Google Cloud

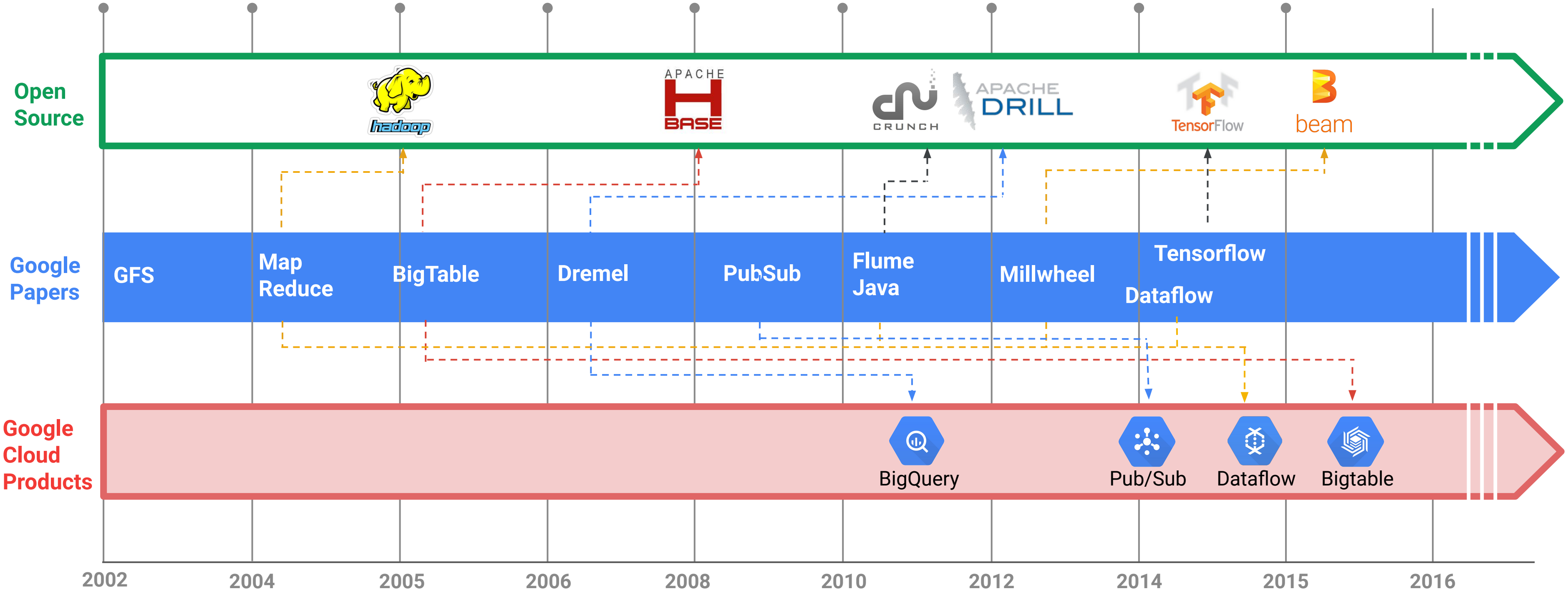
Data Analytics on GCP

Spyrales.fr

Pascal Rabier
2020-04-30

NO MORE TURNING AWAY
BITES THE DUST
WHITE ROOM WHEN I COME HOME COLD AND TIRED
BLACK CURTAINS IT'S GOOD TO WARM MY BONES BESIDE THE FIRE
NEAR THE STATION FAR AWAY, ACROSS THE FIELD, TOLLING ON THE IRON BELL
ROOF COUNTRY, CALLS THE FAITHFUL TO THEIR KNEES
REMEMBER WHEN YOU WERE YOUNG, YOU SHONE LIKE THE SUN
SHINE ON YOU CRAZY DIAMOND
NOW THERE'S A LOOK IN YOUR EYES
DREAM
WAITING
I MEAN A "YES"
PEOPLE TALKING WITHOUT SPEAKING
PEOPLE HEARING WITHOUT LISTENING
SONGS THAT VOICES NEVER SHARE
THREATENED BY SHADOWS AT NIGHT
AND EXPOSED IN THE LIGHT
TIGERS CROUCHED IN JUNGLES
GOODBYE WINDOWS

