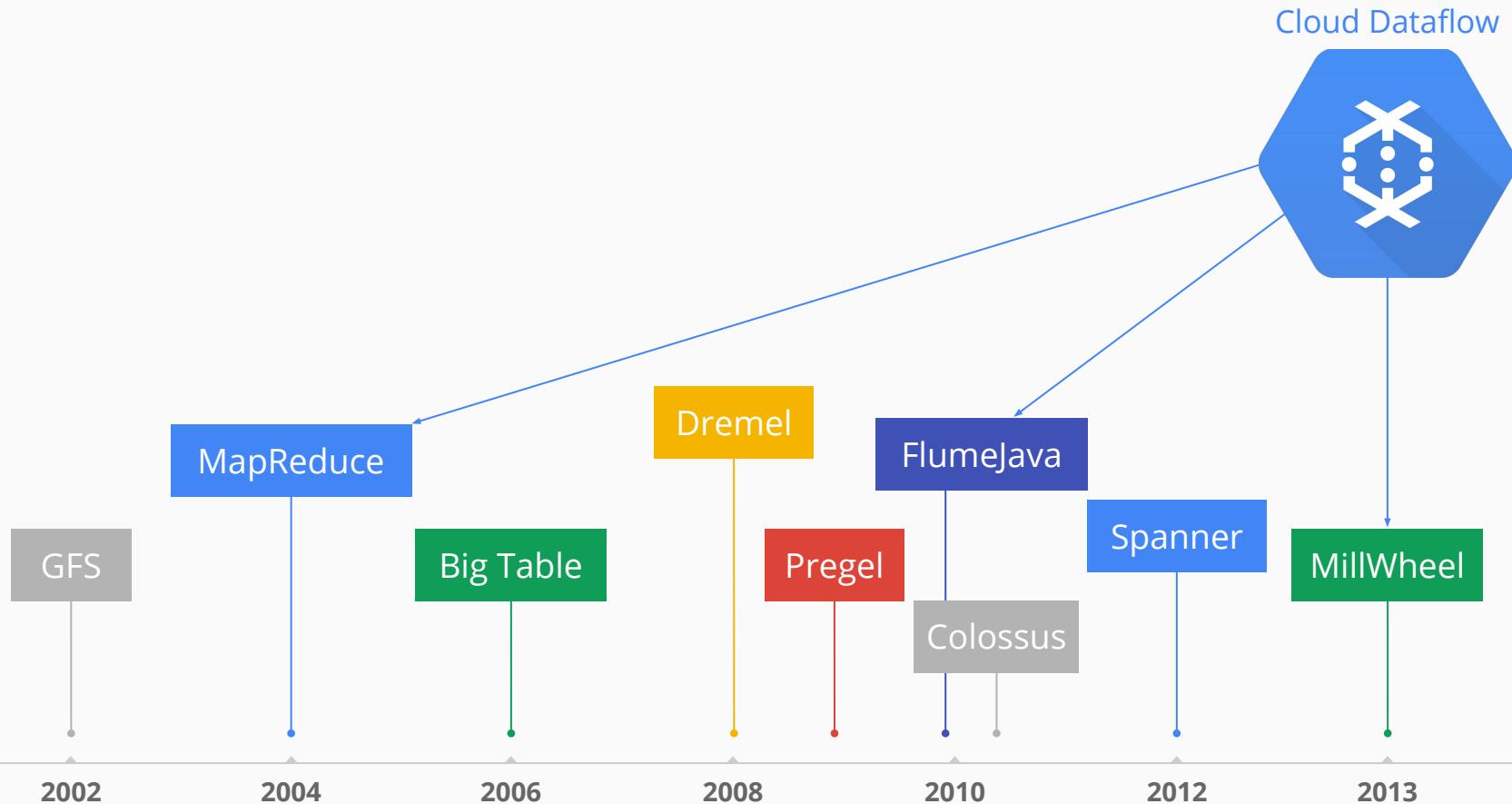
Updated Oct 31, 2019



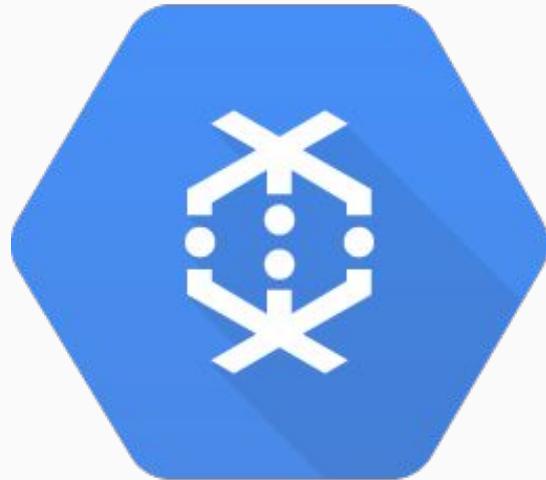
Start

Dataflow

Dataflow - Processing as a Service



Dataflow



Google Cloud
Dataflow



Dataflow is a unified programming model and a managed service for developing and executing a wide range of data processing patterns including ETL, batch computation, and continuous computation.



Cloud Dataflow frees you from operational tasks like resource management and performance optimization.



Servicio completamente gestionado y modelo de programación para el proceso de Big Data

- Gestión de Recursos integrado.
- A demanda.
- Ejecución de los trabajos inteligente.
- Auto escalado.
- Modelo de programación unificado.
- Open Source.
- Monitoraje.
- Integración.
- Procesado confiable y consistente.

Dataflow: Managing Massive Data In/Out



ETL

- Movement
- Filtering
- Enrichment
- Shaping



Analysis

- Reduction
- Batch computation
- Continuous computation



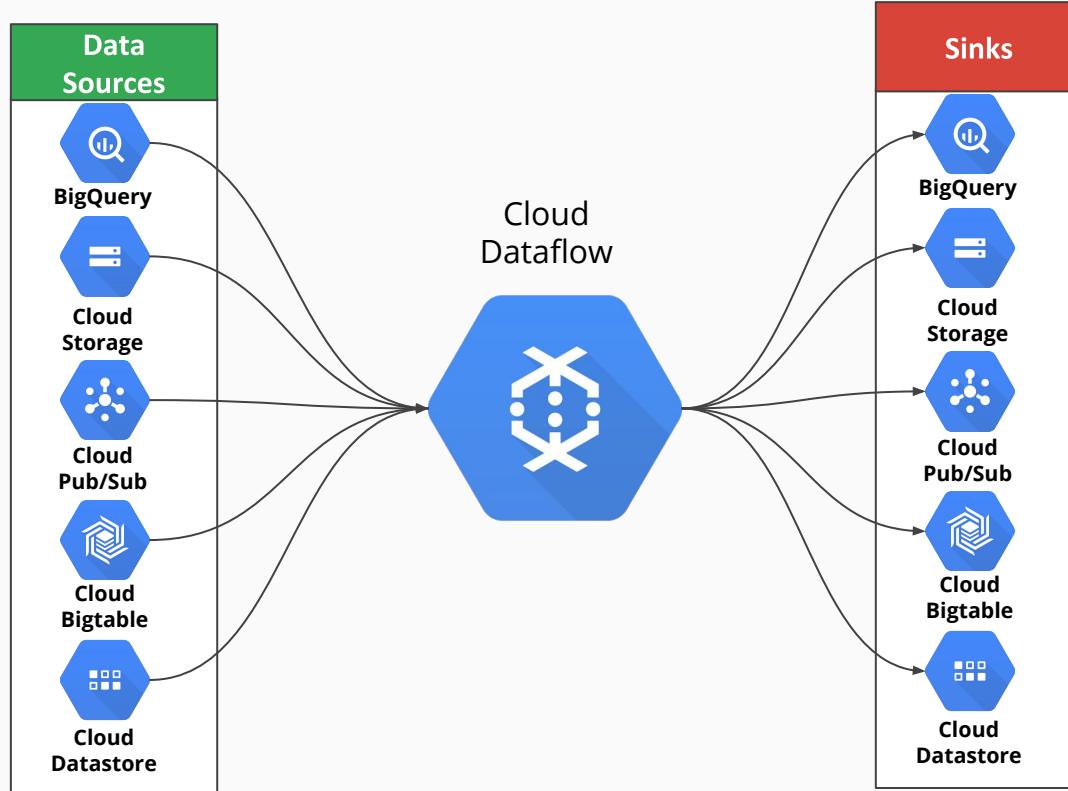
Orchestration

- Composition
- External orchestration
- Simulation

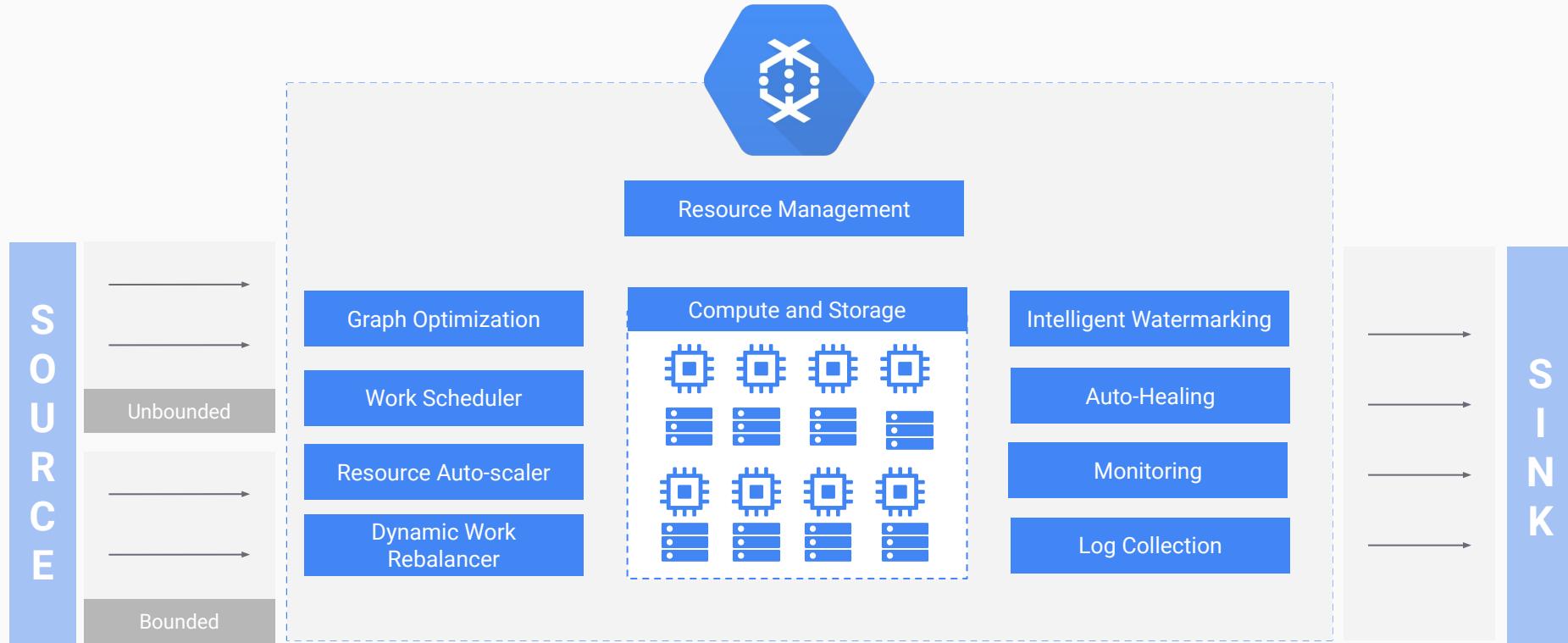
What is Cloud Dataflow?

- Google Cloud Dataflow is a tool for developing and executing a wide range of data processing patterns—for example [extract, transform, and load](#) (ETL)—on very large data sets.
- Use Cloud Dataflow for nearly any kind of data processing task, encompassing both [batch](#) and [streaming data](#) processing.
- Dataflow can handle an unbounded or “infinite” data set from a continuously updating source such as [Google Cloud Pub/Sub](#). That is, Dataflow can process practically any amount of data arriving at any time.
- Dataflow is particularly useful for [embarrassingly parallel](#) data processing tasks, in which the problem can be decomposed into many smaller bundles of data that can be processed independently, making it very fast (in the same way [MapReduce](#) works).

Data Sources and Sinks for Dataflow

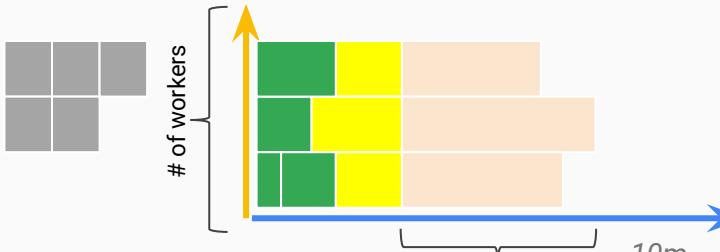


Cloud Dataflow: Under the Hood

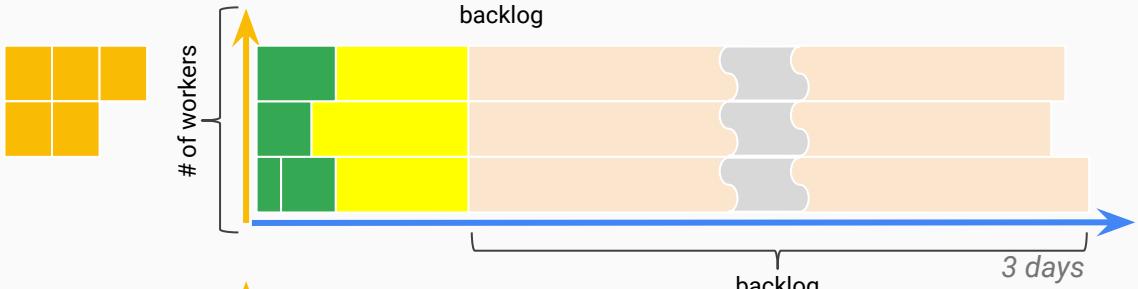


Cloud Dataflow: Autoscaling

Start off with **3** workers,
things are looking okay



Re-estimation shows there's
orders of magnitude more work:
need **100** workers!



You have 100 workers
but you don't have 100 pieces of work!
...and that's really the most important part



Cloud Dataflow: why do people use it?

Usually **the goal** of big data efforts is to reduce the time required to answer questions for making faster, better decisions.

Examples of questions prospects want answered:

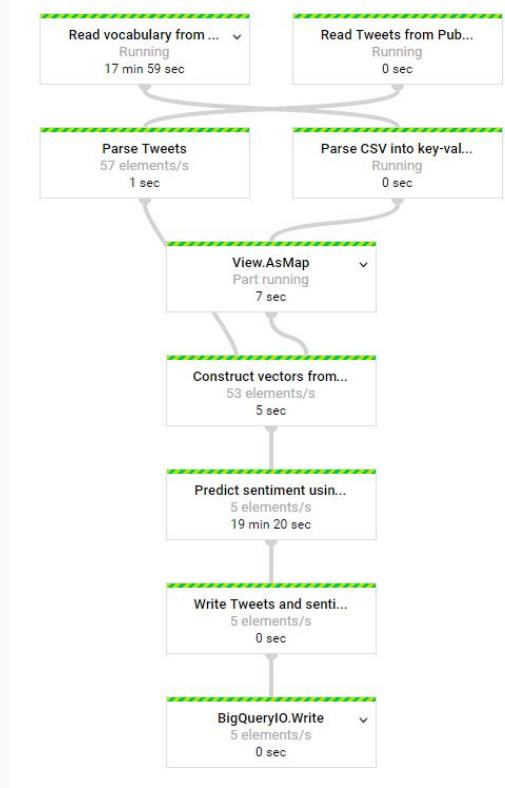
- “How many online sales did I make in the last hour due to advertising conversion?”
- “What was the average viewing time over the past seven days compared to last year?”
- “Which version of my web page do people like better?”
- “Which transactions look fraudulent?”

Cloud Dataflow **speeds up the rate at which questions can be answered** by:

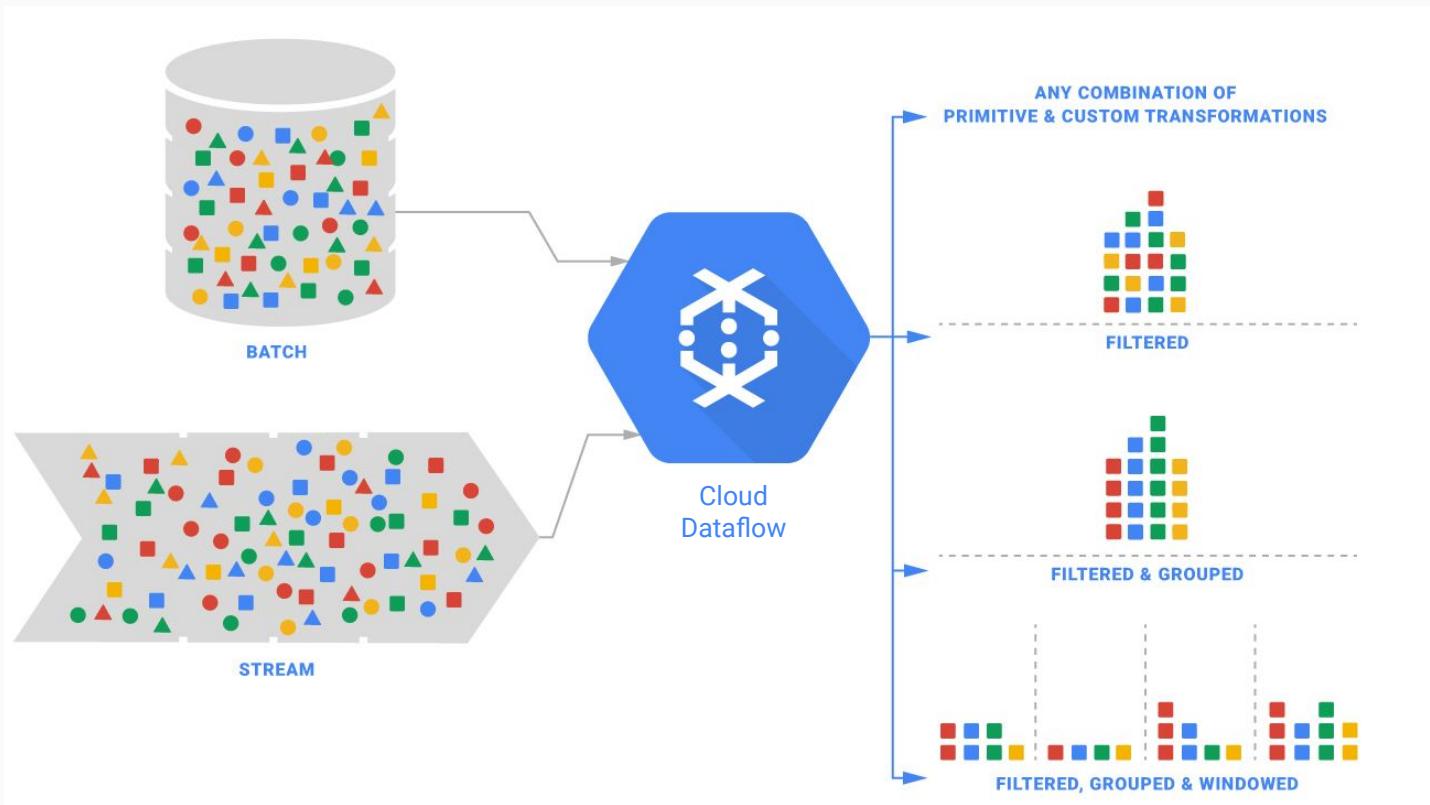
- Integrating data from multiple sources and preparing it for analysis.
- Analyzing event data streams using the Dataflow service.

How Cloud Dataflow works ?

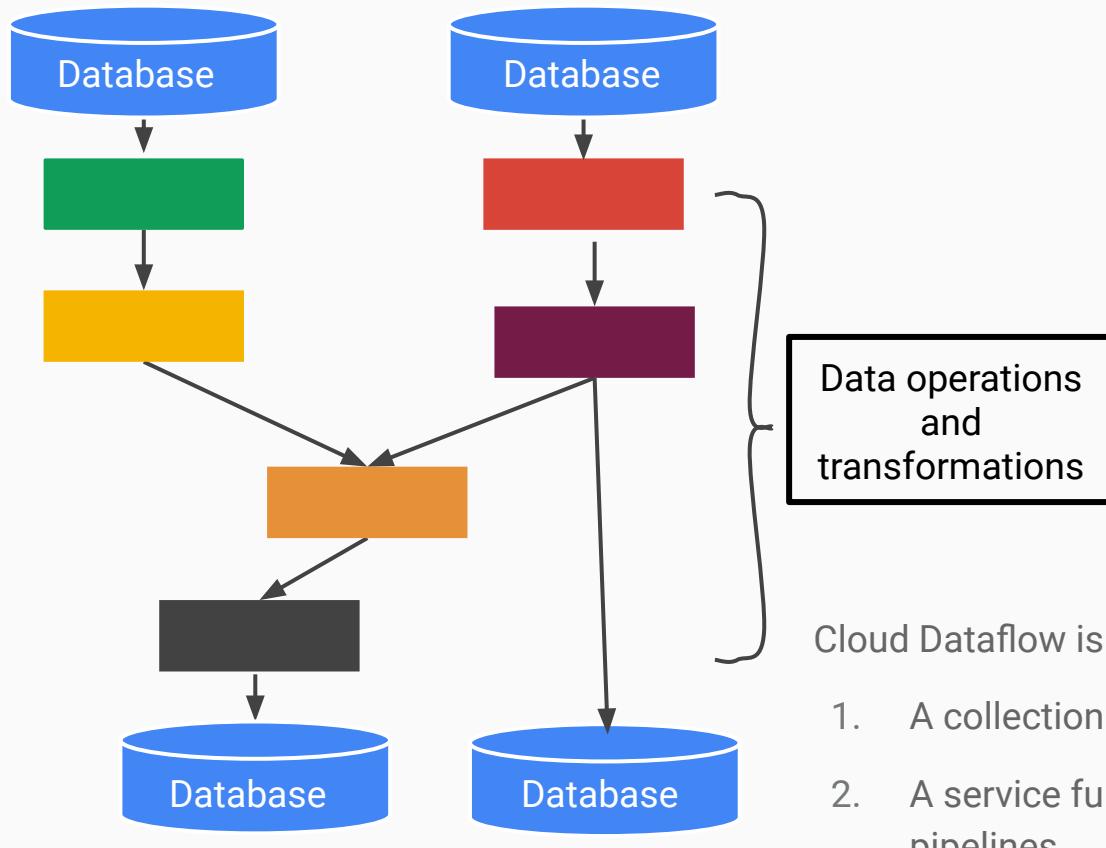
- Dataflow provides an easy way to create your data processing jobs. Each job is represented by a data processing **pipeline** that you create by writing a program.
- Each pipeline reads some input data, performs some transforms on that data to gain useful or actionable intelligence about it, and produces some resulting output data.
- A pipeline's transforms might include filtering, grouping, comparing, or joining data.