15+ Years of Tackling Big Data Problems

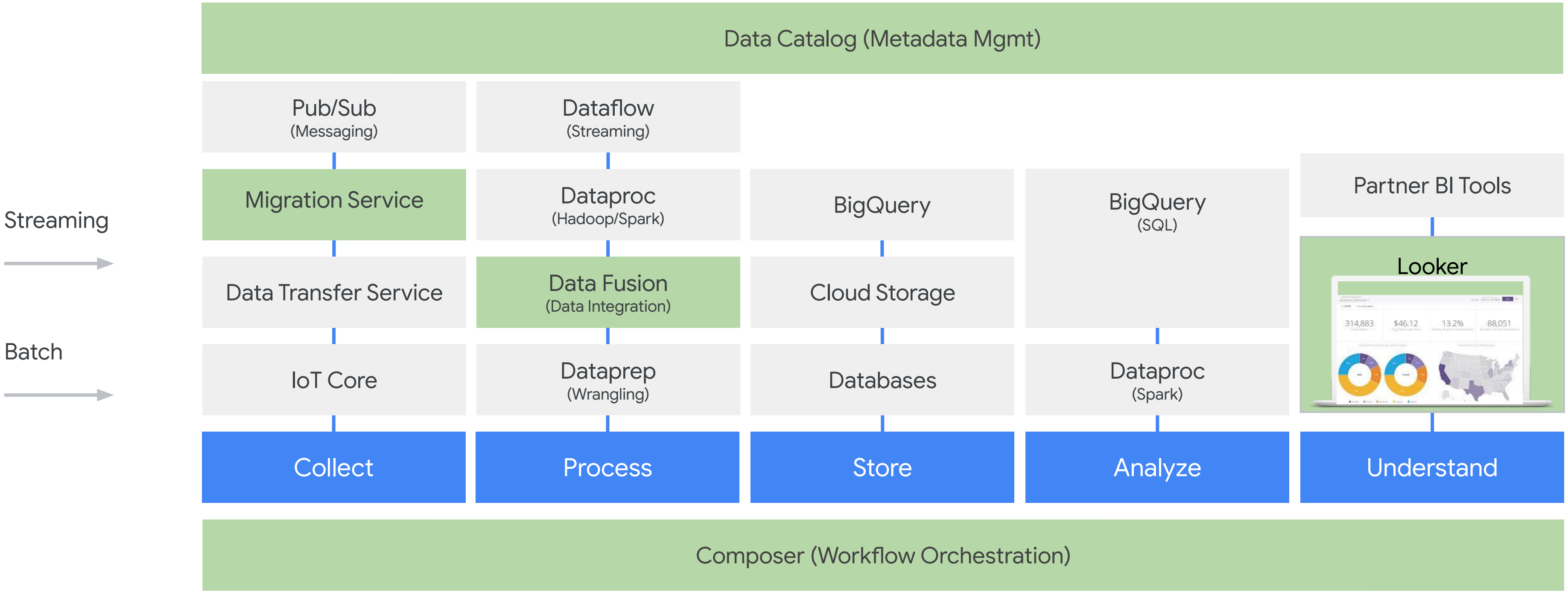


Fully managed storage & database services

Object	Key-value	Non-relational		Relational		Warehouse
Cloud Storage	**App Engine Memcache**	**Cloud Firestore**	**Cloud Bigtable**	**Cloud SQL**	**Cloud Spanner**	**BigQuery**
Binary or object data	Web/mobile applications, gaming	Hierarchical, mobile, web	Heavy read + write, events	Web frameworks	RDBMS+scale, HA, HTAP	Enterprise Data Warehouse
Images, media serving, backups	Game state, user sessions	User profiles, Game State	AdTech, financial, IoT	CMS, eCommerce	Transactions, Ad/Fin/MarTech	Analytics, Dashboards

Google’s Smart Analytics Platform

Collect, process, store, analyze and visualize data and insights



Data Catalog (Metadata Mgmt)

Streaming

Batch

Collect: Pub/Sub (Messaging), Migration Service, Data Transfer Service, IoT Core

Process: Dataflow (Streaming), Dataproc (Hadoop/Spark), Data Fusion (Data Integration), Dataprep (Wrangling)

Store: BigQuery, Cloud Storage, Databases

Analyze: BigQuery (SQL), Dataproc (Spark)

Understand: Partner BI Tools, Looker

Composer (Workflow Orchestration)

Smart Analytics as a Service: Fully Managed. Serverless. Enterprise class. Globally Distributed. Secure

Providing choice to customers

Cloud Native Services

Differentiation

- BigQuery
- Dataflow
- Pub/Sub
- Data Catalog
- looker

Managed Open Source Services

Familiarity

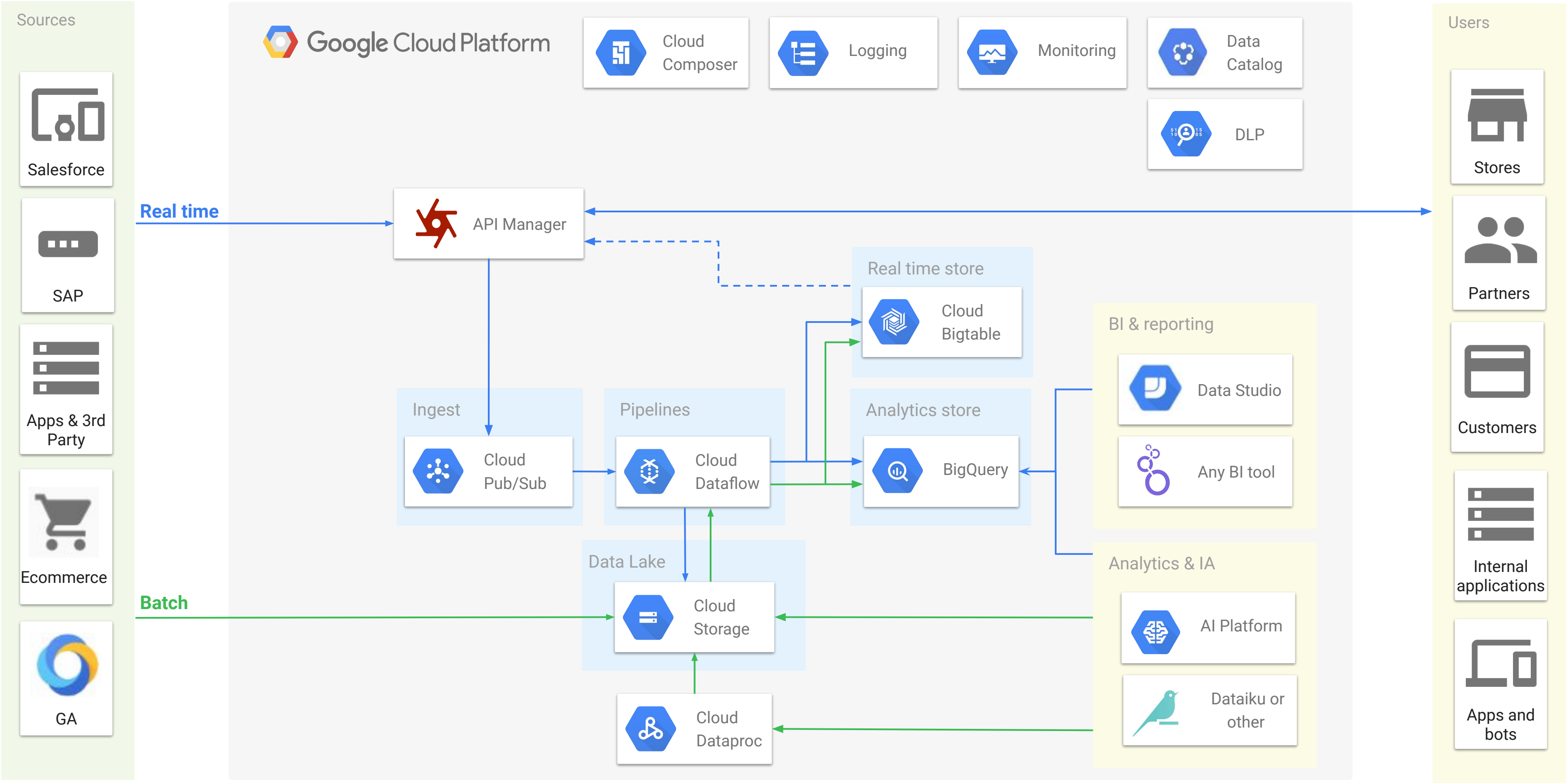
- Composer
- beam
- Dataproc
- Data Fusion

Partner Services

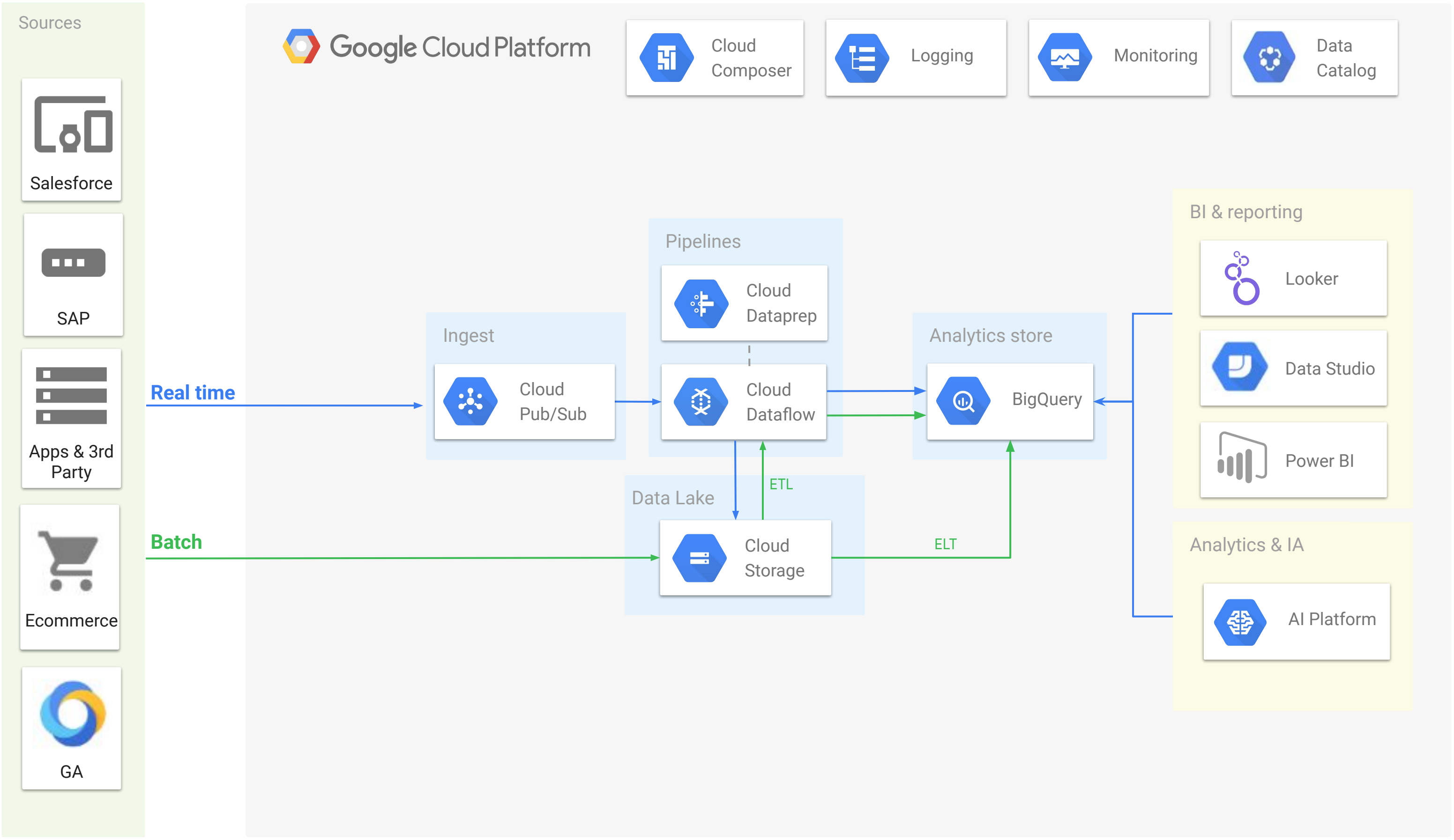
Completeness

- confluent
- Dataprep
- Informatica
- collibra
- Qubole

An example of a Big Data architecture with GCP



First step


Sources
Salesforce
SAP
Apps & 3rd Party
Ecommerce
GA
Real time
Batch
Google Cloud Platform
Cloud Composer
Logging
Monitoring
Data Catalog
Ingest
Cloud Pub/Sub
Pipelines
Cloud Dataprep
Cloud Dataflow
ETL
Data Lake
Cloud Storage
ELT
Analytics store
BigQuery
BI & reporting
Looker
Data Studio
Power BI
Analytics & IA
AI Platform