How Cloud Dataflow works ?



Dataflow: What are data pipelines?



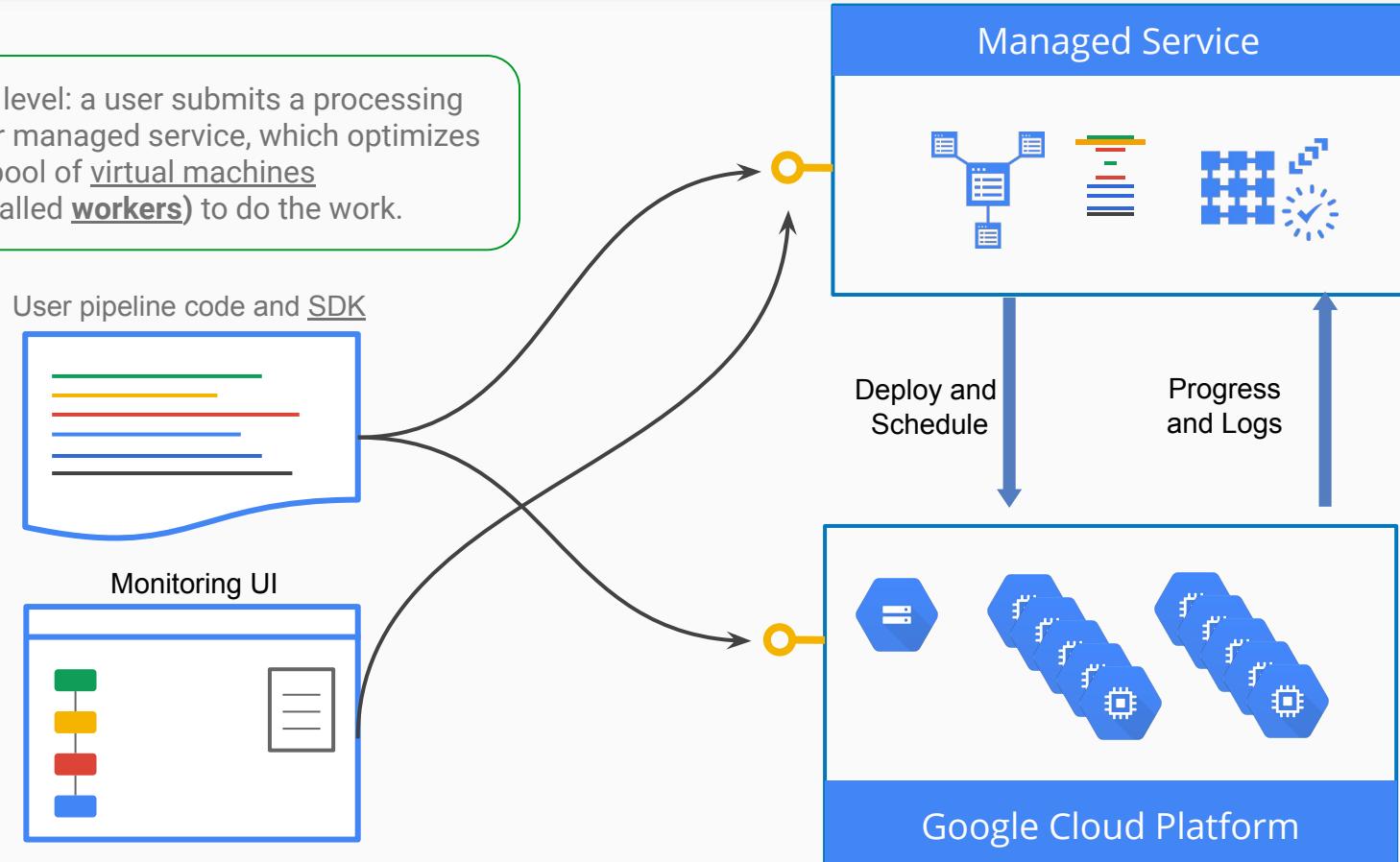
- A pipeline is a defined set of data processing transformations
- Optimized and executed as a unit
- Can include multiple inputs and multiple outputs
- Can perform many mathematical, logical, or transformation operations
- [PCollections](#) conceptually flows through the pipeline

Cloud Dataflow is implemented as two distinct elements:

1. A collection of software development kits.
2. A service fully managed by Google for running data pipelines.

Dataflow: Life of a Pipeline

At a very high level: a user submits a processing pipeline to our managed service, which optimizes it and runs a pool of virtual machines (sometimes called workers) to do the work.



Dataflow Python code

```
40  def run(argv=None):
41      """Main entry point; defines and runs the pipeline."""
42
43      parser = argparse.ArgumentParser()
44      parser.add_argument('--input',
45                          dest='input',
46                          default='gs://table/table.csv',
47                          help='Input file to process.')
48
49      parser.add_argument('--output',
50                          dest='output',
51                          default='project:dataset.table01',
52                          help='Output BigQuery table for results specified as: PROJECT:DATASET.TABLE')
53      known_args, pipeline_args = parser.parse_known_args(argv)
54
55      # Run the pipeline (all operations are deferred until run() is called).
56
57      p = beam.Pipeline(options=PipelineOptions(pipeline_args))
58
59      (p
60       # Read the text file[pattern] into a PCollection.
61       | 'Read from a File' >> beam.io.ReadFromText(known_args.input,skip_header_lines=1)
62       | 'String To BigQuery Row' >> beam.ParDo(FormatDoFn())
63       # Write the output using a "Write" transform
64       | 'Write to BigQuery' >> beam.io.WriteToBigQuery(
65           known_args.output,
66           schema=TABLE_SCHEMA,
67           create_disposition=beam.io.BigQueryDisposition.CREATE_IF_NEEDED,
68           write_disposition=beam.io.BigQueryDisposition.WRITE_APPEND)
69
70      # Run the pipeline (all operations are deferred until run() is called).
71      )
72      p.run()
73
```

Dataflow technologies that customers value

Automated input “sharding”

Removing the requirement to pre-sort input data

Unified programming model

Developers can express the computation needs regardless of batch or streaming data input

Dynamic work rebalancing

Removing the investment in manually balancing the load between resources

Logical monitoring

Provides developers a logical view of pipeline behavior vs. a control view

Auto-scaling of work resources

Removing the investment in manually scaling resources to match needs

On-demand resourcing

All resources are provided on demand, providing nearly limitless resource scale

Why do customers value Dataflow?

Less overhead-one system

Dataflow reduces operational tuning and operation management overhead found in traditional batch or streaming processing systems. Dataflow fully manages resources in Google Cloud Platform on a per-job basis, including spinning up and shutting down workers and accessing Google Cloud Storage buckets for both I/O and temporary file staging.

Price vs. performance

Dataflow provides intelligent optimization (without developer intervention) of resources to provide optimum price and performance.

Dataflow solves this pain points

Lack of existing ETL solution for BigQuery

Attempting to adopt or extend a BigQuery implementation and they need a reliable, scalable way to move, cleanse, and load data into BigQuery.

Dataflow provides optimized ETL for BigQuery.

Hadoop MapReduce cluster at capacity

Have existing investment in on-premise or on-cloud Hadoop, but they're at capacity and can't or don't want to make further investments in hardware or development.

Dataflow provides on demand and nearly limitless elasticity to offload or build new workflows outside of Hadoop.

Struggling to launch streaming Spark cluster

Trying to get a real-time streaming processing application spun up.

Struggling with overhead of self deploying and managing a Spark cluster.

Dataflow provides a unified batch and streaming model that is fully managed.

Developers focus on development, not on operations.

Migrate existing MapReduce

Would like to migrate existing MapReduce or generic batch process to a real-time stream processing model.

Want to take a batch job; for example, process a log file every 24 hours—and migrate it to process a continuous stream of data and report on that data every 5 minutes.

Dataflow solves this pain points

Too much cost with Hadoop cluster

Typically will have 1 (if not 2) dedicated headcount just for cluster management. Dataflow effectively removes this cost, since the service is fully managed.

“Burstiness” of data rates or processing needs

Existing systems; for example, Hadoop cannot dynamically shift resources to meet incoming data rate changes.

This pain point is exacerbated by the fact that their incoming data rates are growing beyond their control, or the business is asking questions faster than they can respond.

Needs large-scale computation to develop metrics for management dashboards

Traditionally, these metrics would be built after data was loaded into their SQL database. This uses more resources, gets slower as data grows, and is troubled by data contention.

Dataflow provides computation language primitives, which can be run as a combination with traditional ETL processing.

Managing two systems: 1 - batch, 1- streaming

More advanced customers may have already implemented both a batch and streaming solution; however, they are then left with managing two different systems.

Dataflow provides a **unified model** for batch and stream processing.

Dataflow use cases

Batch data movement

Moving at rest data from one system to another, such as from Google Cloud Storage to BigQuery

Data reduction and enrichment

Reduce, compress, re-shape existing data into smaller, computed values, such as log files and geo tags

Continuous computation

Analyze real-time streaming inputs, such as click streams

Continuous data movement

Real-time ETL over streaming inputs



Apache Maven is a software project management and comprehension tool. Based on the concept of a project object model (POM), Maven can manage a project's build, reporting and documentation from a central piece of information.

Recommended Workloads

WORKLOADS	CLOUD DATAPROC	CLOUD DATAFLOW
Stream processing (ETL)		✓
Batch processing (ETL)	✓	✓
Iterative processing and notebooks	✓	
Machine learning with Spark ML	✓	
Preprocessing for machine learning		✓ (with Cloud ML Engine)

Ejercicio DataFlow

Run a big data text processing pipeline in Cloud Dataflow

21 min

Updated Oct 8, 2019

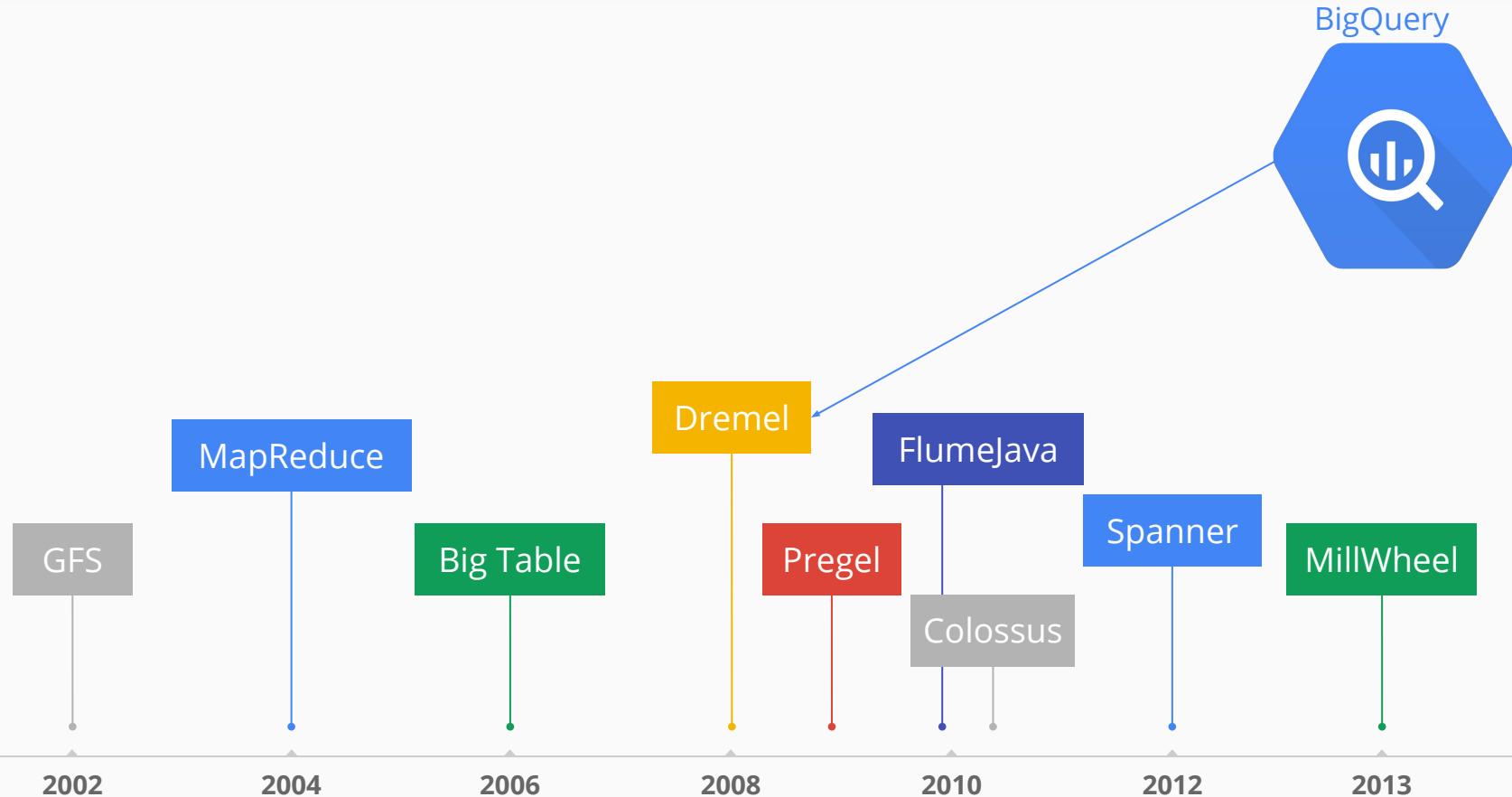


Start

BigQuery



BigQuery - Analytics as a Service



BigQuery: 100% serverless data warehouse



Google
BigQuery



Google Cloud's Enterprise Data
Warehouse for Analytics



Petabyte-Scale and Fast
Convenience of SQL



Encrypted, Durable and Highly
Available



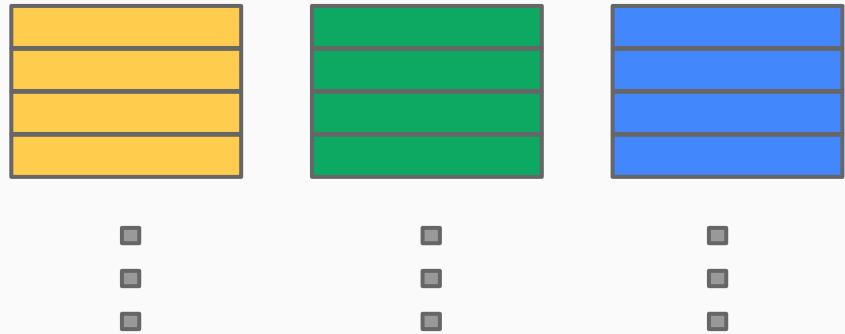
Fully Managed and Serverless



BigQuery: Columnar Storage



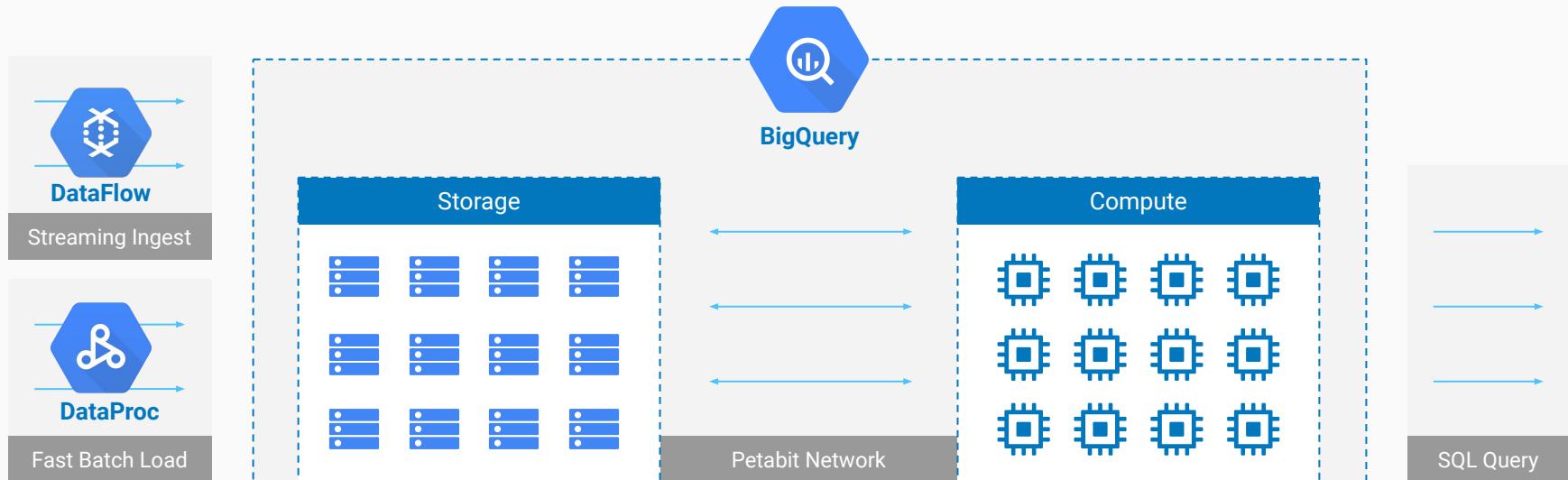
Record-Oriented Storage



Column-Oriented Storage

BigQuery: Massive Parallel Processing Engine

BigQuery = **Massively Parallel Processing** query with the petabit network and thousands of servers



BigQuery - UI



Google Cloud Platform iyv-cloud

BigQuery FEATURES & INFO SHORTCUTS + COMPOSE NEW QUERY

Query history Saved queries Job history Transfers Scheduled queries Reservations BI Engine Resources + ADD DATA Search for your tables and datas...

iyv-cloud bicing bqml_tutorial simpsons bigquery-public-data

Query editor

1

Run Save query Save view Schedule query More

usage

QUERY TABLE COPY TABLE DELETE TABLE EXPORT

Schema Details Preview

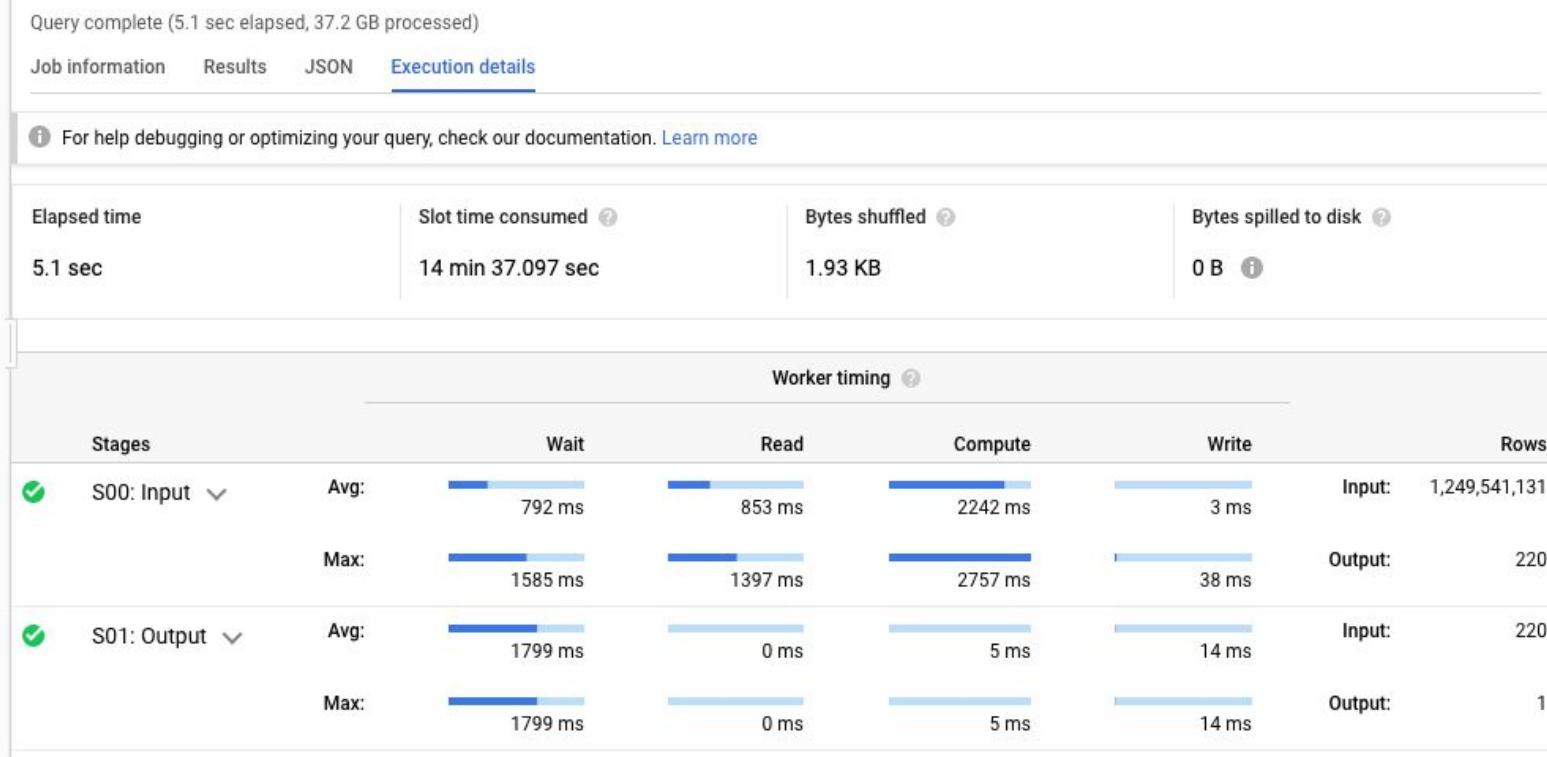
Field name	Type	Mode	Description
station_id	INTEGER	NULLABLE	
name	STRING	NULLABLE	

BigQuery - Sample Query



```
SELECT COUNT(*)  
FROM  
`bigquery-samples.wikipedia_benchmark.Wiki1B`  
WHERE REGEXP_CONTAINS(title, "G.*o.*o.*g")
```

BigQuery - Query Plan Explanation



Bigquery: Pricing



Storage

- Amount of data in table
- Ingest rate of streaming data
- Automatic discount for old data



Processing

- On-demand OR Flat-rate plans
- On-demand based on amount of data processed
- 1 TB/month free
- Have to opt-in to run [high-compute queries](#)



Free:

- Loading
- Exporting
- Queries on metadata
- Cached queries
- Queries with errors
- 10GB storage

Bigquery: Documents

Google Cloud Platform

Why Google Products Solutions Launcher Pricing Customers Documentation Support Partners CONTACT SALES

BigQuery > Documentation

Cloud BigQuery Product Overview Documentation

Quickstarts All Quickstarts Using the Web UI Using the Command-Line Tool

How-to Guides All How-to Guides Loading Data Into BigQuery Using the BigQuery Data Transfer Service Querying Data Using External Data Sources Using Views Updating Data Exporting Data Labeling Datasets Monitoring and Logging BigQuery API Basics BigQuery Web UI bq Command-Line Tool Migrating to Standard SQL

BigQuery is Google's fully managed, petabyte scale, low cost analytics data warehouse. BigQuery is NoOps—there is no infrastructure to manage and you don't need a database administrator—so you can focus on analyzing data to find meaningful insights, use familiar SQL, and take advantage of our pay-as-you-go model.