BigQuery

BigQuery

Google Cloud Platform's **enterprise data warehouse** for analytics

Gigabyte- to **petabyte-scale** storage and SQL queries

Encrypted, durable,
And highly available

Fully managed and **serverless** for maximum agility and scale

Unique

Real-time insights from streaming data

Unique

Built-in **ML** for out-of-the-box predictive insights

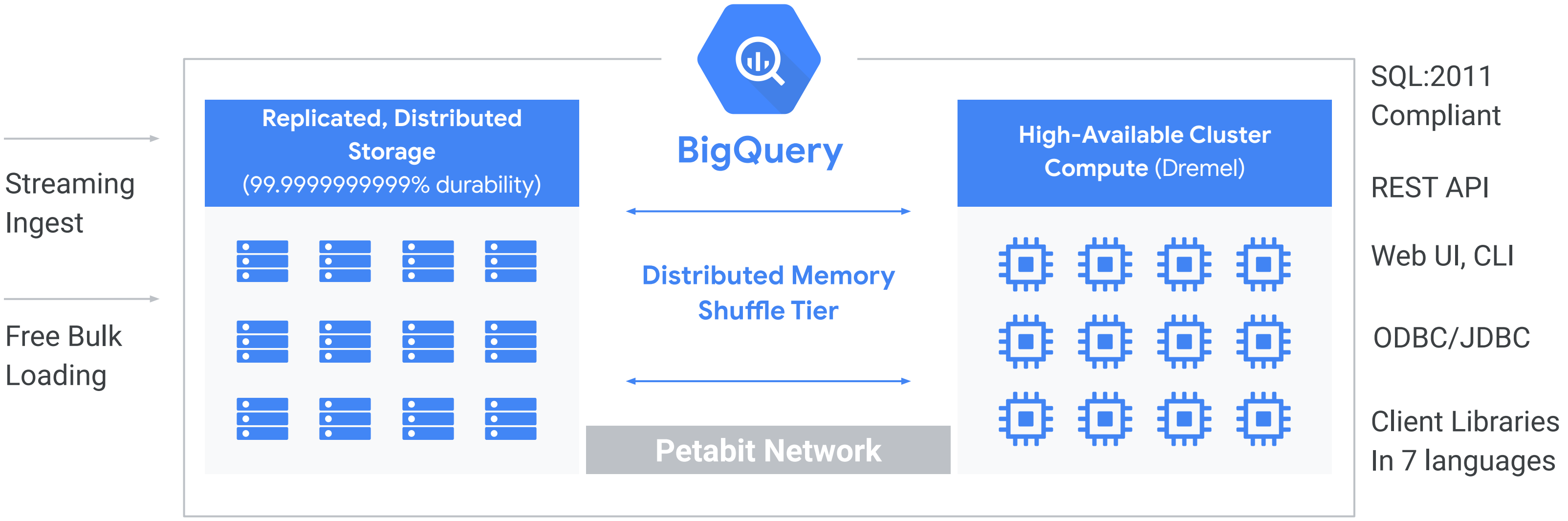
Unique

High-speed, in-memory **BI Engine** for faster reporting and analysis

Unique

BigQuery | Architecture

Decoupled storage and compute for maximum flexibility

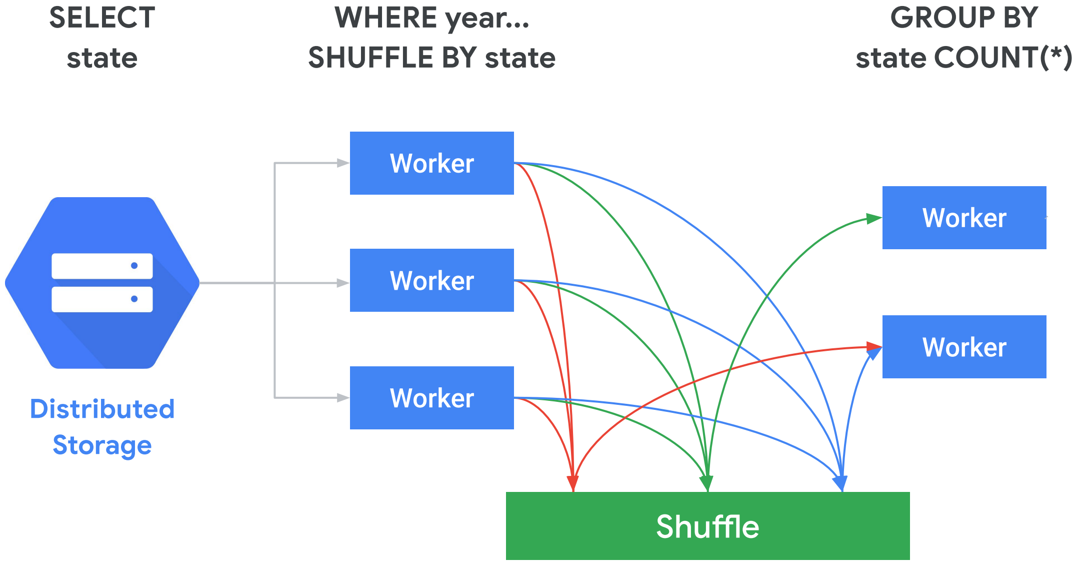

Streaming
Ingest
Free Bulk
Loading
Replicated, Distributed Storage
(99.9999999999% durability)