Quickstarts Learn in 5 minutes

How-to Guides Perform specific tasks

APIs & Reference API, Web UI, and Command-line

Concepts Develop a deep understanding of BigQuery

Tutorials Walkthroughs of common applications

Resources Pricing, quotas, release notes, and other resources

¿Te ha resultado útil esta página? Danos tu opinión:

ENVIAR COMENTARIOS

CONSOLA ::

Bigquery: Public Dataset

Google Cloud Platform

Why Google Products Solutions Launcher Pricing Customers Documentation Support Partners CONTACT SALES

Real-time data analysis with Kubernetes, Google Cloud Pub/Sub, and BigQuery
Real-time logs analysis using Fluentd and BigQuery
Analyzing Financial Time Series using BigQuery

Resources All Resources
Pricing and Quotas
Release Notes
Support
Public Datasets

Visión general
1000 Cannabis Genomes Project
Bay Area Bike Share Trips Data
Chicago Crime Data
Chicago Taxi Trips
EPA Historical Air Quality Data
GDELT Book Corpus
GitHub Data
Hacker News
Healthcare Common Procedure Coding System (HCPCS) Level II
IRS 990 Data
Major League Baseball
Medicare
NOAA GHCN Weather
NOAA GSOD Weather
NOAA ICOADS
NYC 311 Service Requests
NYC Citi Bike Trips
NYC TLC Trips
NYC Tree Census
NYPD Motor Vehicle Collisions
OpenAQ: Real-time Air Quality Data
Open Images

BigQuery > Documentation

Google BigQuery Public Datasets

A public dataset is any dataset that is stored in BigQuery and made available to the general public. This page lists a special group of public datasets that Google BigQuery hosts for you to access and integrate into your applications. Google pays for the storage of these data sets and provides public access to the data via BigQuery. You pay only for the queries that you perform on the data (the first 1 TB per month is free, subject to [query pricing details](#)).

Public datasets hosted by BigQuery

1000 Cannabis Genomes Project
Genomic open dataset of approximately 850 strains of Cannabis via the Open Cannabis Project.

Bay Area Bike Share Trips
This data includes all Bay Area Bike Share trips from August 2013 to the present, and is updated daily.

Chicago Crime Data
This dataset reflects reported incidents of crime that occurred in the City of Chicago from 2001 to the present.

Chicago Taxi Trips
This dataset includes taxi trips from 2013 to the present, reported to the City of Chicago in its role as a regulatory agency.

EPA Historical Air Quality Data
This dataset is provided by the EPA and includes 16 measurements of air quality in the United States from 1990 to the present.

GDELT Book Corpus
A dataset that contains 3.5 million digitized books stretching back two centuries, encompassing the complete English-language public domain collections of the Internet Archive (1.3M volumes) and HathiTrust (2.2 million

STAR RATING: ★ ★ ★ ★ ★ ENVIAR COMENTARIOS

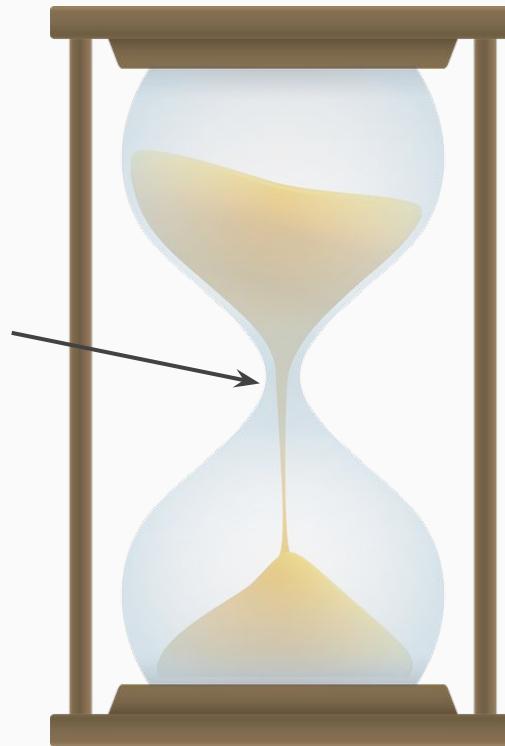
Contenido
Public datasets hosted by BigQuery
1000 Cannabis Genomes Project
Bay Area Bike Share Trips
Chicago Crime Data
Chicago Taxi Trips
EPA Historical Air Quality Data
GDELT Book Corpus
GitHub Data
Hacker News
Healthcare Common Procedure Coding System (HCPCS) Level II
IRS Form 990 Data
Major League Baseball Data
Medicare Data
NOAA GHCN
NOAA GSOD
NOAA International Comprehensive Ocean-Atmosphere Data Set
NYC 311 Service Requests
NYC Citi Bike Trips
NYC TLC Trips
NYC Tree Census
NYPD Motor Vehicle Collisions
OpenAQ: Real-time Air Quality Data
Open Images Data
RxNorm
San Francisco 311 Service Requests Data
San Francisco Fire Department Service Calls Data

Bigquery: Best Practises - Slot Estimation

Slots Available

*More Slots = Faster
Depletion Rate*

*More Concurrent Queries
= Slower Depletion Rate*

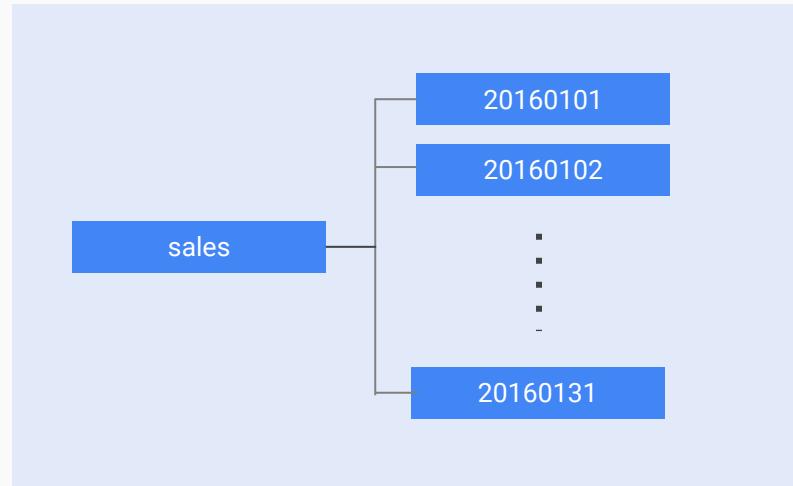


Guideline: It is recommended to plan for 2,000 BigQuery slots for every 50 medium-complexity queries simultaneously

50 queries = 2,000 Slots

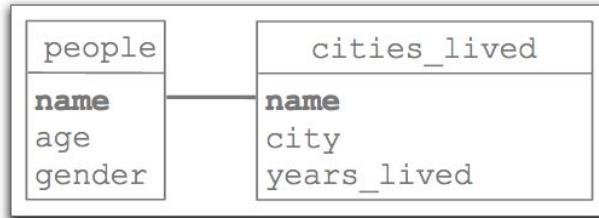
Bigquery: Best Practises - Use time partition

```
SELECT ...  
FROM sales  
WHERE _PARTITIONTIME  
BETWEEN TIMESTAMP("20160101")  
AND TIMESTAMP("20160131")
```

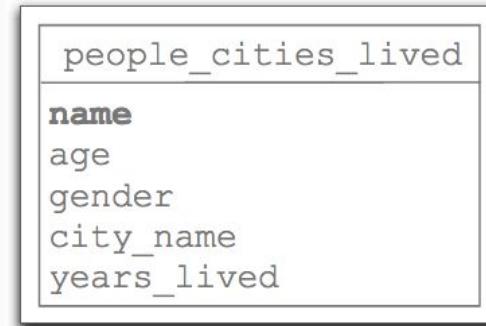


Bigquery: Best Practises - Denormalization

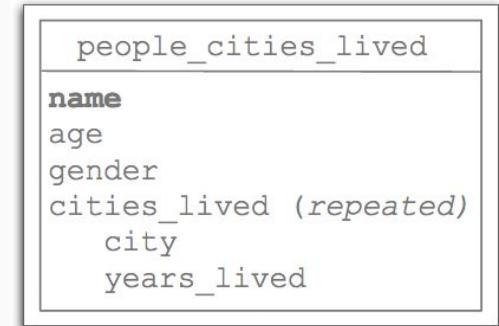
Relational Database Design



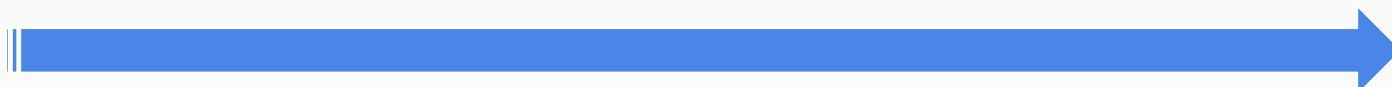
BigQuery (Object) Design



Flattened (CSV)



Nested/Repeated (JSON)



Less Performant

High Performing

Bigquery: Best Practises - Nested Schemas

- **Avoid costly joins**
- **No performance punishment for SELECT (DISTINCT Company ID)**

Organization Details with Nested Historical Transactions			
Company ID	Company Name	Transactions.Amount	Code.Expense
161218560	NY Association Inc.	\$10,000	Lobbying
		\$5,000	Legal
		\$1,000	Insurance
123435560	ACME Co.	\$7,000	Travel

Ejercicio BigQuery

Inicio

Query GitHub data using BigQuery

17 min

Updated Oct 8, 2019



Start

Load and Analyze Data in BigQuery

Updated Nov 8, 2019

 [Start](#)

Query the Wikipedia dataset in BigQuery

27 min

Updated Nov 1, 2019

[Start](#)

- 1 Create a Dataset (EU located)
- 2 Create a Table with 2 fields, Name (String) and Value (Integer)
- 3 Upload a CSV file for Netflix TV Shows

Dataprep



Cloud Dataprep: self-service data preparation



Google Cloud
Dataprep

- ✓ Visually explore and interact with data
- ✓ Instantly understand data distribution and patterns
- ✓ Quickly identify data quality issues
- ✓ Get automatic data transformation suggestions
- ✓ Standardize, structure and join datasets easily with a guided approach

Dataprep



Instant Data Exploration

Visually explore and interact with data in seconds. Instantly understand data distribution and patterns



Intelligent Data Cleansing

Automatically identifies data anomalies. Get data transformation suggestions based on your usage pattern.



Serverless

Cloud Dataprep is a serverless service, so you do not need to create or manage infrastructure.



Supports Common Data Sources of Any Size

Process diverse datasets - structured and unstructured. Transform data stored in CSV, JSON, or relational table formats. Prepare datasets of any size, megabytes to terabytes



Powerful

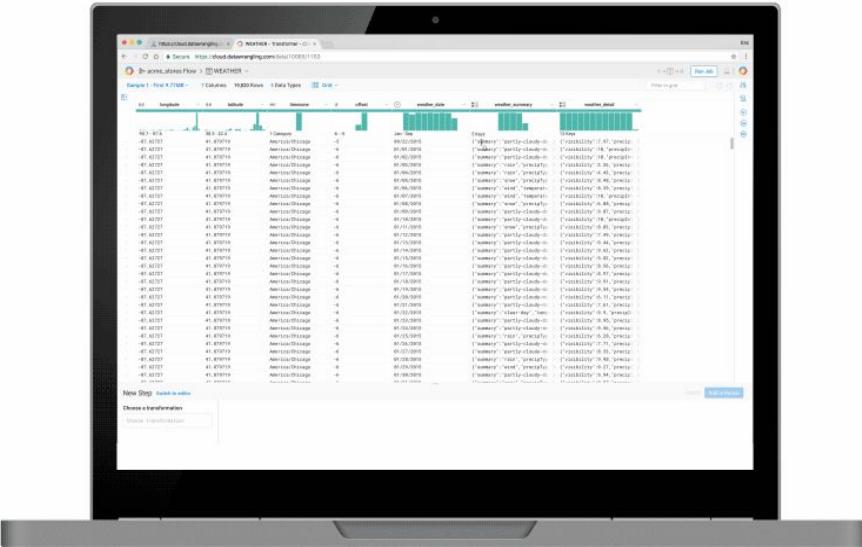
Cloud Dataprep is built on top of the powerful Google Cloud Dataflow service. Cloud Dataprep is auto-scalable and can easily handle processing massive data sets.



Integrated with Google Cloud Platform

Easily process data stored in Google Cloud Storage, Google BigQuery or from your desktop. Export clean data directly into BigQuery for further analysis.





Easy data preparation with Google Cloud Dataprep

[Google Cloud Dataprep](#) (Private Beta) is a new managed data service, built [in collaboration with Trifacta](#), that enables analysts and data scientists to visually explore and prepare data for analysis in seconds via the following features:

- **Anomaly detection:** Cloud Dataprep automatically detects schema, type, distributions and missing/mismatched values. It utilizes machine learning to suggest corrective data transformations.
- **Drag-and-drop development:** An intuitive user experience eliminates the need for coding so that users can focus on analysis.
- **Out-of-box integration with GCP:** Users can more securely read raw data from [Google Cloud Storage](#) or BigQuery, or upload from their local machine and write back cleaned data into BigQuery for further analysis.
- **Fully managed infrastructure:** Like our other managed services, IT resource provisioning and management are handled for you automatically and elastically.

Dataprep

B Marketing Ads Data > advertising_display_clicks

Grid Columns Full Dataset - 5.29MB 11 Columns 30,541 Rows 5 Data Types Run Job

Filter in grid

ser_id	event_time	url	src	device_type
series	Feb 2005 - Jun 2015	30,292 Categories	4 Categories	smartphone
8000-399	2013-11-11 15:38:00	http://seattletimes.com/turpis/a/pede/posuere/nonummy/integer/non.xml?pid=1365984800000-506	smartphone	smartphone
8000-29	2014-01-24 02:15:00	https://cafepress.com/vivamus.json?pid=1319500000000-208	smartphone	smartphone
8000-178	2010-04-19 19:09:00	http://mapquest.com/ugue/quam/sollicitudin/vitae/consectetuer/eget/rutrum.jsp?pid=1287094400000-560	computer	computer
8000-157	2012-05-24 06:40:00	http://bravesites.com/dapibus/duis/at/velit/eu.png?pid=1125878400000-283	computer	computer
8000-377	2006-01-21 04:28:00	http://google.co.jp/rhonus/dui/vel/sem/sed.html?pid=1121644800000-186	smartphone	smartphone
8000-56	2012-09-25 23:00:00	http://tripod.com/adipiscing/lorem/vitae/mattis/nibh/ligula/nec.jpg?pid=1220558480000-511	smartphone	smartphone
8000-219	2015-06-25 10:40:00	https://wmj.com/sem/laus/morbi/sem/mauris.png?pid=1253836800000-46	smartphone	smartphone
8000-55	2007-03-17 18:26:00	http://newsolutions.com/aliquam/ugue/quon/sollicitudin/vitae.jpg?pid=1150675200000-675	other	other
8000-386	2012-07-07 19:04:00	http://wfoo.com/sit/ amet.aspx?pid=1322697600000-250	other	other
8000-286	2006-11-08 14:34:00	https://theglobeandmail.com/sem/accumsan/felis/ut.json?pid=1157241600000-568	smartphone	smartphone
8000-498	2012-01-17 02:28:00	http://cam.ac.uk/neque/vestibulum/eget/vulputate.aspx?pid=1109721600000-237	other	other
8000-153	2011-12-16 28:38:00	https://odnoklassniki.ru/maecenas.jpg?pid=1281312000000-61	other	other
8000-583	2006-11-08 28:49:00	https://ft.com/hac.png?pid=1140393680000-91	tablet	tablet
8000-484	2012-03-22 18:33:00	http://blogspot.com/sem.aspx?pid=1204934400000-354	computer	computer
8000-446	2006-01-27 14:47:00	http://redcross.org/ut/suscipit/a.xml?pid=1105747200000-171	smartphone	smartphone
8000-268	2014-06-24 02:45:00	http://amazon.com/in.json?pid=1121644000000-33	computer	computer

New Step Switch to editor

Choose a transformation

Choose a transformation

Cancel Add to Recipe

Dataprep

Cloud Dataprep Deployment-Demo FLOWS DATASETS JOBS

Marketing Ads Data

```
graph LR; sales[sales_scaled] --> ad1[advertising_display_clicks]; sales --> ad2[advertising_display_clicks]; ad1 --> sales2[sales_scaled - 2]; ad2 --> sales2;
```

Details

advertising_display_clicks

Edit Recipe Run Job See Recipe

ABC event_type	ABC event_subtype	# advertiser_id
view	view	164332
view	view	164332
click	click	164332
view	view	164332
click	click	164332
view	view	164332
click	click	164332

Size 12 columns · 5 types
Used in 1 Flow
Run 0 Jobs [More details](#)
Updated Today at 1:34 PM
Created Today at 1:25 PM

Ejercicio Dataprep

Quickstart

This page shows you how to get started using the Cloud Dataprep Web application.

AI Notebooks



Google Cloud for Data Scientist 

File Edit View Insert Runtime Tools Help

 COMMENT

+ CODE + TEXT ▲ CELL ▾ CELL CONNECT ▾

Table of contents Code snippets Files X

Before you begin

Provide your credentials to the runtime

Use BigQuery via magics

Use BigQuery through google-cloud-bigquery

Declare the Cloud project ID which will be used throughout this notebook

Sample approximately 2000 random rows

Describe the sampled data

View the first 10 rows

▼ **Before you begin**

1. Use the [Cloud Resource Manager](#) to Create a Cloud Platform project if you do not already have one.
2. [Enable billing](#) for the project.
3. [Enable BigQuery](#) APIs for the project.

▼ **Provide your credentials to the runtime**

```
[ ] from google.colab import auth  
auth.authenticate_user()  
print('Authenticated')
```

→ Authenticated

AI Notebooks



- ✓ For data scientists
- ✓ Notebook interface
- ✓ Leverage existing Jupyter modules and knowledge
- ✓ Suitable to interactive data science and machine learning
- ✓ Closely integrated with BigQuery and CloudML

Instancias administradas de notebooks de JupyterLab

Notebooks de AI Platform es un servicio administrado que ofrece un entorno integrado de JupyterLab en el que los desarrolladores de aprendizaje automático y los científicos de datos pueden crear, en un solo clic, instancias que ejecutan JupyterLab y tienen preinstalados los marcos de trabajo de ciencia de datos y aprendizaje automático más recientes. Notebooks está integrado en BigQuery, Cloud Dataproc y Cloud Dataflow, lo que facilita pasar de la transferencia de datos al procesamiento previo y la exploración, y, luego, el entrenamiento y la implementación de modelos.





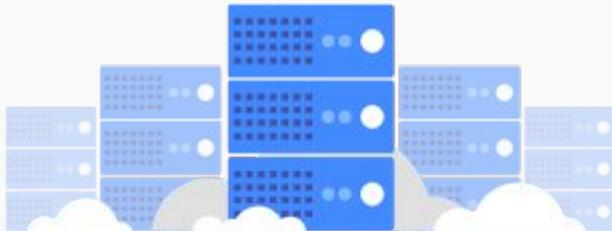
Ponte en marcha rápidamente

Puedes implementar instancias nuevas de JupyterLab con un clic y comenzar a analizar tus datos de inmediato. Cada instancia viene preconfigurada con versiones optimizadas de las bibliotecas más populares de ciencia de datos y aprendizaje automático, de modo que no tengas que preocuparte por crear y administrar las VM.

Usa frameworks de trabajo populares de código abierto

Notebooks no tiene una curva de aprendizaje, ya que usa la interfaz estándar de la industria de JupyterLab y tiene preinstaladas versiones optimizadas de bibliotecas populares, como TensorFlow, PyTorch, scikit-learn, Pandas, NumPy, SciPy y Matplotlib.





Scale on Demand

Puedes comenzar en un nivel bajo y escalar verticalmente con solo agregar CPU, RAM y GPU. Cuando tengas demasiados datos para una sola máquina, podrás cambiar sin interrupciones a servicios distribuidos como BigQuery, Cloud Dataproc, Cloud Dataflow y Entrenamiento y predicción de AI Platform.

AI Notebooks

AI Platform Notebooks te permite crear y administrar instancias de máquina virtual (VM) que vienen ya empaquetadas con [JupyterLab](#). Las instancias de AI Platform Notebooks son compatibles con los marcos de trabajo de TensorFlow y PyTorch, y tienen un conjunto ya instalado de paquetes de aprendizaje profundo de Python y R. Puedes configurar instancias solo para CPU o habilitadas para GPU a fin de optimizar tu flujo de trabajo.

AI Platform Notebooks facilita el trabajo de crear y configurar una [máquina virtual de aprendizaje profundo](#), ya que proporciona imágenes verificadas, optimizadas y probadas para el marco de trabajo elegido.

Tus instancias de notebook están protegidas con la autenticación y autorización de Google Cloud Platform (GCP) y se encuentran disponibles mediante una URL de instancia de notebook. Las instancias de notebook también se integran a [GitHub](#), de modo que puedas sincronizar tu notebook con un repositorio de GitHub sin dificultades.

Ejercicio AI Notebooks

Analyzing Natality Data using AI Platform Notebooks and BigQuery

14 min

Updated Oct 16, 2019



Start

Composer



Cloud Composer

Managed Apache Airflow to make workflow creation and management easy, powerful, and consistent.



How Cloud Composer helps you

- **Comprehensive GCP integration:** Orchestrate your full GCP pipeline through Cloud Composer
- **Hybrid and multi-cloud environments:** Connect your pipeline and break down silos through a single orchestration service
- **Easy workflow orchestration:** Get moving quickly by coding your workflows using simple Python
- **Open source at its core:** Ensure freedom from lock-in through Apache Airflow's open source portability

Key Cloud Composer features

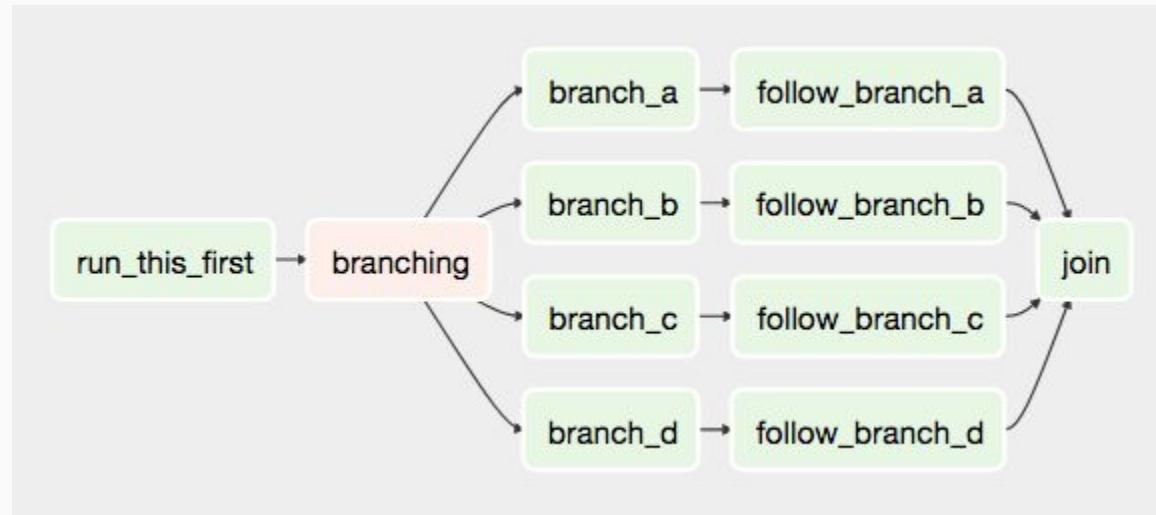
- Client tooling including the Google Developer Console and Cloud SDK
- Easy and controlled access to the Airflow web UI through Cloud Identity-Aware Proxy
- Streamlined Airflow runtime and environment configuration, such as plugin support
- Stackdriver logging and monitoring
- Identity access management (IAM)
- Simplified DAG (workflow) management
- Python (PyPi) package management

Composer - Airflow DAGs

DAG = Directed Acyclic Graph(Vertex, Edge) where

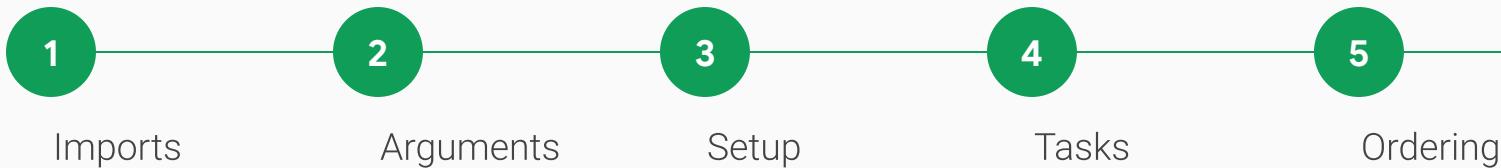
V = Task | DAG (SubDagOperator)

E = Dependency (e.g., $v1 >> v2$: $v1$ is the upstream dependency of $v2$)



Understanding Airflow DAGs

Airflow DAGs are Python files with 5 main sections



Imports

Import Python dependencies needed for the workflow.

```
import datetime
from airflow import DAG
from airflow import models
from airflow.contrib.operators import bigquery_operator
from airflow.utils import trigger_rule
```

Arguments

Define default and DAG-specific arguments. Can utilize environment variables and Jinja templating.

```
default_dag_args = {
    # Setting start date as yesterday starts the DAG
    # immediately when it is detected in the Cloud
    # Storage bucket.
    'start_date': yesterday,
    'email_on_failure': False,
    'email_on_retry': False,
    # If a task fails, retry it once after waiting
    # at least 5 minutes
    'retries': 1,
    'retry_delay': datetime.timedelta(minutes=5),
}

bq_dataset_name = 'airflow_bq_dataset_{{ ds_nodash }}'
bq_github_table_id = bq_dataset_name + '.github_commits'
output_file = 'gs://my_bucket/github_commits.csv'
```

Setup

Give the DAG a name,
configure the schedule,
and set the DAG
settings.

```
dag = DAG(  
    'demo_dag',  
    # Continue to run DAG once per day  
    schedule_interval = datetime.timedelta(days = 1),  
    default_args = default_dag_args  
)
```

Tasks

Tasks that do two things:

1. Run a BigQuery query that exports the results to a new dataset
2. Exports the new dataset to Cloud Storage

```
# Perform query of Airflow GitHub commits
bq_airflow_commits_query =
    bigquery_operator.BigQueryOperator(
        task_id = 'bq_airflow_commits_query',
        bql = """ SELECT commit, subject, message
                  FROM [bigquery-public-data:github_repos.commits]
                 WHERE repo_name contains 'airflow'
                """
        ,
        destination_dataset_table = bq_github_table_id
    )

# Export query result to Cloud Storage
export_commits_to_gcs =
    bigquery_to_gcs.BigQueryToCloudStorageOperator(
        task_id = 'export_airflow_commits_to_gcs',
        source_project_dataset_table = bq_github_table_id,
        destination_cloud_storage_uris = [output_file],
        export_format = 'CSV'
    )
```

Ordering

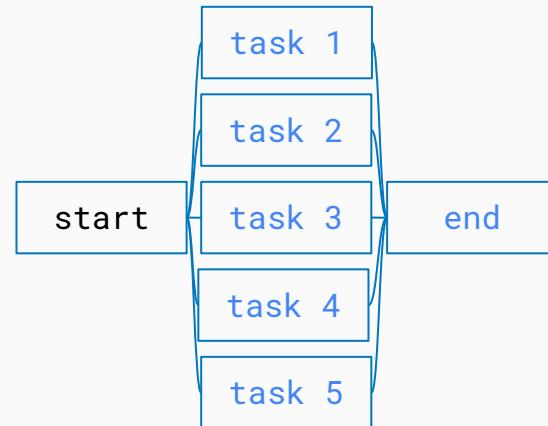
Set the order the tasks should be executed in.
Ordering can be simple (all at once, in serial) or complex (graph.)

```
# Define task ordering
bq_airflow_commits_query >> export_commits_to_gcs
```

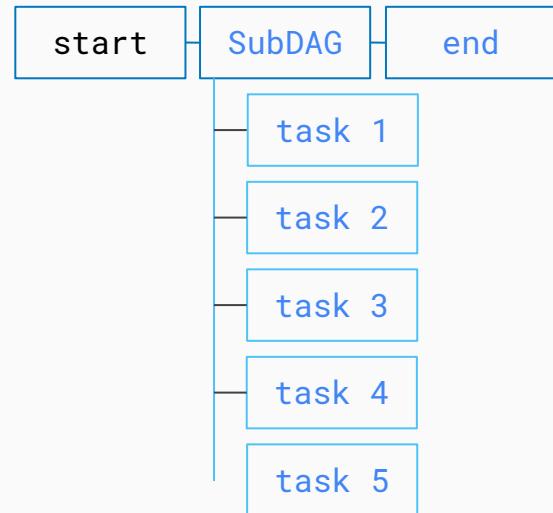
SubDAGs

SubDAGs can be used to simplify and separate DAGs into clean and logical parts.

One DAG, many tasks



One DAG with SubDAG



Composer - Steps



Composer - Operators

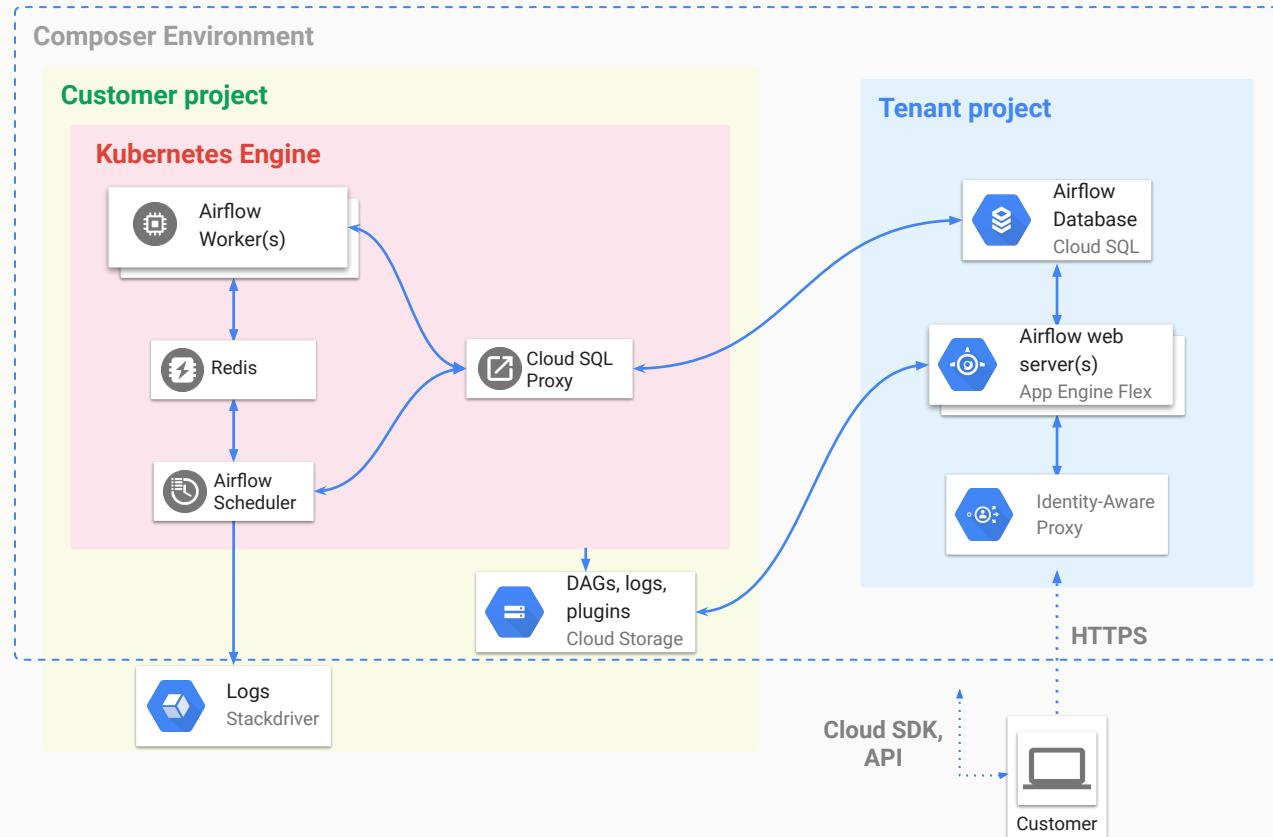
An operator describes a single task in a workflow.

Types:

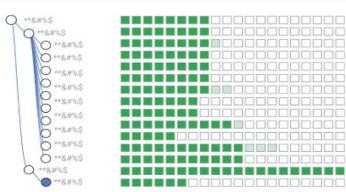
- Action operator that performs or tells the system to perform an action
- Transfer operator that moves data from one system to another
- Sensor are a special type of operators that keeps running until a certain criteria is met (e.g. availability of data).

Airflow already supports BigQuery, Dataflow, Dataproc, Datastore, Cloud Storage, Pub/Sub and Cloud ML engine.

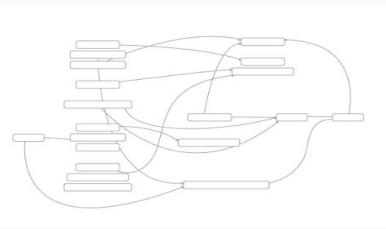
Composer - Architecture



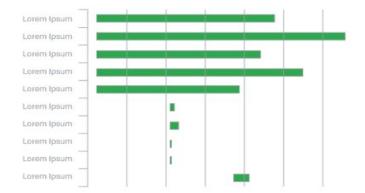
Composer - DAG Views on Airflow UI



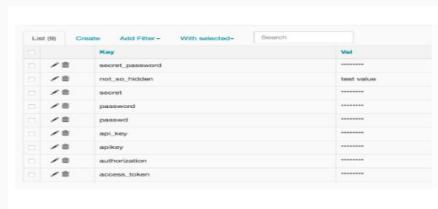
Tree views show a DAG across time, aiding in identifying blockers for late pipelines.



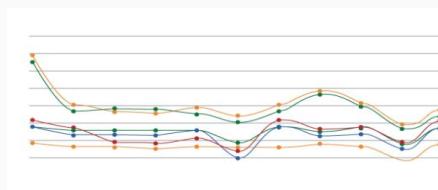
Graph views give a comprehensive look at dependencies and status of a work-flow.



Gantt views allow you to analyze task duration to spot bottlenecks



Variable views allow you to inspect the key-value pair of variables used in jobs



Gantt views allow you to analyze task duration to identify bottlenecks

0100101
11010101
0111100
10001101

Code views give you the clearest view of what is happening within your pipeline

Composer - Integrations

Cloud Storage
Unified object storage for developers and enterprises.



Cloud ML Engine
Build superior machine learning models and deploy them into production.



Cloud Pub/Sub
Ingest event streams from anywhere, at any scale, for simple, reliable, real-time stream analytics.



BigQuery
A fast, highly-scalable, cost-effective, and fully-managed enterprise data warehouse for analytics at any scale.

Cloud Composer

Cloud Dataflow
Simplified stream and batch data processing, with equal reliability and expressiveness.



Cloud Dataproc
A faster, easier, more cost-effective way to run Apache Spark and Apache Hadoop.

Composer - Connectors

Broad community of development

Diverse needs of the OSS community ensure that connectors are built to a variety of services, with more added on an ongoing basis

Easy-to-use experience

Drop do-it-yourself orchestration between services that hamper future flexibility by using established connectors

Connect data across environments

Take advantage of what you feel are best in breed solutions from a mix of your on-premises and cloud environments

Public Cloud Integration

Azure Blob Storage
AWS EMR
AWS S3
AWS EC2
AWS Redshift
Databricks SubmitRunOperator

Workflow Orchestration Cloud Composer



On-prem integration

GCP Integration

Cloud Storage
BigQuery
Cloud Dataflow
Cloud ML Engine
Cloud Dataproc
Cloud Pub/Sub
Cloud Datastore

Cloud Composer Data Workflow Orchestration

End-to-End GCP Integration

Author full workflows for GCP with a single tool, Integration with BigQuery, Dataflow, Dataproc, Datastore, Cloud Storage, Pub/Sub, Cloud ML Engine

Hybrid and Multi-Cloud Environments

Break down data silos and unite data sources in one tool, Connect data pipeline between different clouds and different on-premises tools

Easy Workflow Orchestration

Author workflows in Python, One-click deployment and auto-synchronization, Rich library of connectors, Easy troubleshooting and increased reliability through graph-based (DAG) UI

Open Source at its Core

Built upon Apache Airflow, Freedom from lock-in, Community continuing contributions for non-GCP platforms GCP contributing back our developments

Ejercicio Composer

Inicio

Intro to Cloud Composer

34 min Updated May 20, 2019

</>

Start

Data Fusion



Cloud Data Fusion



Cloud Data Fusion is a **fully-managed, cloud native, enterprise data integration service** for quickly building and managing data pipelines.

Data Fusion - CDAP



CDAP
cdap.io

Google acquired Cask -
May 14th 2018

Cloud Data Fusion is Google's [cloud native](#) manage service for [enterprise data integration](#) powered by CDAP.

CDAP is an 100% open-source framework to build data analytics applications for on-premise and cloud.



API



Runtime



Pre-built
Accelerators



Library



Apache 2.0
License



In production at
scale for +4.5
years / Dev +7
years

Developer, Data scientist and Business Analyst



Need to cleanse, match, de-dupe, blend, transform, partition, transfer, standardize, automate, & monitor

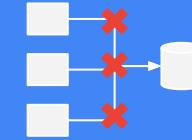
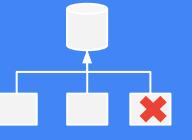
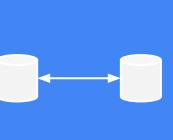


Use Cloud Data Fusion to visually build integration pipeline, test, debug and deploy



Run it at scale on GCP, operationalize (monitor, report) pipelines, inspect rich integration metadata

Data Fusion

Business Imperatives	Improve Decisions & Regulatory Compliance	Modernize Business and Reduce IT costs	Acquire and Merge	Increase Business Profitability	Extend or Migrate or both to Cloud
IT Initiatives	Business Intelligence / Analytics / AI	Legacy Retirement	Data and Application Consolidation	Customer, Supplier, Product Hubs	iPaaS, SaaS
 Data Fusion Use cases	 Data Warehouse	 Data Migration	 Data Consolidation	 Master Data Management	 Data Consistency
Data Services					

Data Fusion - Benefits



Integrate with any data



Increase productivity



Reduce complexity



Increase flexibility

The screenshot shows the Data Fusion Preparation interface. At the top, there's a blue header bar with the title "Data Fusion | Preparation" and a "DASHBOARD" button. Below the header, there's a navigation bar with a file icon and the path "sales_clean.txt" under "File System". The main area is divided into two tabs: "Data" (selected) and "Insights". The "Data" tab displays a table with 11 rows of data. The columns are labeled: Double, String, String, String, String, Integer, Long, and Integer. The "String" columns are labeled "price", "city", "zip", "type", "beds", "baths", "size", "lot_size", and "stories". The "Double" column is labeled "Double". The "Integer" column is labeled "Integer". The "Long" column is labeled "Long". The "Data" table has 11 rows, each representing a data point from the "sales_clean.txt" file.

	Double	String	String	String	String	Double	Integer	Long	Integer
	▼ price □	▼ city □	▼ zip □	▼ type □	▼ beds □	▼ baths □	▼ size □	▼ lot_size □	▼ stories □
1	1000000.0	Santa Clara	95050	Condo	2	2.5	1410	1422	3
2	1000000.0	Santa Clara	95050	Condo	3	2.5	1670	1740	2
3	1000000.0	Santa Clara	95050	Condo	3	2.5	1708	1750	2
4	1000000.0	Santa Clara	95051	Single-Family Home	3	2.0	1068	5600	1
5	1000000.0	Palo Alto	94306	Condo	2	1.5	998	499	2
6	1000000.0	Sunnyvale	94089	Single-Family Home	3	2.0	1108	5824	1
7	1000000.0	Santa Clara	95054	Single-Family Home	3	2.0	1612	6250	1
8	1000000.0	Mountain View	94040	Condo	2	2.0	1206	1880	1
9	1000000.0	Sunnyvale	94085	Condo	2	2.5	1198	1082	3
10	1000000.0	Sunnyvale	94085	Condo	3	3.5	1513	1575	3
11	1000000.0	Santa Clara	95054	Single-Family Home	3	2.0	1097	6200	1

Data Fusion - Development

Rich graphical interface

100+ plugins - connectors, transforms & actions

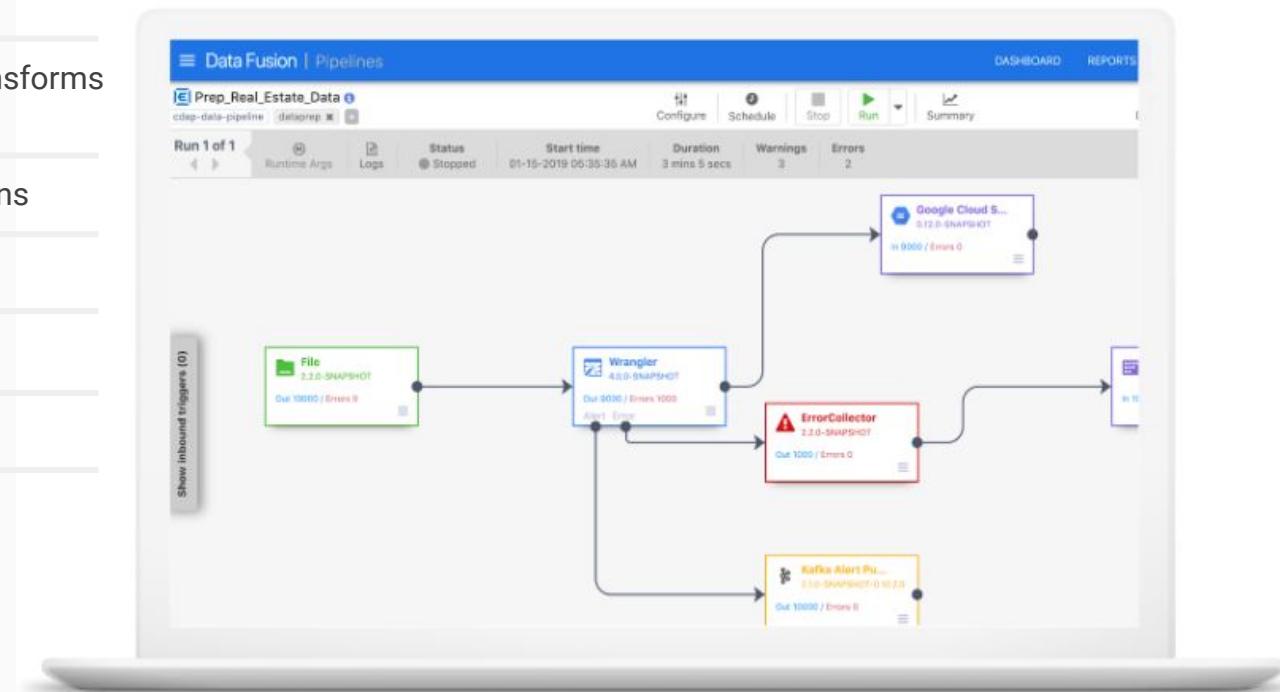
Code free visual transformations

1000+ transforms, data quality

Test and debug pipeline

Pre-built pipelines

Developer SDK



Data Fusion - Execution & Management

Cloud Dataproc - Batch & Realtime

Schedules and Triggers

Control and data flow pipelines support

Dashboard and Reports

Execute existing **Spark** and **MapReduce** jobs

Private IP support

Service Networking - VPC Peering

REST API and CLI

The screenshot shows the Data Fusion Control Center interface. At the top, there is a search bar, a filter dropdown set to "Sort by Newest", and a title "Data Fusion | Control Center". Below this, a section titled "Entities in namespace 'default'" displays a grid of entities. The entities are categorized into Application, Dataset, and Data Pipeline types. Each entity has a summary table with columns for Programs, Operations, and Writes, along with a "View Details" button.