BigQuery
Distributed Memory Shuffle Tier
Petabit Network
High-Available Cluster Compute (Dremel)
SQL:2011 Compliant
REST API
Web UI, CLI
ODBC/JDBC
Client Libraries In 7 languages

BigQuery remote memory shuffle

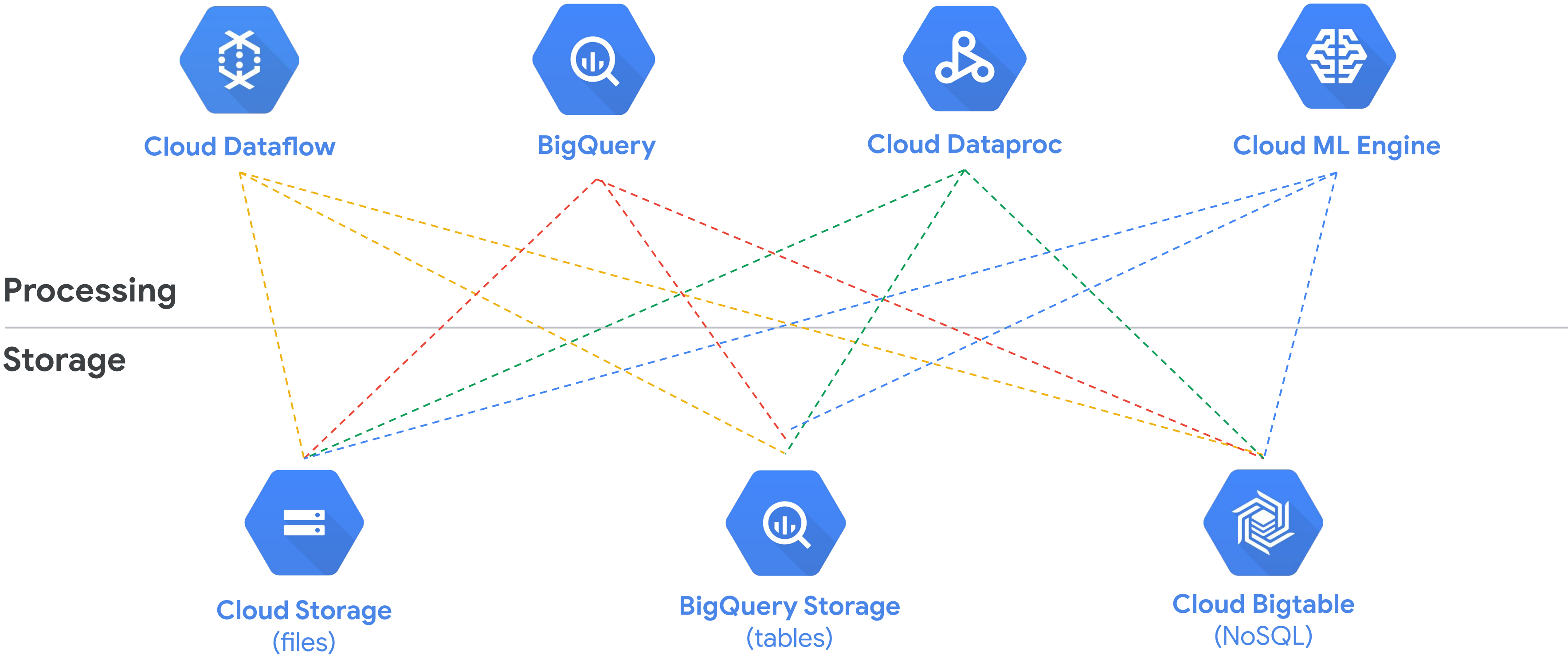
Faster performance for complex queries

Join and aggregate more data

Better scalability



Separation of storage and compute



Cloud Dataflow

BigQuery

Cloud Dataproc

Cloud ML Engine

Processing

Storage

Cloud Storage
(files)

BigQuery Storage
(tables)

Cloud Bigtable
(NoSQL)

BigQuery platform interoperability

BigQuery Storage API

Use BigQuery Storage like GCS for Dataflow and Dataproc, break down the Data Warehouse storage wall