Type	Name	Snapshot	Programs	Operations	Writes
Application	ModelManagementApp	1.0.0-SNAPSHOT	1	0	0
Dataset	experiment_model_meta	1.0.0-SNAPSHOT	1	0	0
Dataset	experiment_model_components	1.0.0-SNAPSHOT	1	0	0
Dataset	experiment_n	1.0.0-SNAPSHOT	1	0	0
Dataset	File1	1.0.0-SNAPSHOT	1	0	0
Data Pipeline	StreamExample	1.0.0-SNAPSHOT	2	0	0
Application	dataprep	1.5.0-SNAPSHOT	1	1	0
Dataset	connections	1.0.0-SNAPSHOT	1	0	0
Dataset	schemaRegistry	1.0.0-SNAPSHOT	1	0	0
Dataset	workspace	1.0.0-SNAPSHOT	1	8,917	717

Data Fusion - Integration Metadata

Tags and Properties support

Pipeline

Dataset

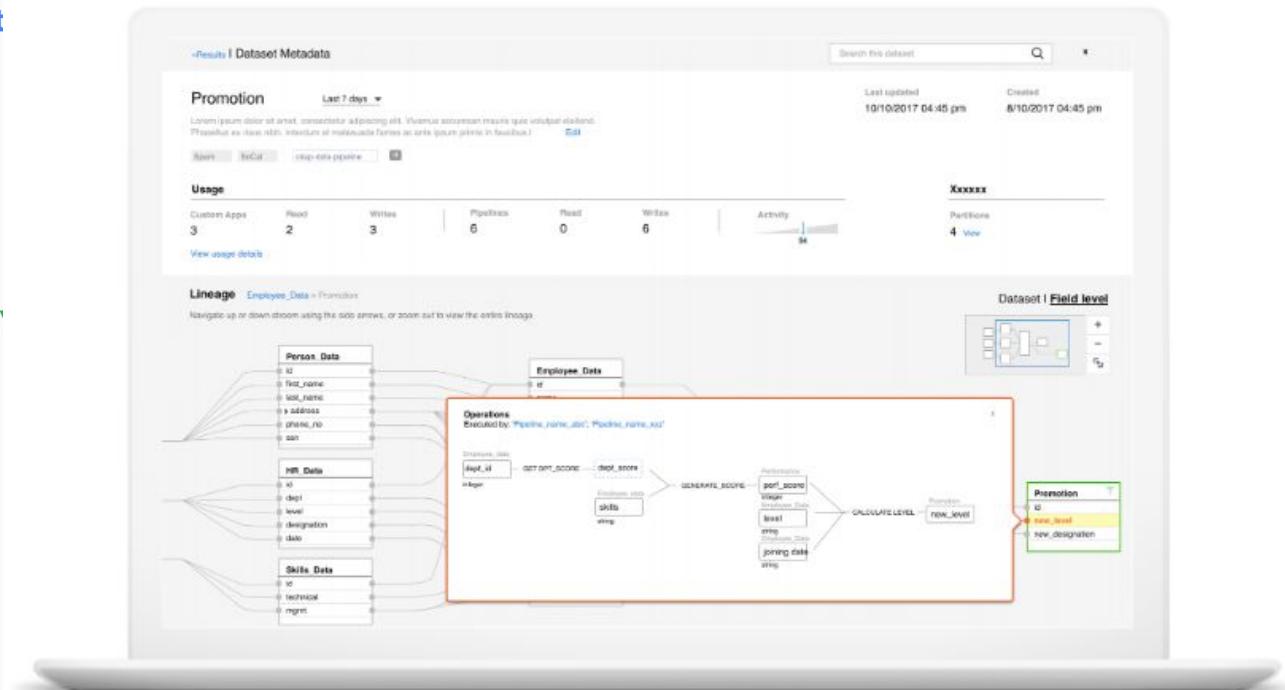
Schema

Search integrated entities by:

Keyword

Schema name and type

Dataset level and Field
level Lineage



Data Fusion - Extensible

Pipeline templatization

Conditional Pipeline Triggers

Plugin Management

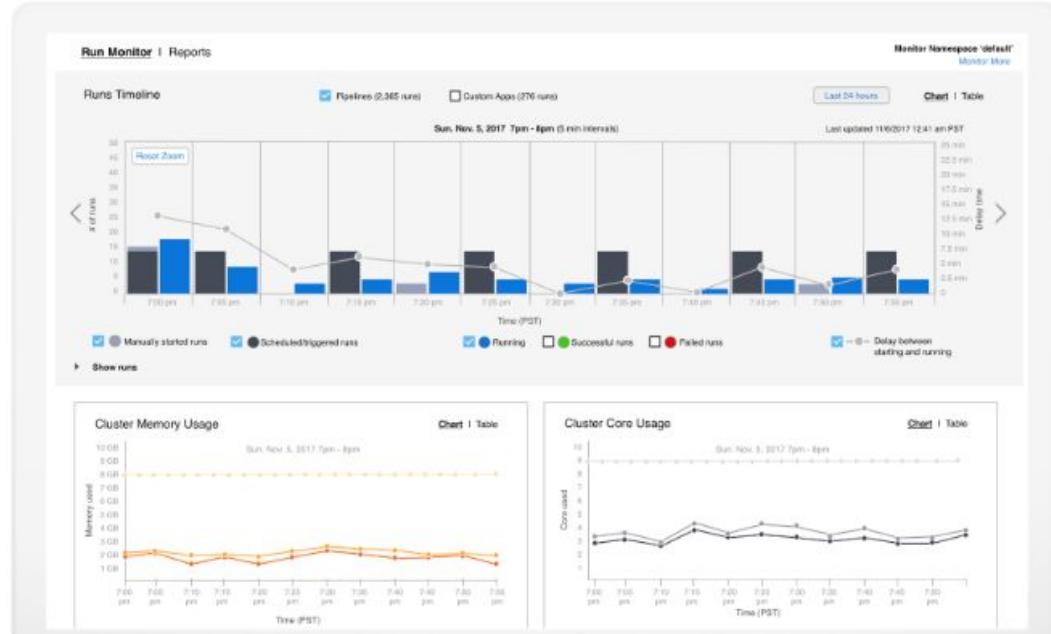
Plugin templatization

Plugin UI Widget

Custom Provisioners

Custom Compute Profiles

Hub Integration



Ejercicio Data Fusion

Ingest CSV data to BigQuery using Cloud Data Fusion - Batch ingestion

Updated Jan 21, 2020



Start

Data Catalog



Data Catalog is a **fully managed** and **highly scalable** data discovery and metadata management service



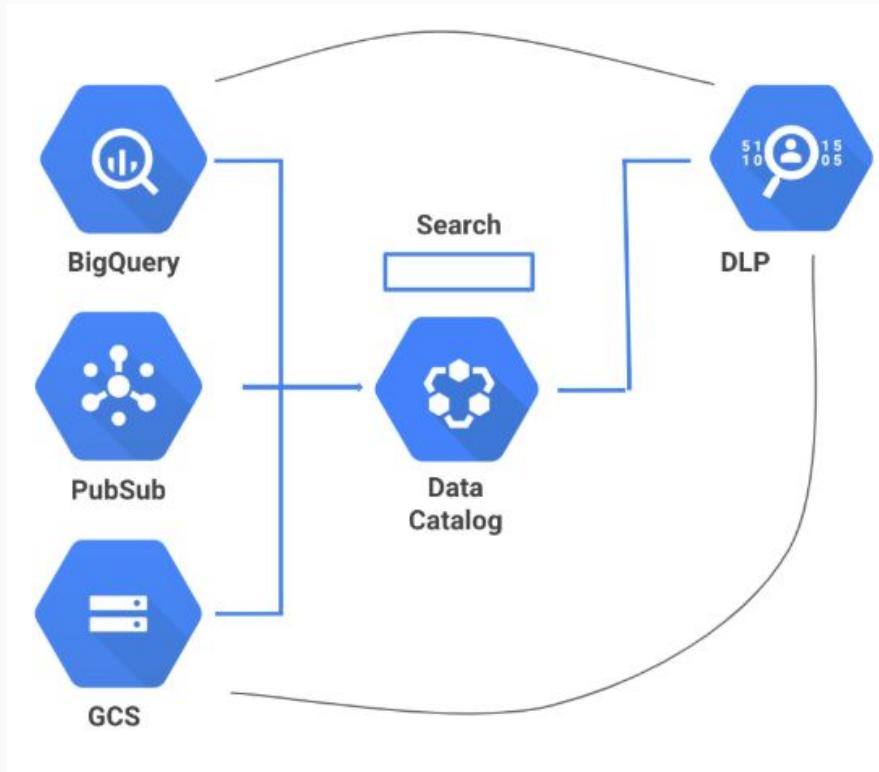
Organizations faced with a wealth of data spread across disjoint systems need an **effective solution for data discovery**

Offers **unified data discovery** of all data assets, spread across multiple projects and systems

Empowers users to **annotate business metadata** in a collaborative manner

Provides the **foundation** for data **governance**

Data Catalog feature highlights



1. Provides a simple search interface for data discovery
2. Supports UI and API for all metadata operations
3. Supports business metadata through schematized tags
4. Auto-ingests technical metadata from GCP data assets
5. Enforces ACL controls on metadata
6. Auto-tags PII data through DLP integration

Data Catalog - Architecture

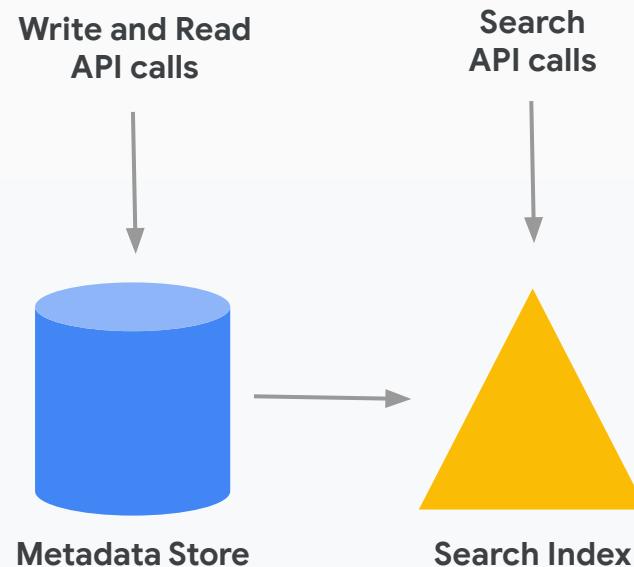
Data Catalog leverages the architecture and innovation of Google's internal metadata management service that is in wide use for many years.

Spanner, the globally distributed, strongly consistent database for storing all metadata entries

Real-time and batch syncers for auto-ingestion of technical metadata

Google search index with built-in ACL checks for data discovery -

Leverages the same technology that powers **Gmail** and **Google Drive**



Data Catalog - Data Discovery

UI enables search, read, and write access
to all metadata

Simple keyword search interface enables both, business and technical users

Facet search enables power users

type:view

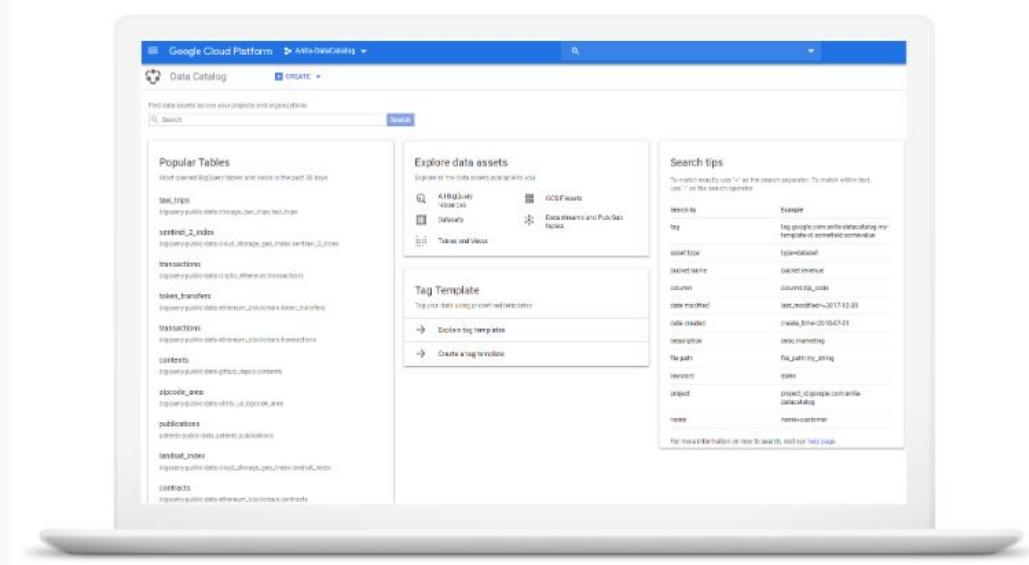
column:keyword

tag:approved_for_use: true

tag:data_classification:confidential

tag:has_pii:true

tag:type_pii:ssn



Programmatic access through API

Read, Write, and Search API for full metadata access

API powers bulk metadata updates

- Beta API with Python, Java, and Node.js language libraries
-

API enables Enterprise Applications and Custom Frontends

- GOJEK and others

```
# Create Tag Template
template = datacatalog_v1beta1.types.TagTemplate()
template.display_name = 'A test for v1beta1 using only primitive types'

template.fields['a-boolean'].display_name = 'BOOL field'
template.fields['a-boolean'].type.primitive_type = \
    datacatalog_v1beta1.enums.FieldType.PrimitiveType.BOOL.value

template.fields['a-double'].display_name = 'DOUBLE field'
template.fields['a-double'].type.primitive_type = \
    datacatalog_v1beta1.enums.FieldType.PrimitiveType.DOUBLE.value

template.fields['a-string'].display_name = 'STRING field'
template.fields['a-string'].type.primitive_type = \
    datacatalog_v1beta1.enums.FieldType.PrimitiveType.STRING.value

template.fields['a-timestamp'].display_name = 'TIMESTAMP field'
template.fields['a-timestamp'].type.primitive_type = \
    datacatalog_v1beta1.enums.FieldType.PrimitiveType.TIMESTAMP.value

template.fields['an-enum'].display_name = 'ENUM field'
template.fields['an-enum'].type.enum_type.allowed_values.append('VALUE_A')
template.fields['an-enum'].type.enum_type.allowed_values.append('VALUE_B')

client.create_tag_template(
    parent=client.project_path('YOUR_PROJECT_ID'),
    tag_template_id='YOUR_TEMPLATE_ID',
    tag_template=template)

# Create Tags
# =====
tag = datacatalog_v1beta1.types.Tag()
tag.template = client.tag_template_path('YOUR_PROJECT_ID', 'YOUR_TEMPLATE_ID')
tag.fields['a-boolean'].bool_value = True
tag.fields['a-double'].double_value = 3.1415
tag.fields['a-string'].string_value = 'tag-string'
tag.fields['a-timestamp'].timestamp_value.FromJsonString('2019-03-29T15:45:00-07:00')

bq_dataset_entry_id = 'YOUR_BQ_DATASET_ENTRY_ID'
client.create_tag(parent=client.entry_path('YOUR_PROJECT_ID', 'global', '@bigquery', bq_dataset_entry_id), tag=tag)
```

Technical and Business Metadata

Technical Metadata (auto-ingested from data source)

- Table names, column names
- Table descriptions, column descriptions
- Date created, date modified

Business Metadata (user provided / inferred)

- Table has PII
- Data quality owner
- 'Delete by' date
- 'Retain till' date
- Business logic used for computing a column
- Data quality score

Metadata access governed by IAM

User-1 permissions

- read access to all datasets in BQ

User-1 view in Data Catalog

- Discover all datasets
- Read access to all datasets

Dataset A

Dataset B

Dataset C

Dataset D

Dataset E

User-2 permissions

- read access to A, B, and C
- metadata-read only access to dataset D
- no read or metadata-read access to dataset E

Dataset A

Dataset B

Dataset C

Dataset D

Dataset E

User-2 view in Data Catalog

- Can discover datasets A, B, C, and D
- Discovers dataset D but needs to request read access to dataset D
- No metadata exfiltration for dataset E