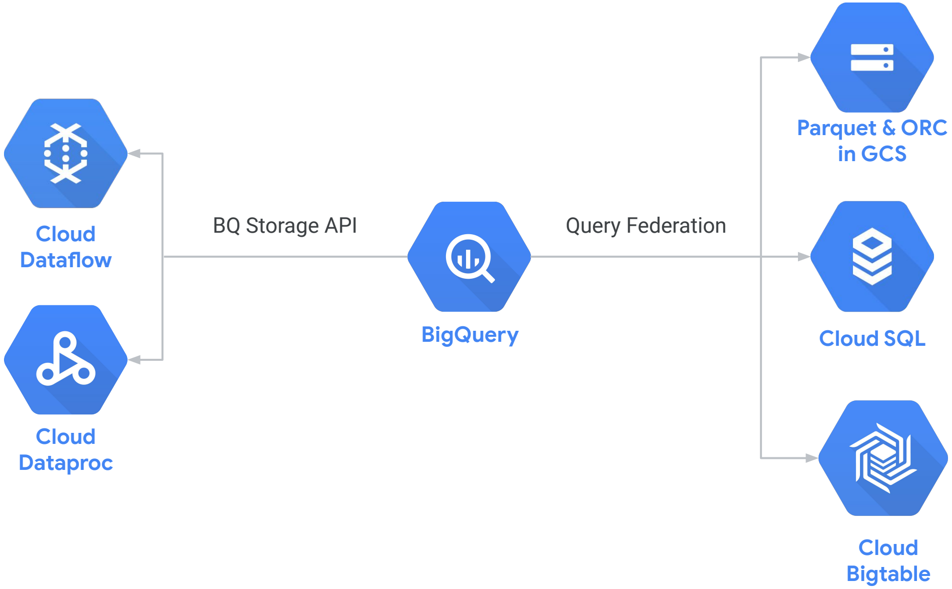
Run high-performance **dataframes** on BigQuery

Cloud SQL and Cloud Bigtable Federation

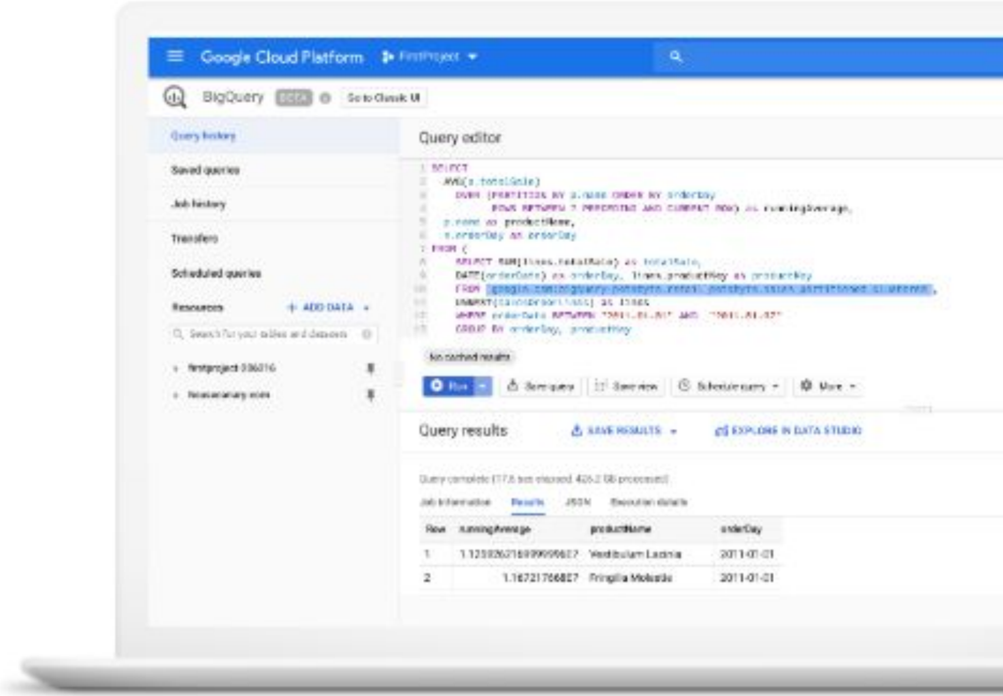
Query your Cloud SQL and Cloud Bigtable instances directly from BigQuery, without moving data around.

Parquet & ORC Federation

Query Parquet and ORC files directly in GCS



You don’t have to take my word for it



```
SELECT
  AVG(s.totalSale)
    OVER (PARTITION BY s.name ORDER BY orderDay
          ROWS BETWEEN 7 PRECEDING AND CURRENT ROW) as runningAverage,
  p.name as productName,
  s.orderDay as orderDay
FROM (
    SELECT SUM(lines.totalSale) as totalSale,
    DATE(orderDate) as orderDay, lines.productKey as productKey
    FROM [illegible],
    UNNEST(lineItems) as lines
    WHERE orderDate BETWEEN "2011-01-01" AND "2011-01-07"
    GROUP BY orderDay, productKey
```

Query complete (17.8 sec elapsed, 426.2 GB processed)

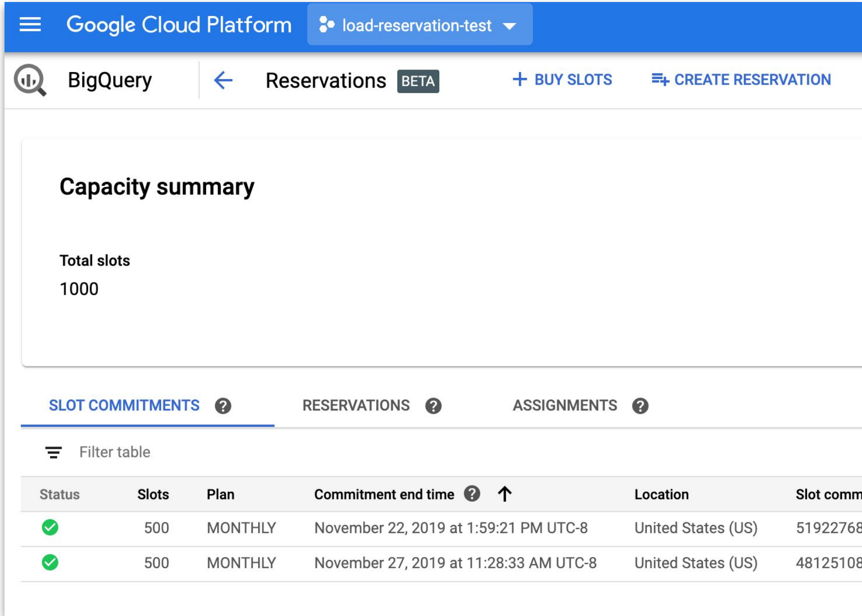
Row	runningAverage	productName	orderDay
1	1.1233262716999996E7	Vestibulum Lacinia	2011-01-01
2	1.16721766827	Fringilla Molestie	2011-01-01

Google Cloud

Enterprise-grade Workload management With Reservations

BigQuery Reservations allows customers to:

- Control flat-rate spend
- Buy slots in Web UI in seconds
- Efficiently manage workloads in BigQuery
- Automatically share any unused capacity



Capacity summary

Total slots
1000

Status	Slots	Plan	Commitment end time	Location	Slot comm
✓	500	MONTHLY	November 22, 2019 at 1:59:21 PM UTC-8	United States (US)	51922768
✓	500	MONTHLY	November 27, 2019 at 11:28:33 AM UTC-8	United States (US)	48125108

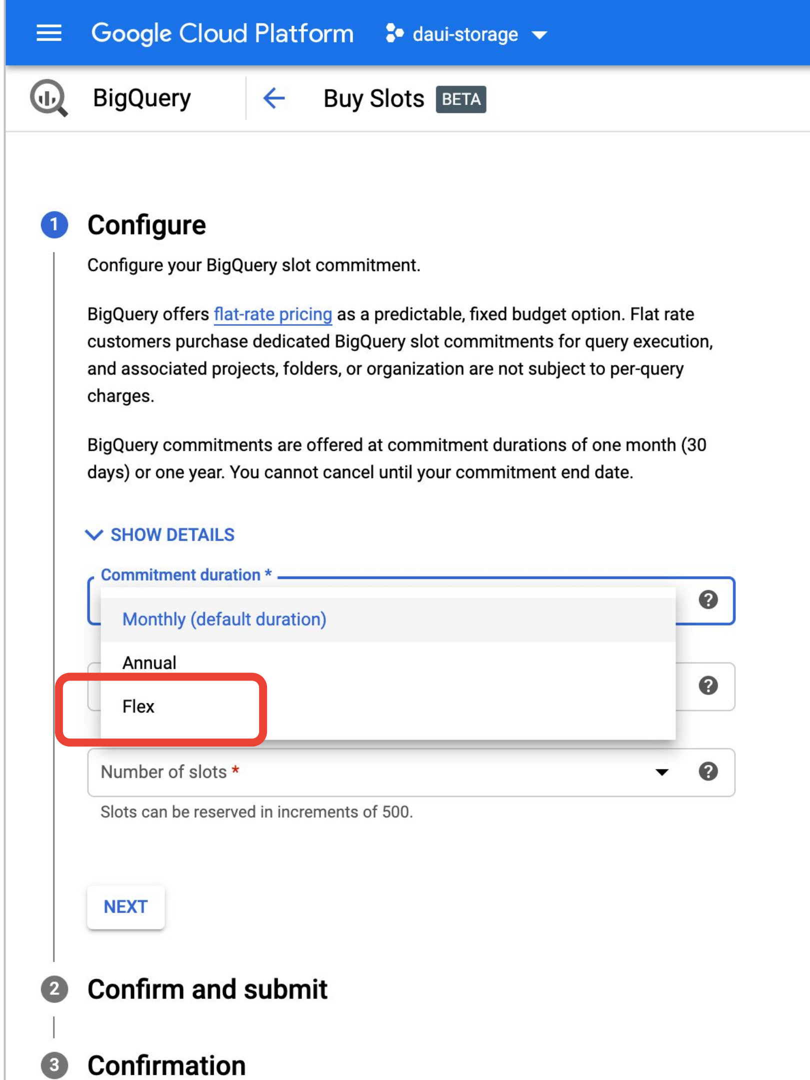
Google Cloud

Introducing Flex slots

- A new commitment type
  - Alongside monthly & annual
- Pricing
  - $30 per slot per month*
- More flexible
  - 60 second minimum
- Combine with monthly/annual
- Available in all BQ Reservations regions!
- Available in **BigQuery Reservations** today!

Google Cloud

*May vary per region



1 **Configure**

Configure your BigQuery slot commitment.

BigQuery offers flat-rate pricing as a predictable, fixed budget option. Flat rate customers purchase dedicated BigQuery slot commitments for query execution, and associated projects, folders, or organization are not subject to per-query charges.