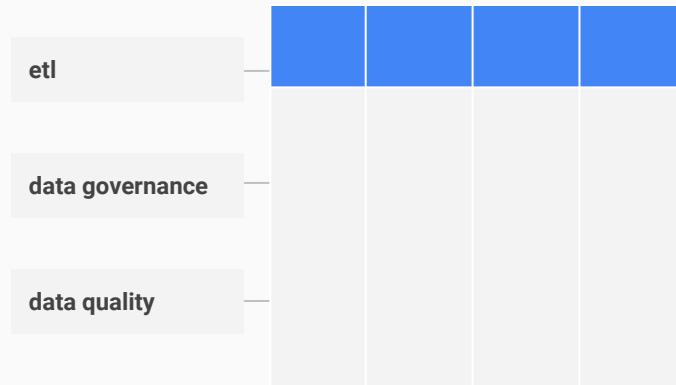
ACL controls on Business Metadata

Data governor permissions

- access to all tags

Data Governor view in Data Catalog

- can discover all tags

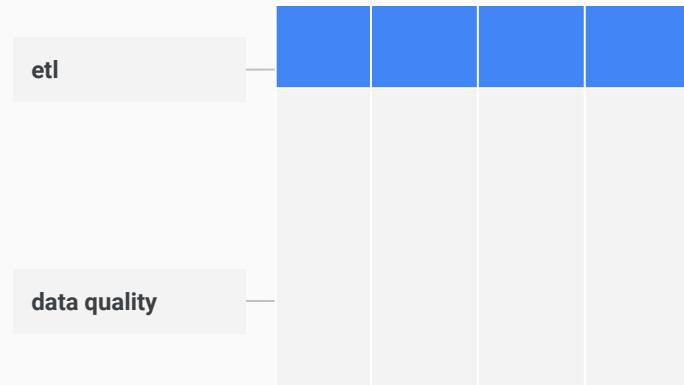


Data Analyst permissions

- no access to 'data governance' tags

Data Analyst view in Data Catalog

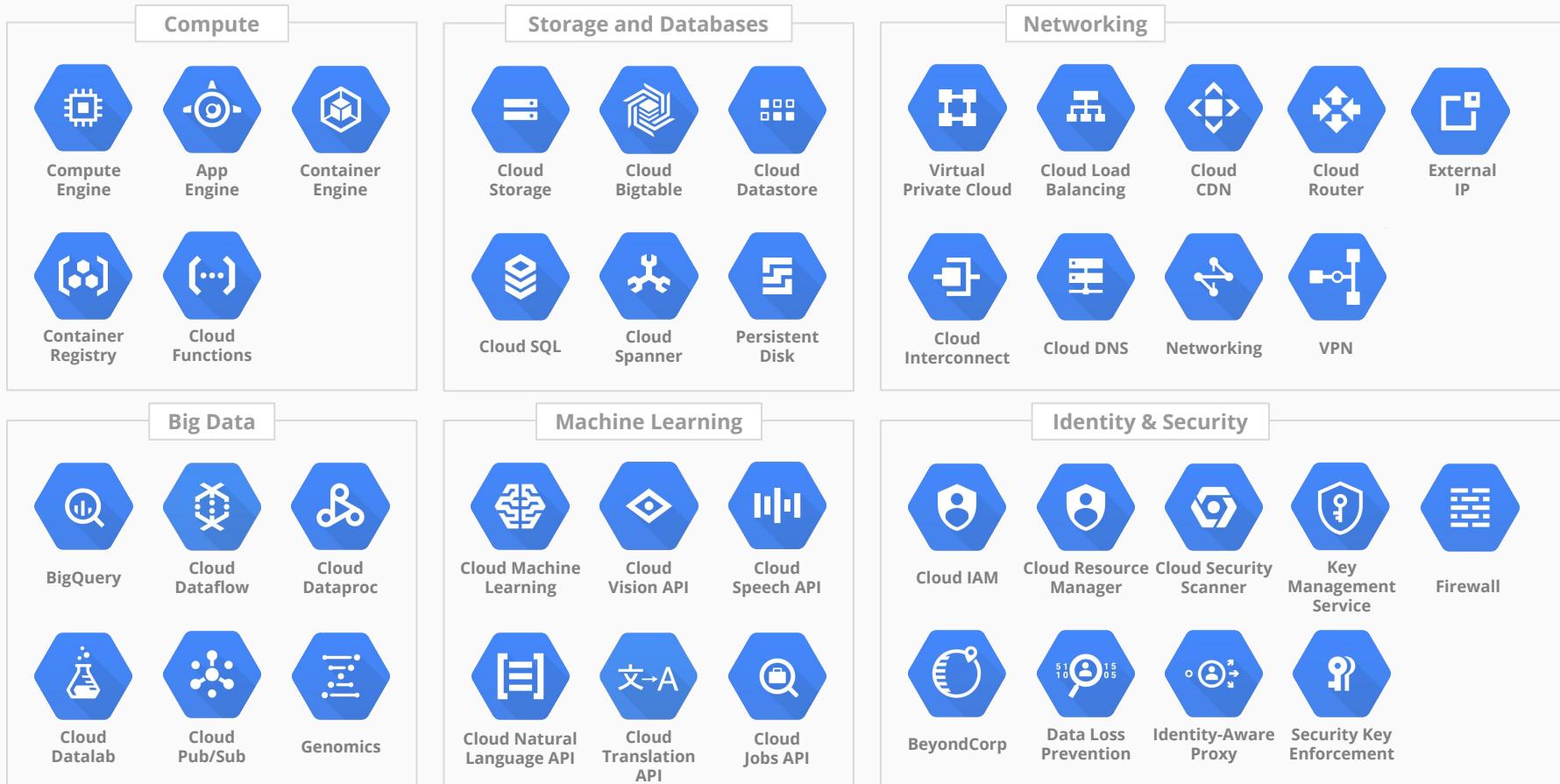
- can discover all accessible tags



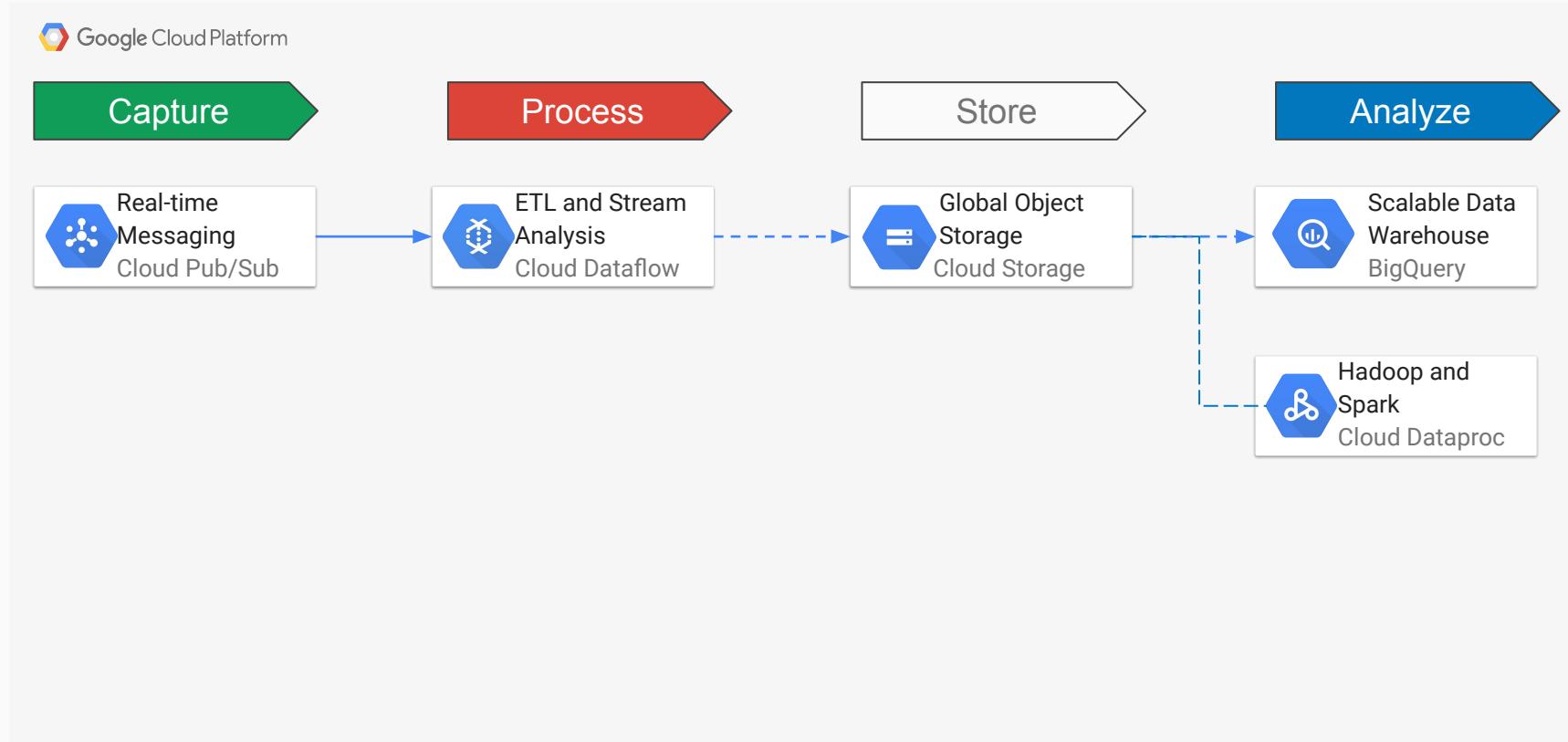
Arquitecturas de ejemplo

Google Cloud Platform Icons

<https://cloud.google.com/icons/>

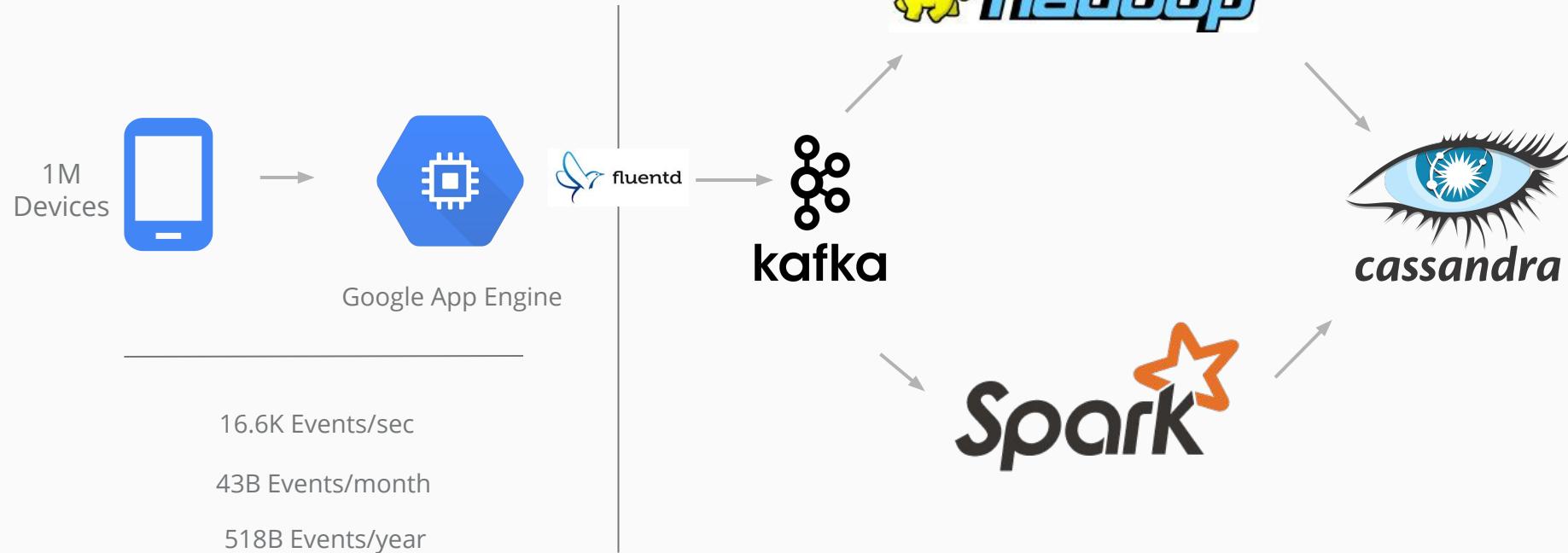


Sample Architecture - Big Data on GCP



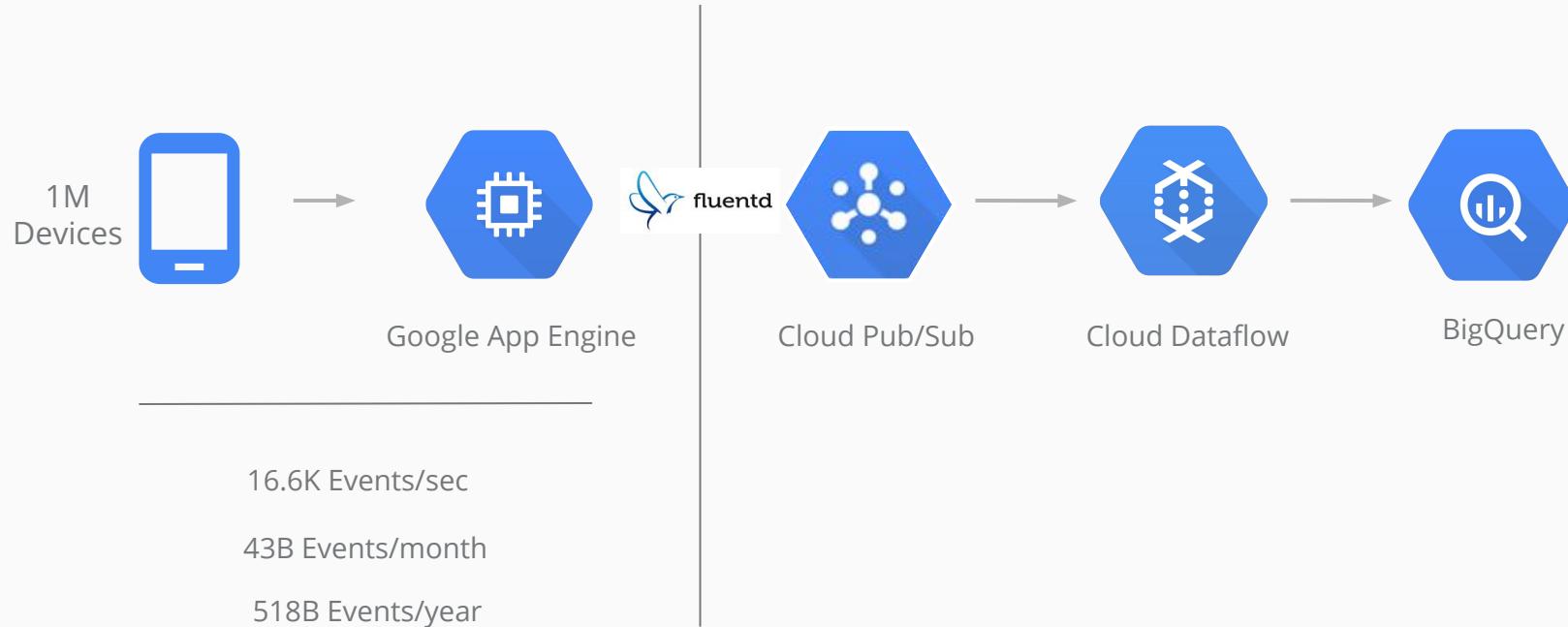
Sample Architecture - Data Analysis System (Open Source)

How many active users did I have in the last minute?

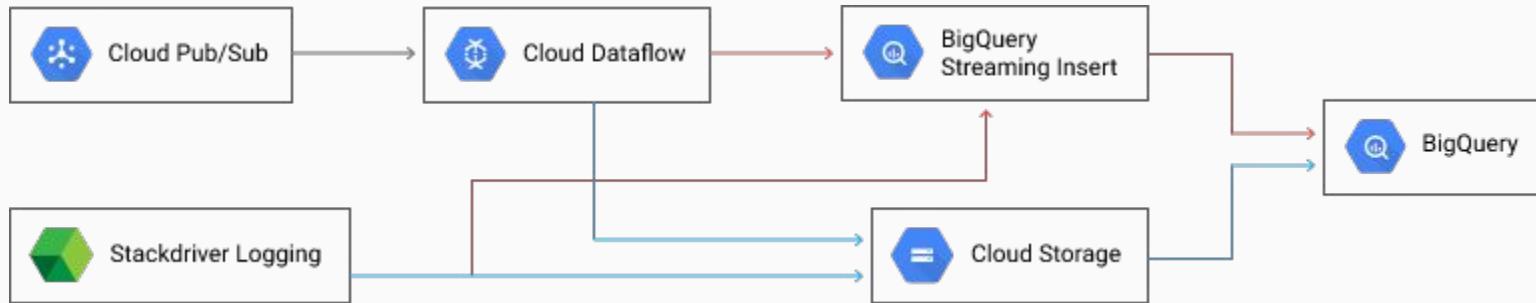


Sample Architecture - Data Analysis System (GCP)

How many active users did I have in the last minute?

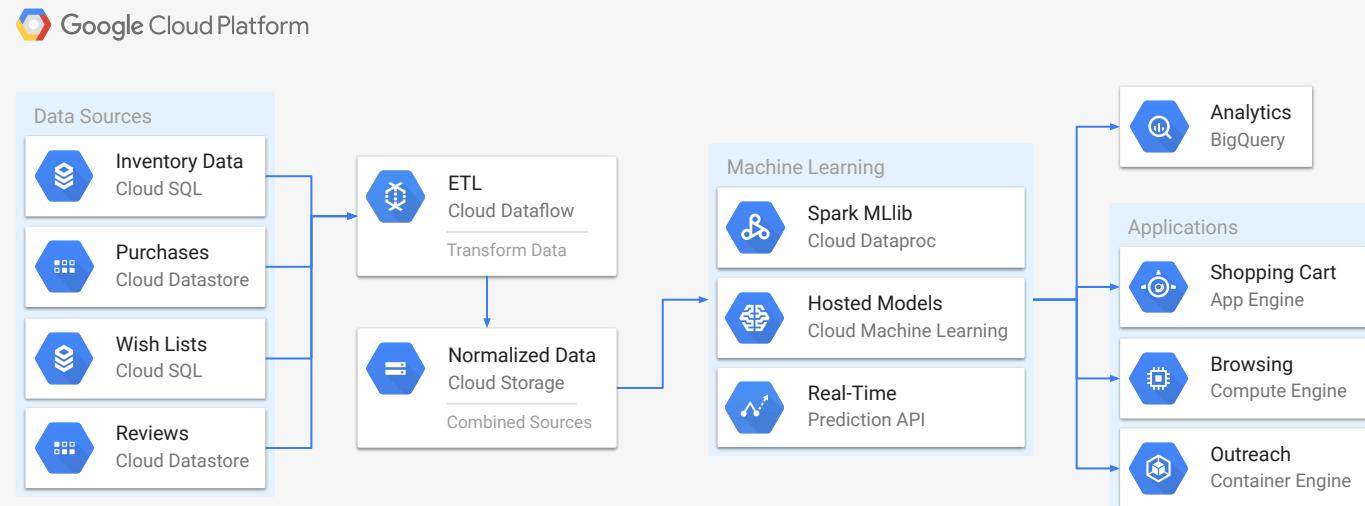


Sample Architecture - Logs Analysis

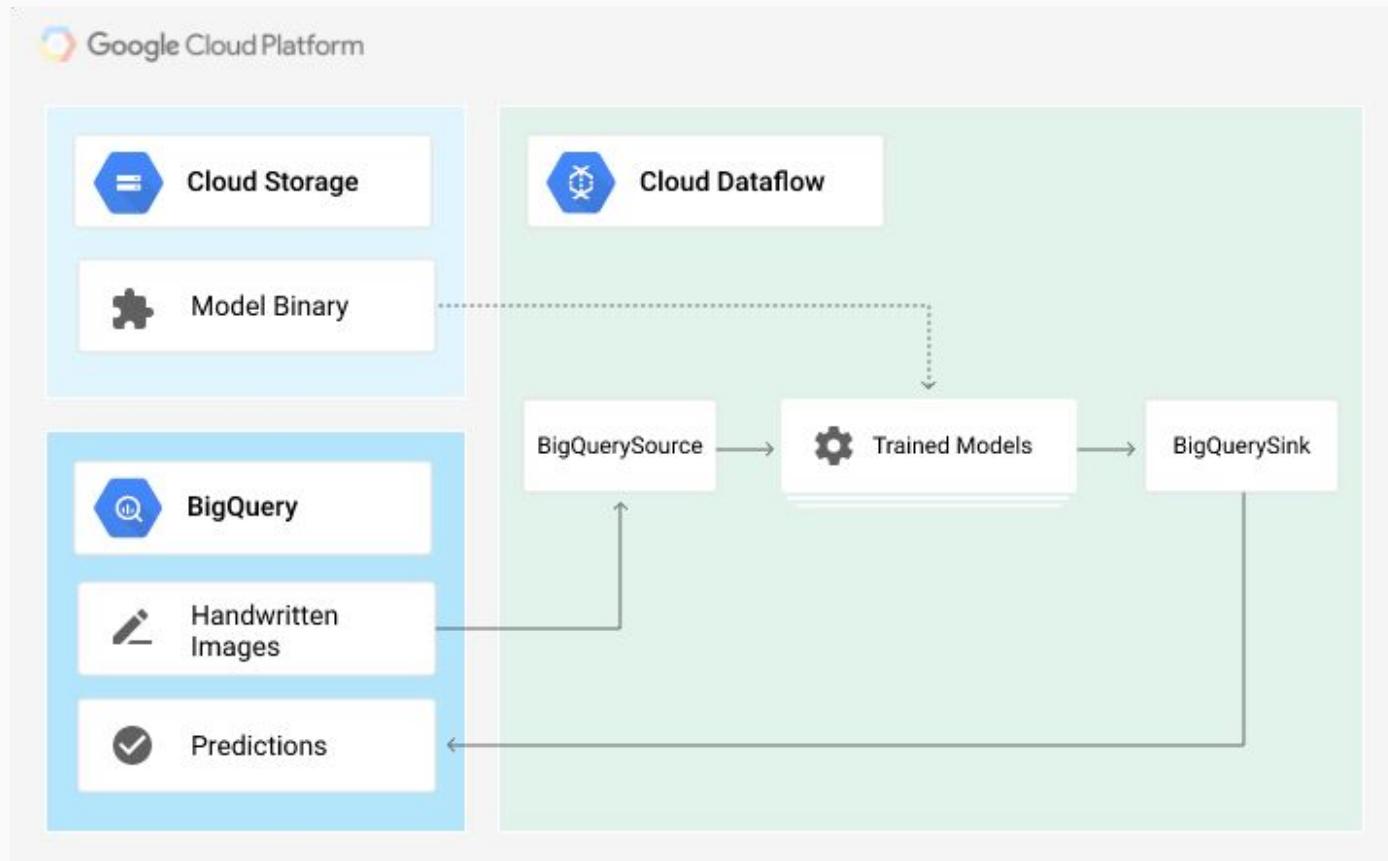


Samples - Retail > Recommendation Engines

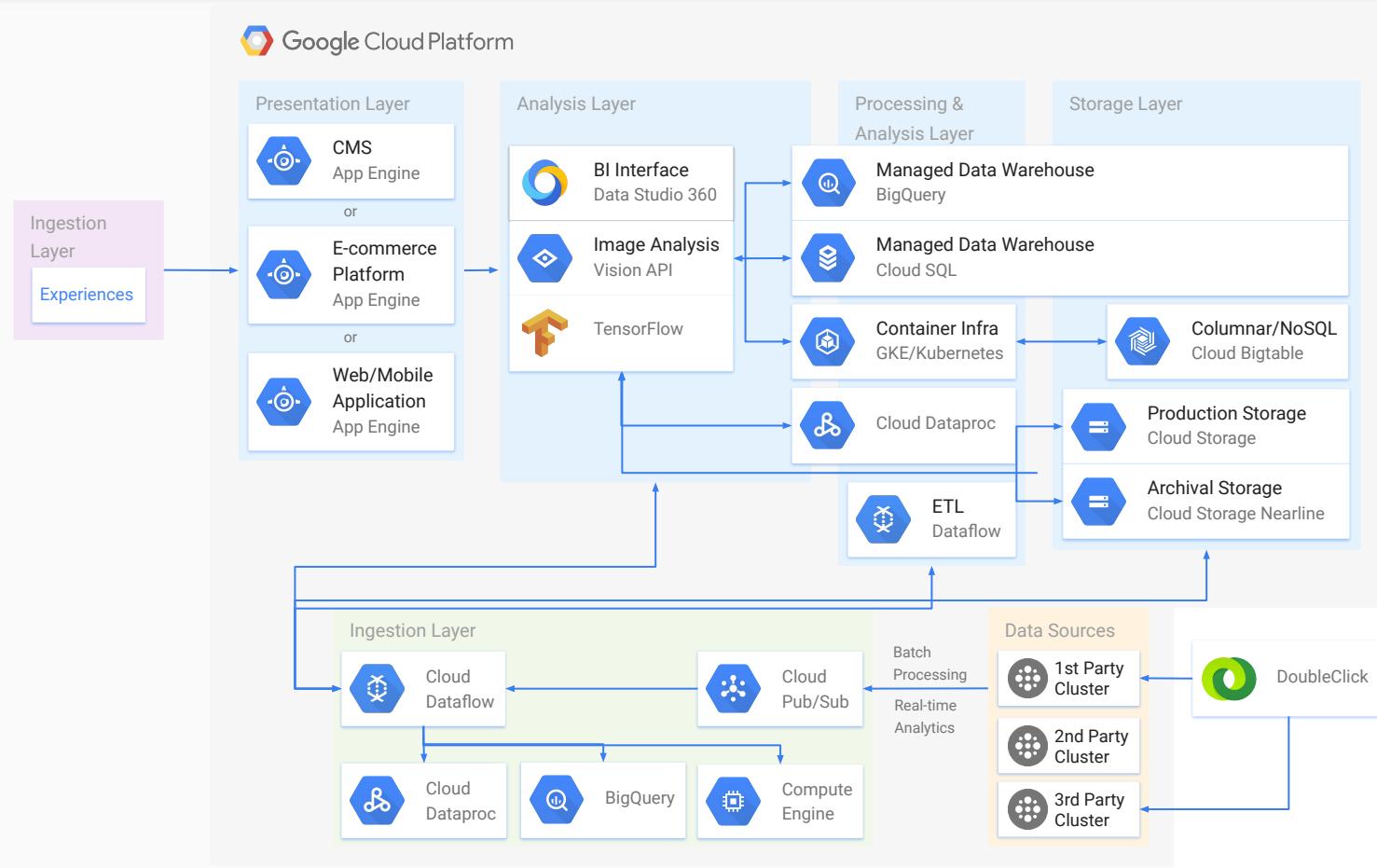
Architecture: Retail > Recommendation Engines



How to use Google Cloud Dataflow with TensorFlow for batch predictive analysis

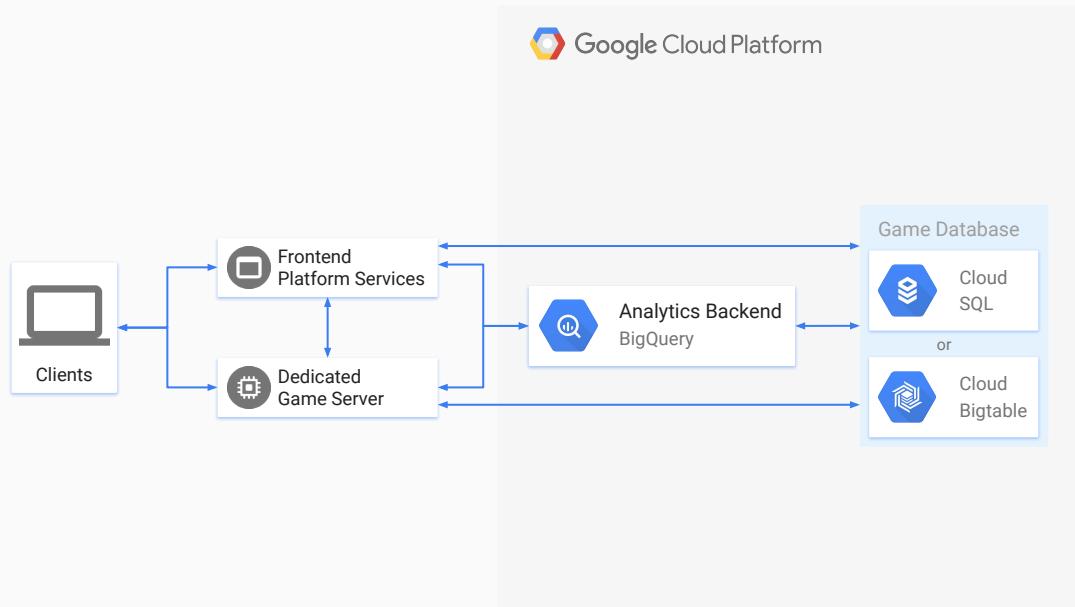


Samples - Digital Marketing > Publisher side analysis



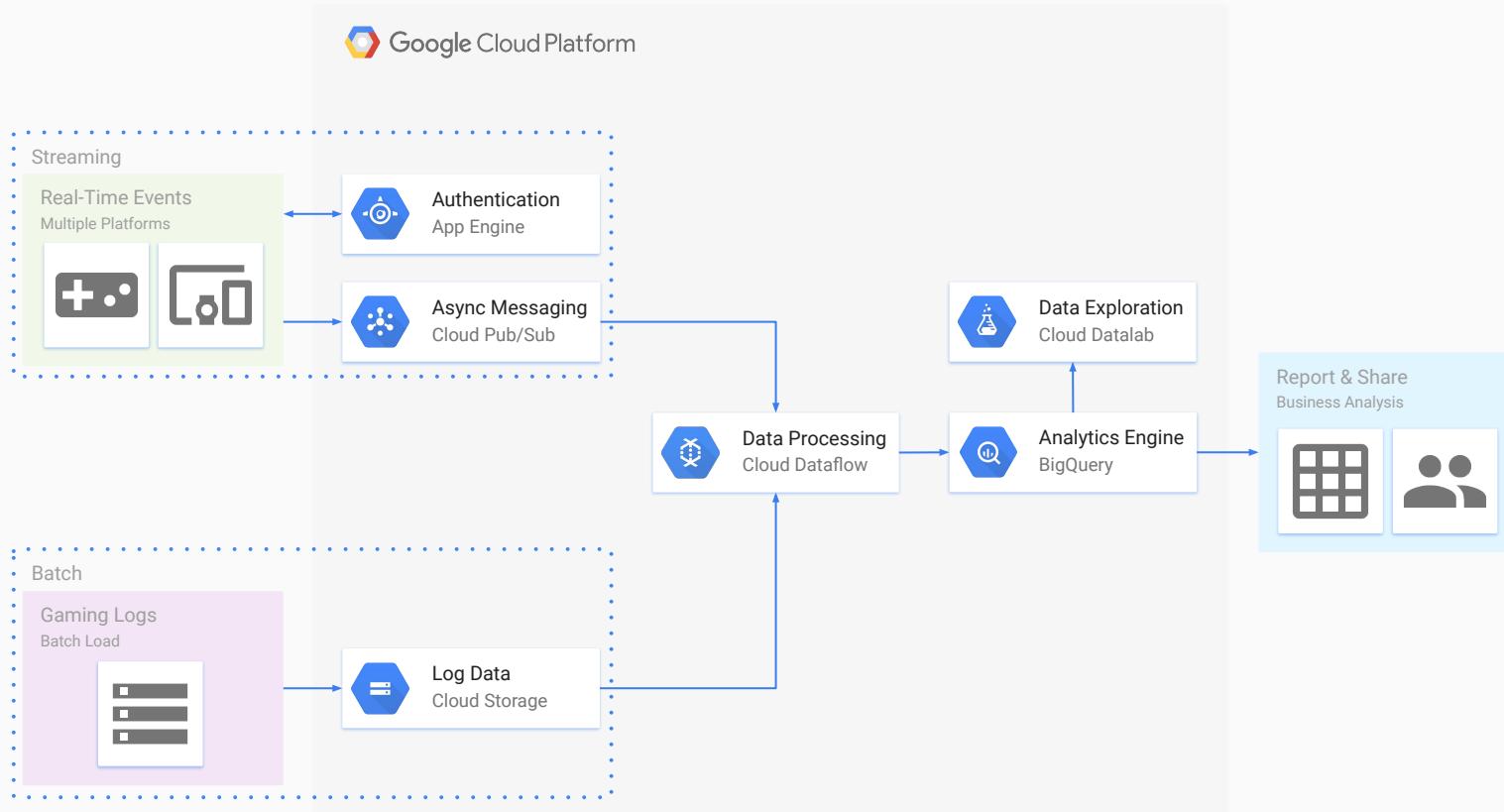
Samples - Gaming > Backend Database

Architecture: Gaming > Backend Database



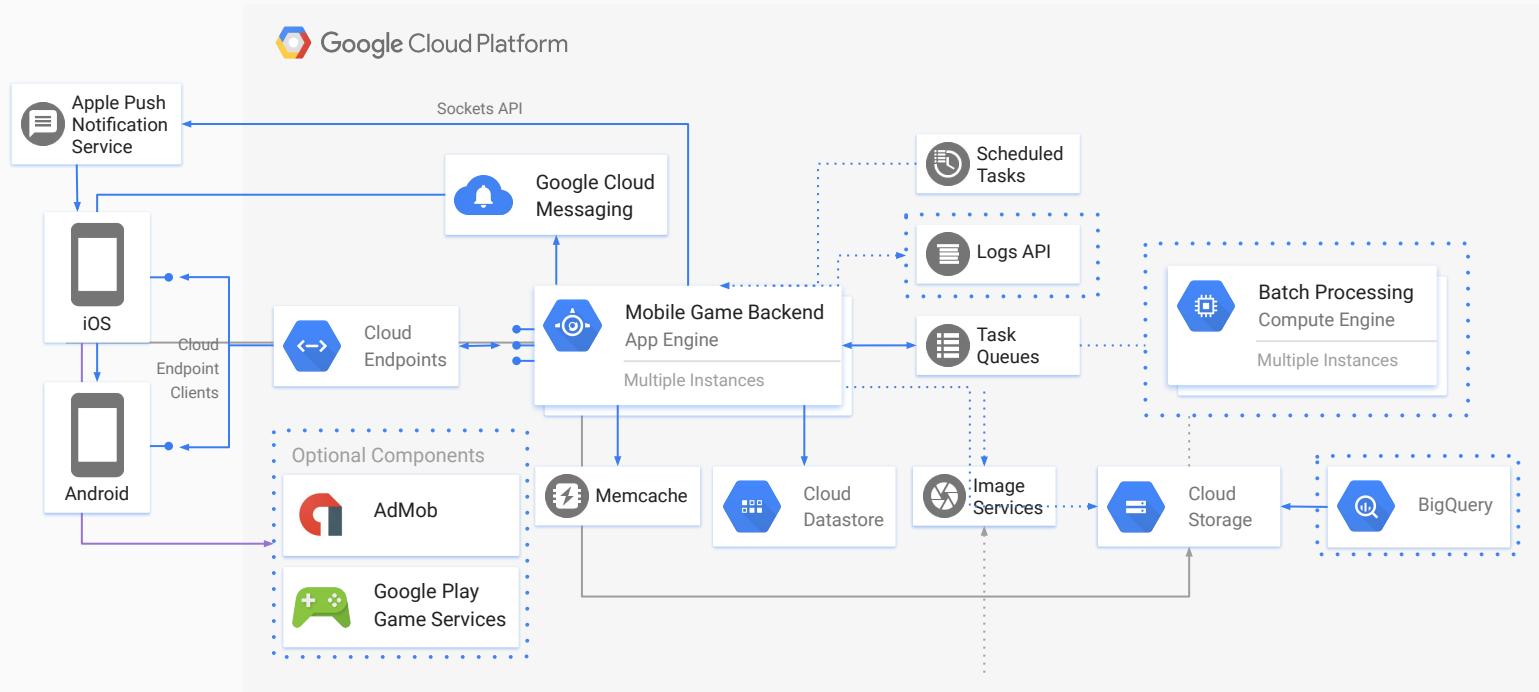
Samples - Gaming > Gaming Analytics

Architecture: Gaming > Gaming Analytics

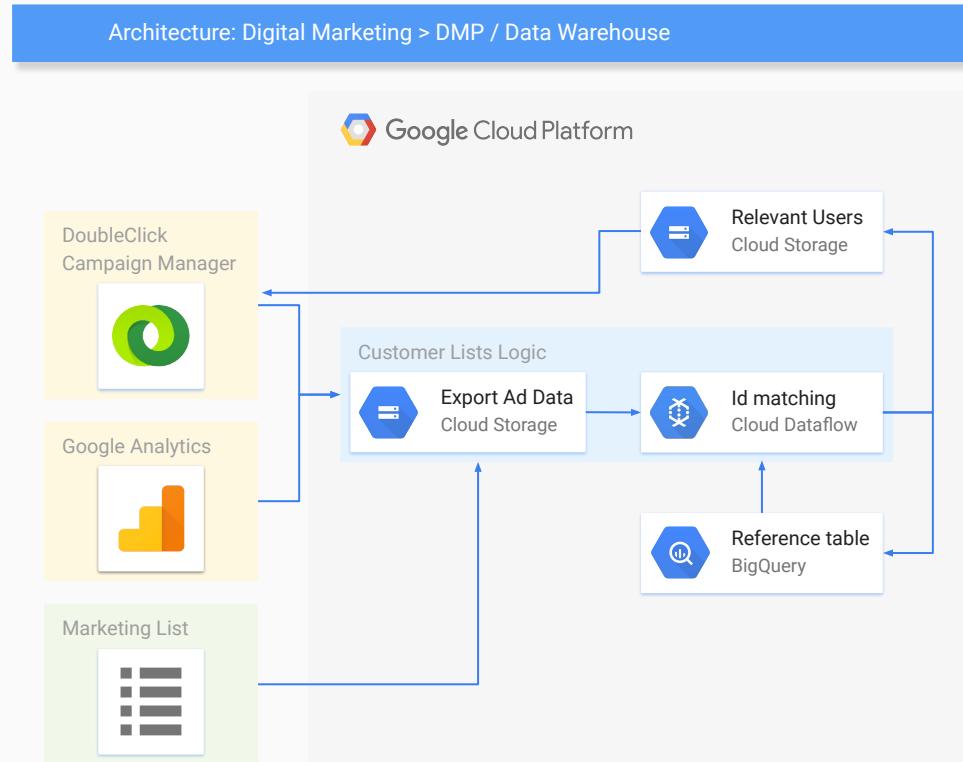


Samples - Gaming > Mobile Game Backend

Architecture: Gaming > Mobile Game Backend

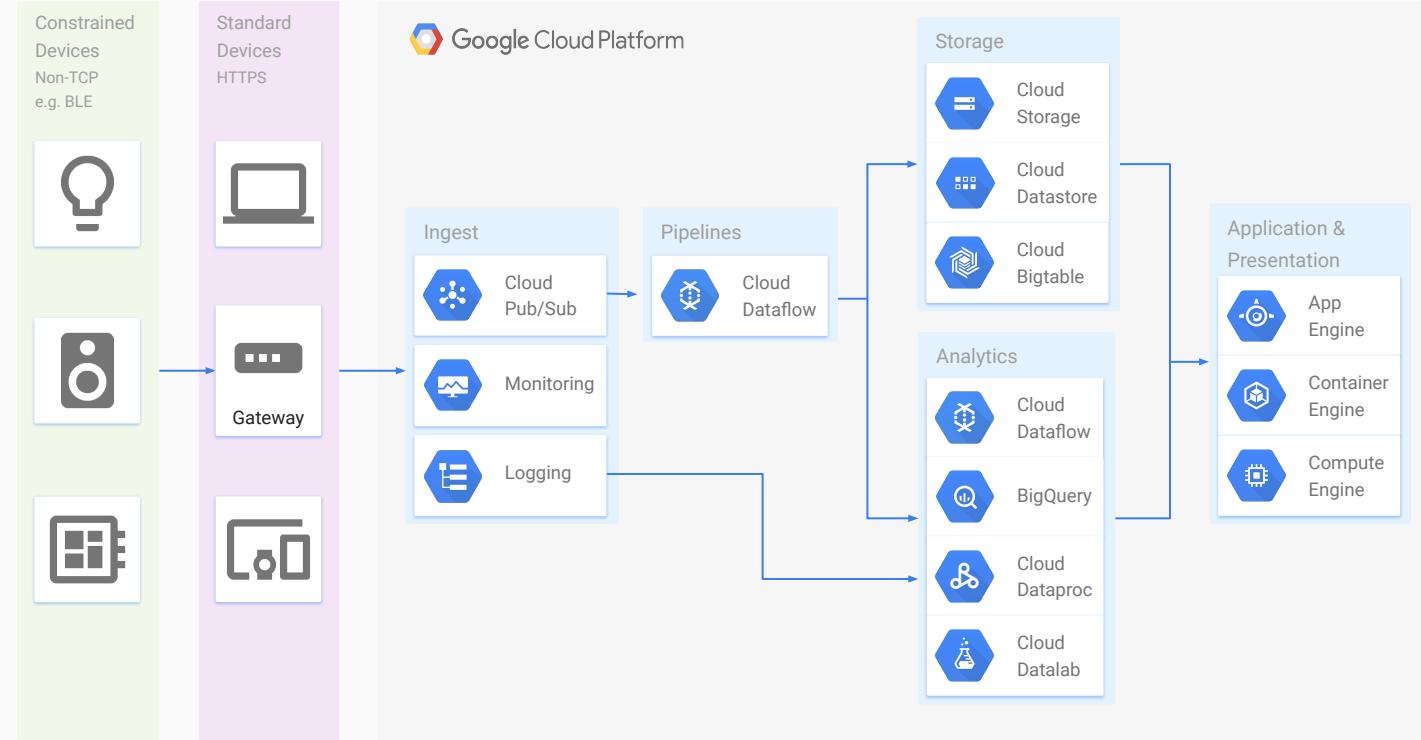


Samples - Digital Marketing > DMP / Data Warehouse



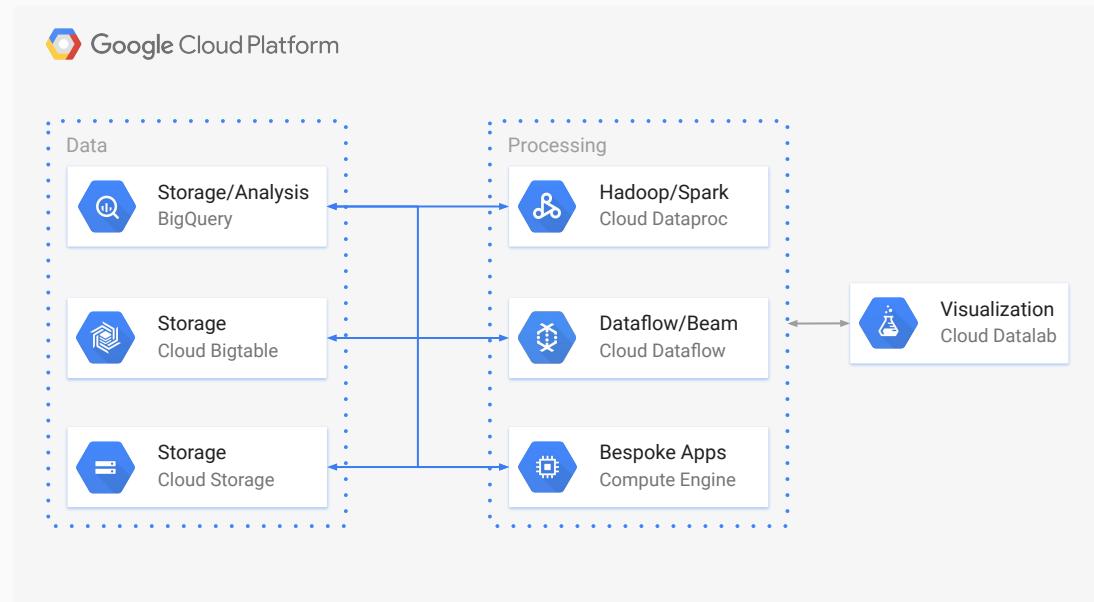
Samples - Internet of Things > Sensor stream ingest and processing

Architecture: Internet of Things > Sensor stream ingest and processing



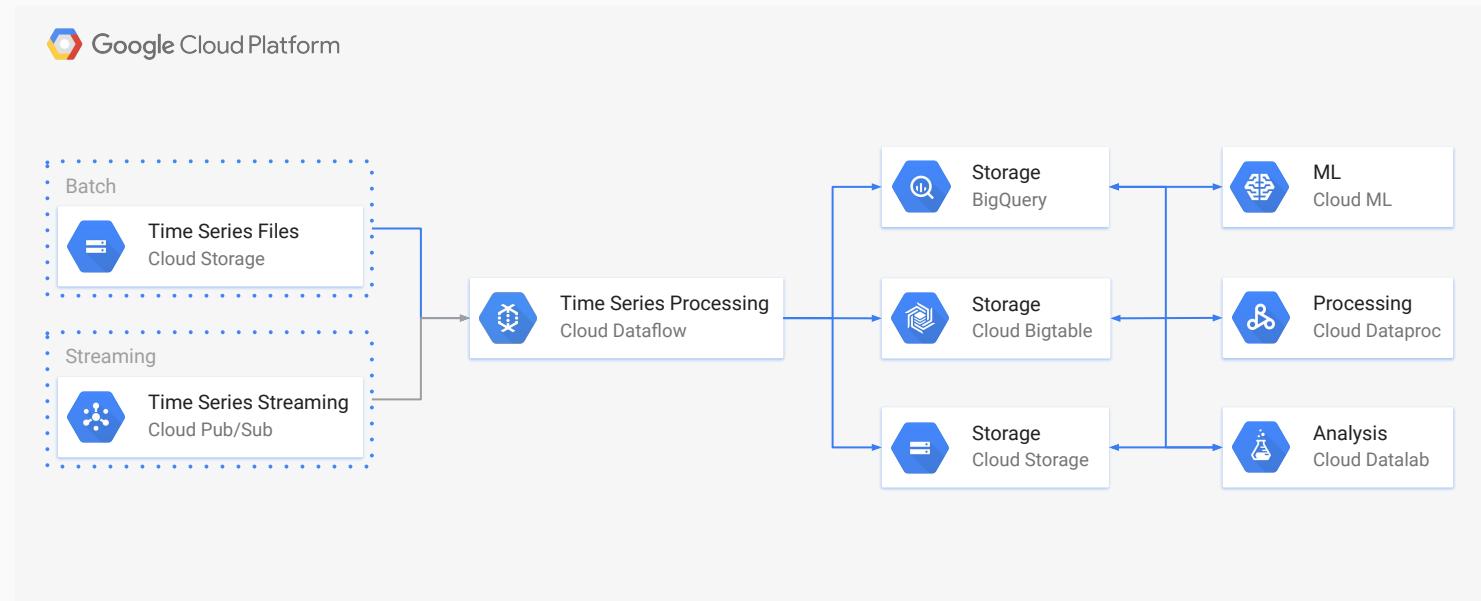
Samples - Financial Services > Monte Carlo Simulations

Architecture: Financial Services > Monte Carlo Simulations



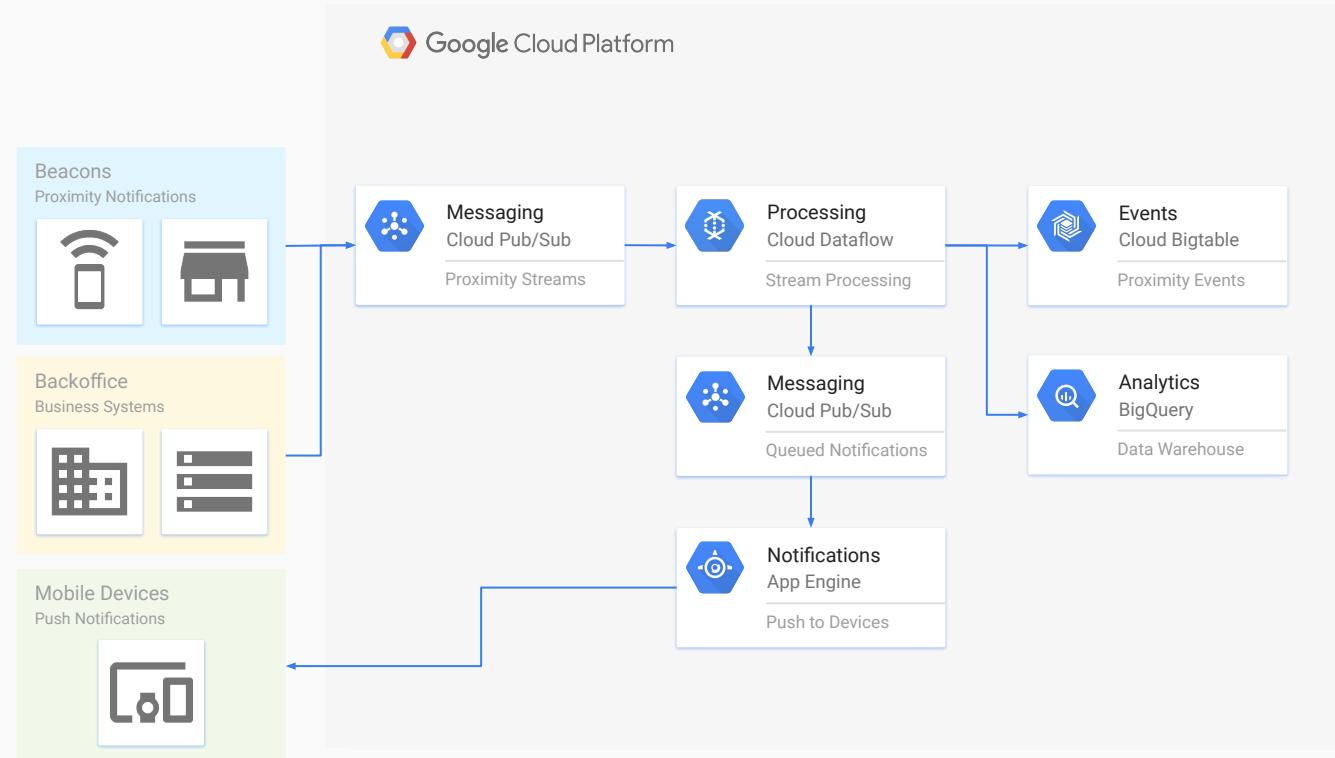
Samples - Financial Services > Time Series Analysis

Architecture: Financial Services > Time Series Analysis

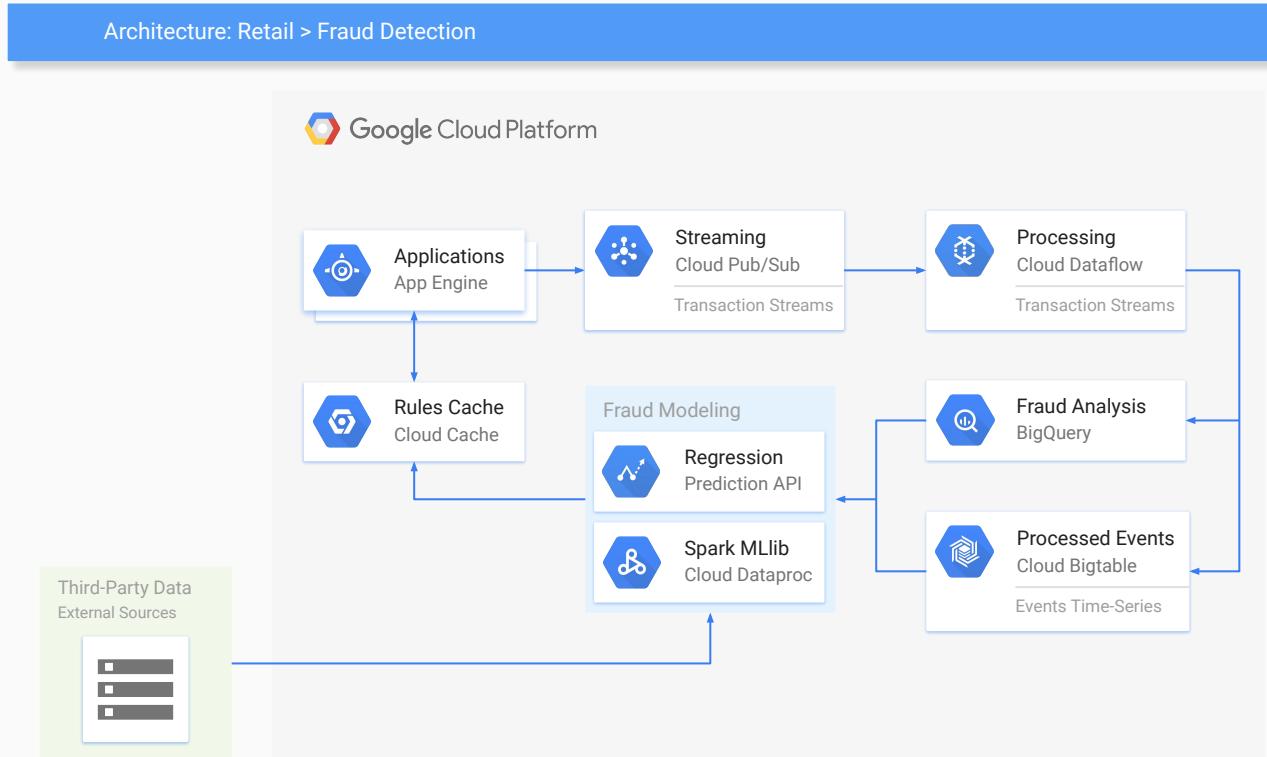


Samples - Retail > Beacons and Targeted Marketing

Architecture: Retail > Beacons and Targeted Marketing

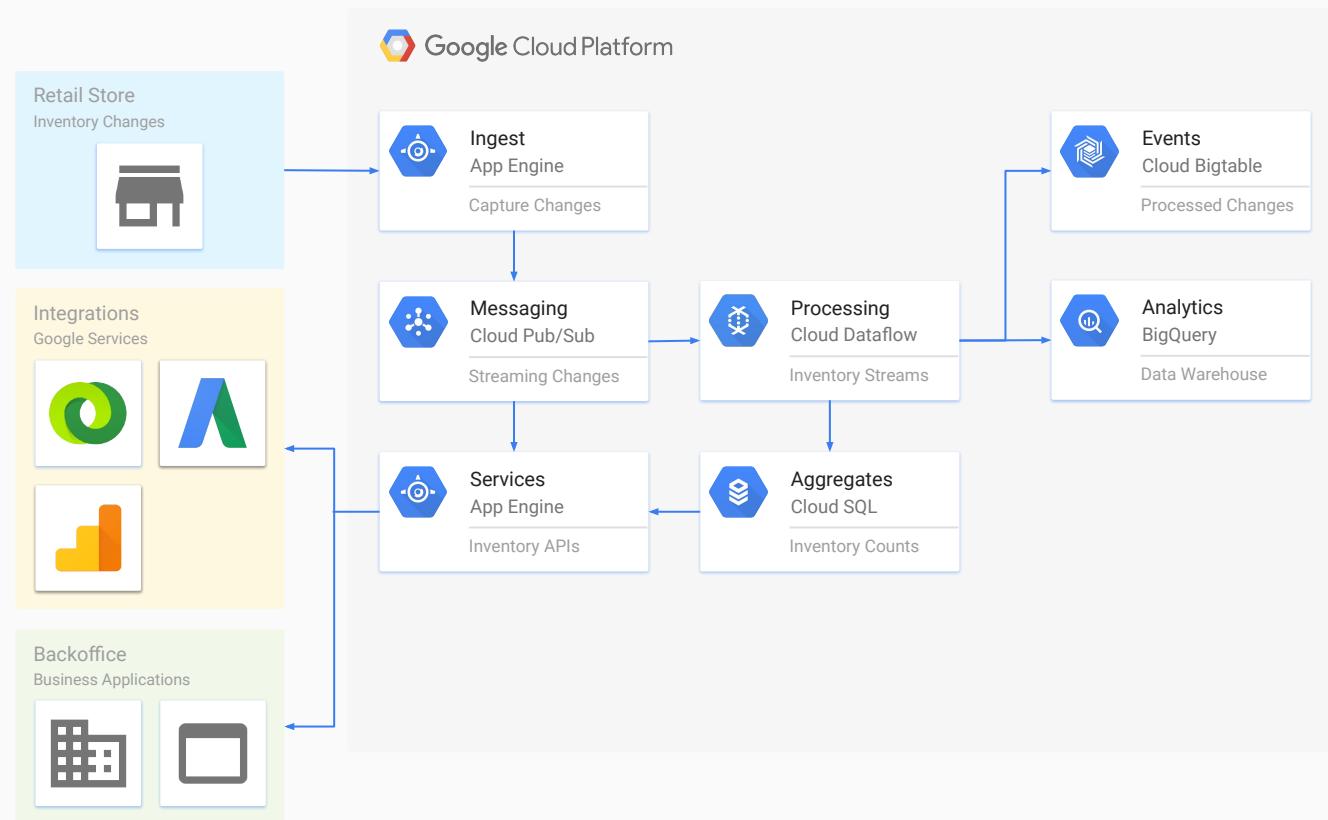


Samples - Retail > Fraud Detection



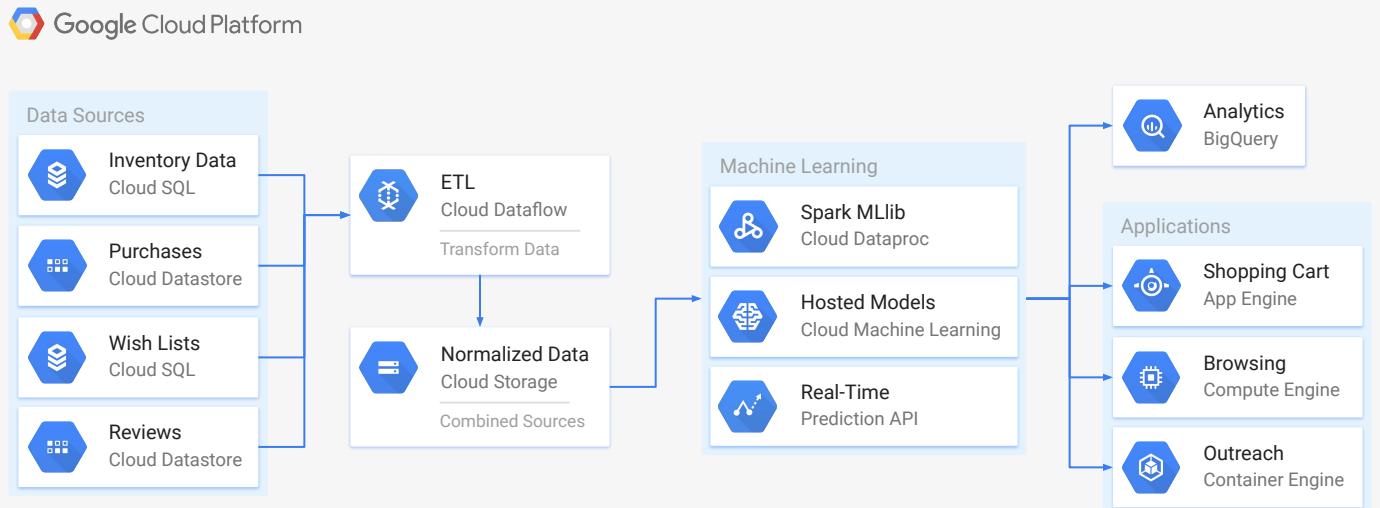
Samples - Retail > Real-Time Inventory

Architecture: Retail > Real-Time Inventory



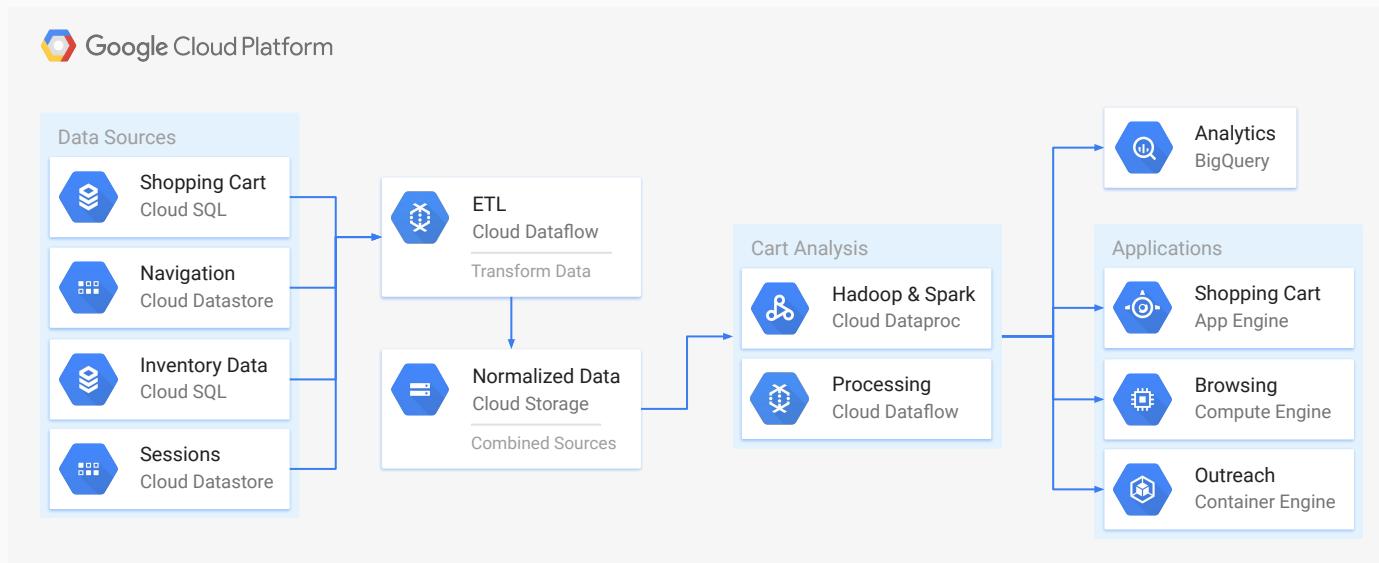
Samples - Retail > Recommendation Engines

Architecture: Retail > Recommendation Engines

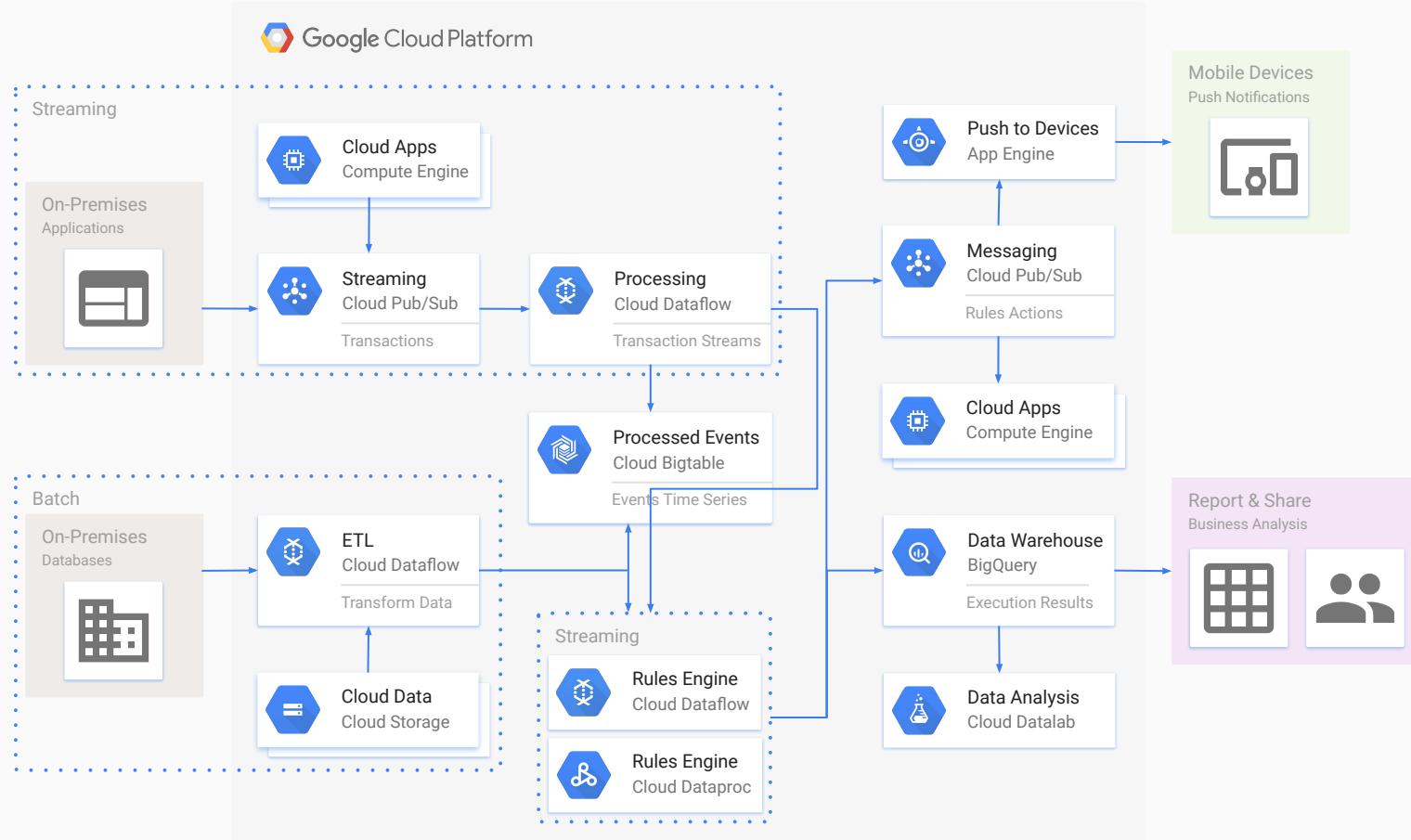


Samples - Retail > Shopping Cart Analysis

Architecture: Retail > Shopping Cart Analysis

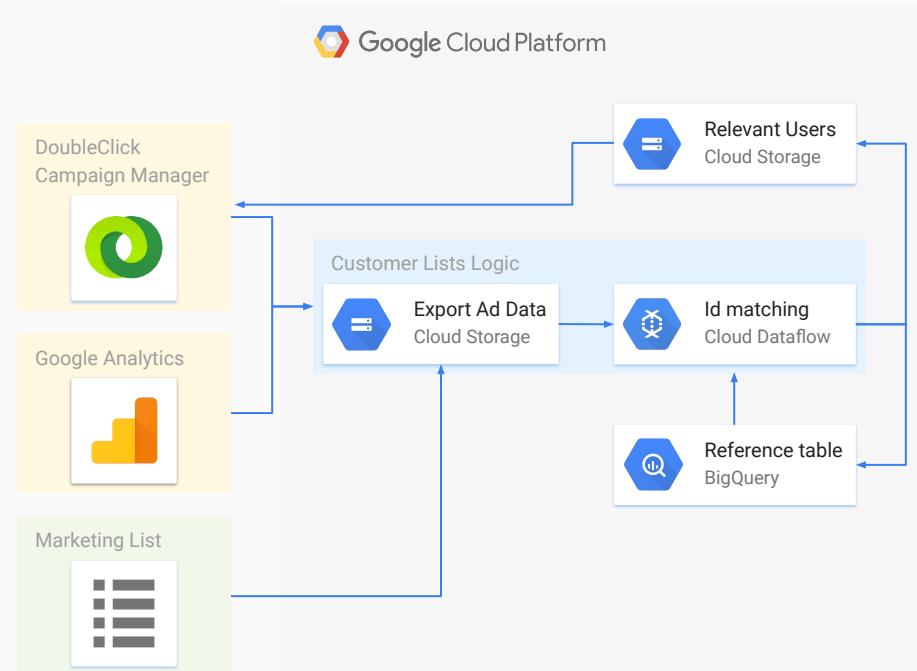


Samples - Big Data > Complex Event Processing



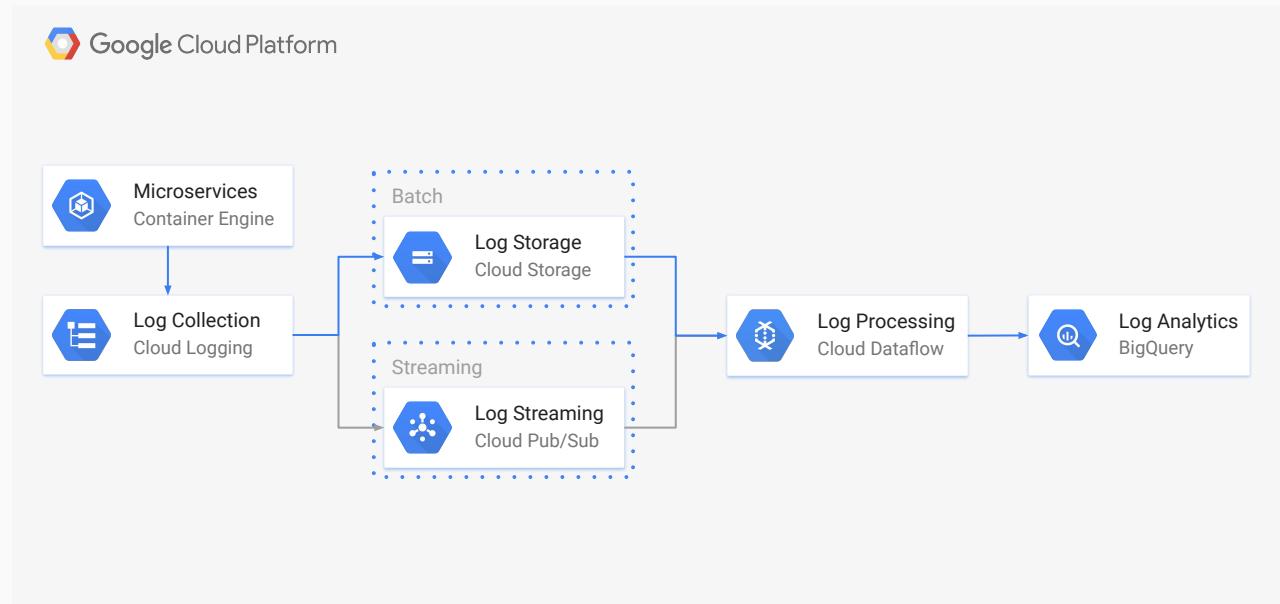
Samples - Big Data > DMP / Data Warehouse

Architecture: Big Data > DMP / Data Warehouse



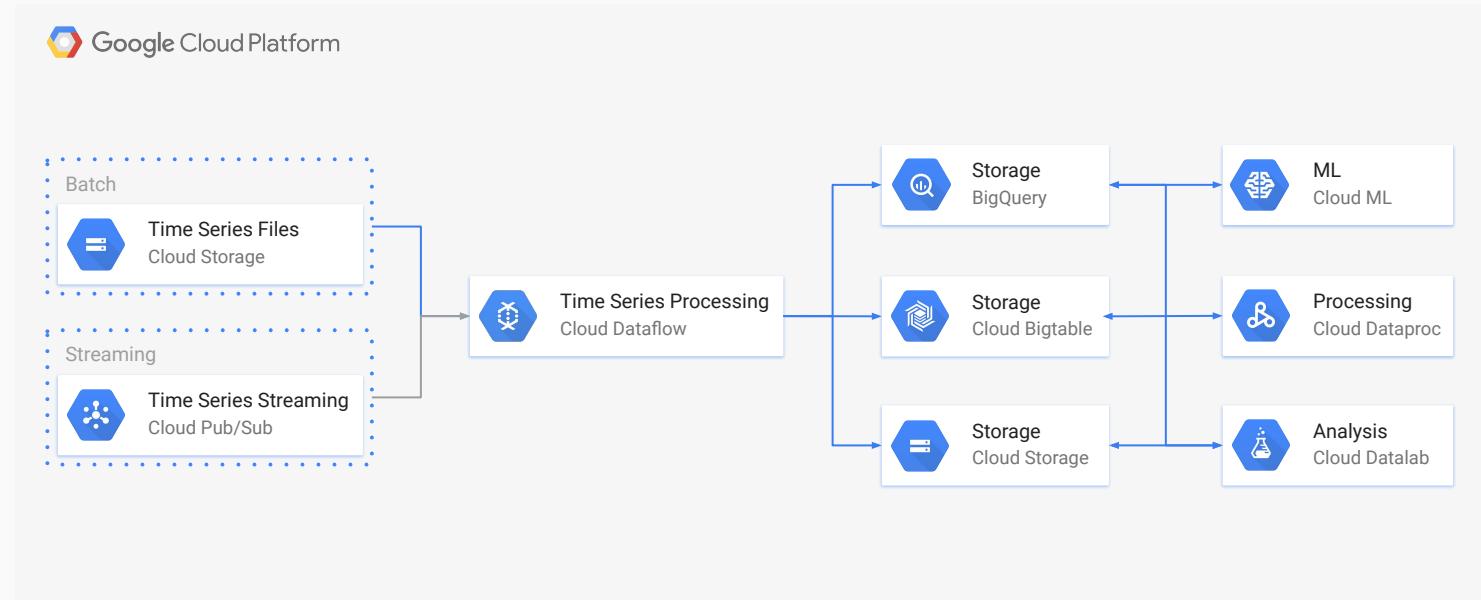
Samples - Big Data > Log Processing

Architecture: Big Data > Log Processing



Samples - Big Data > Time Series Analysis

Architecture: Big Data > Time Series Analysis



¿ Preguntas ?

Ismael Yuste

linkedin.com/in/ismaelyuste/

@IsmaelYuste