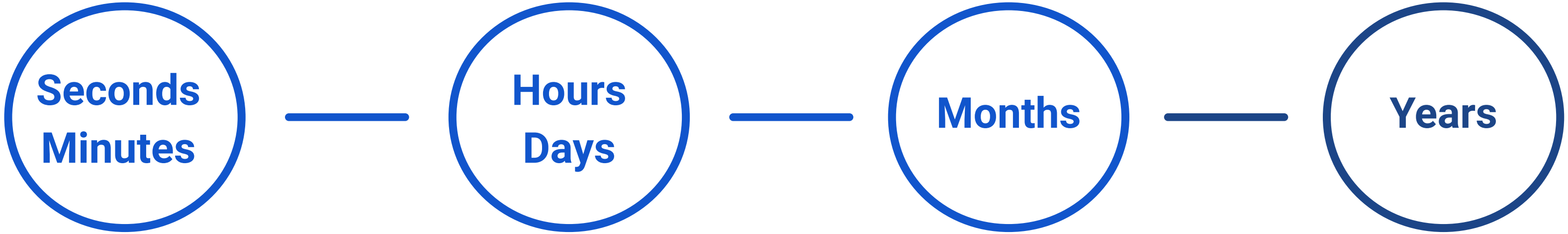
BigQuery commitments are offered at commitment durations of one month (30 days) or one year. You cannot cancel until your commitment end date.

Slots can be reserved in increments of 500.

2 **Confirm and submit**

3 **Confirmation**

BigQuery Commitment Types and Use Cases



Seconds
Minutes

Hours
Days

Months

Years

Flex Slots

Rapidly respond to business demands and Evaluate performance.

Flex Slots

Plan for business-critical calendar events.

Monthly Commitments

Ideal balance between flexibility and cost.

Annual Commitments

The most cost-effective option for steady-state workloads.

Google Cloud

BigQuery workload management

Customers can programmatically perform workload management using Reservations:

Create and delete reservations

Move projects between reservations

Move slots between reservations

Idle slots are seamlessly and automatically shared in real-time

Example

At 3am an important workload in project_d needs to run

At 3am we create a reservation

Move 1000 slots to the reservation

Move project_d into reservation

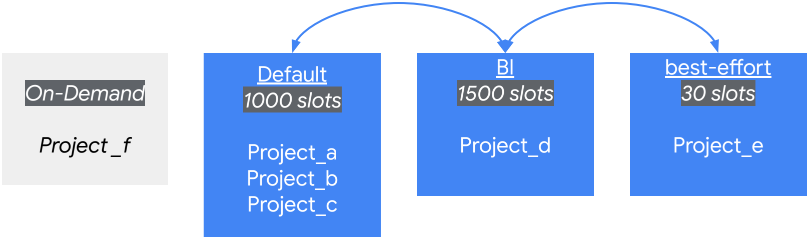
At 6am we delete the reservation

Move 1000 slots back

Move project_d back

Project_d was guaranteed 1000 slots 3am-6am

Idle slots seamlessly shared



On-Demand

Project _f

Default

1000 slots

Project_a

Project_b

Project_c

BI

1500 slots

Project_d

best-effort

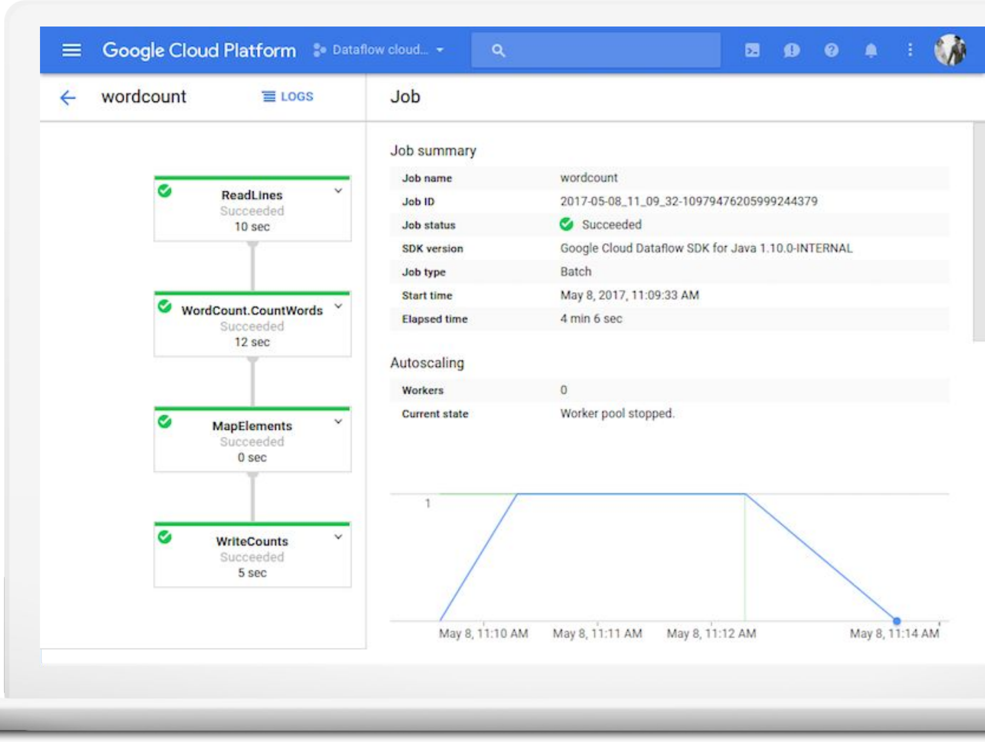
30 slots

Project_e

ETL / Scale streaming analytics pipelines with Cloud Dataflow

Streaming analytics service that minimizes processing time and cost with autoscaling while blending **batch and stream** processing.

- Fastest stream and batch processing on one service
- Lower TCO for streaming analytics
- Automatically burst resources when data spikes
- Build and monitor Apache Beam pipelines



Job

Job summary

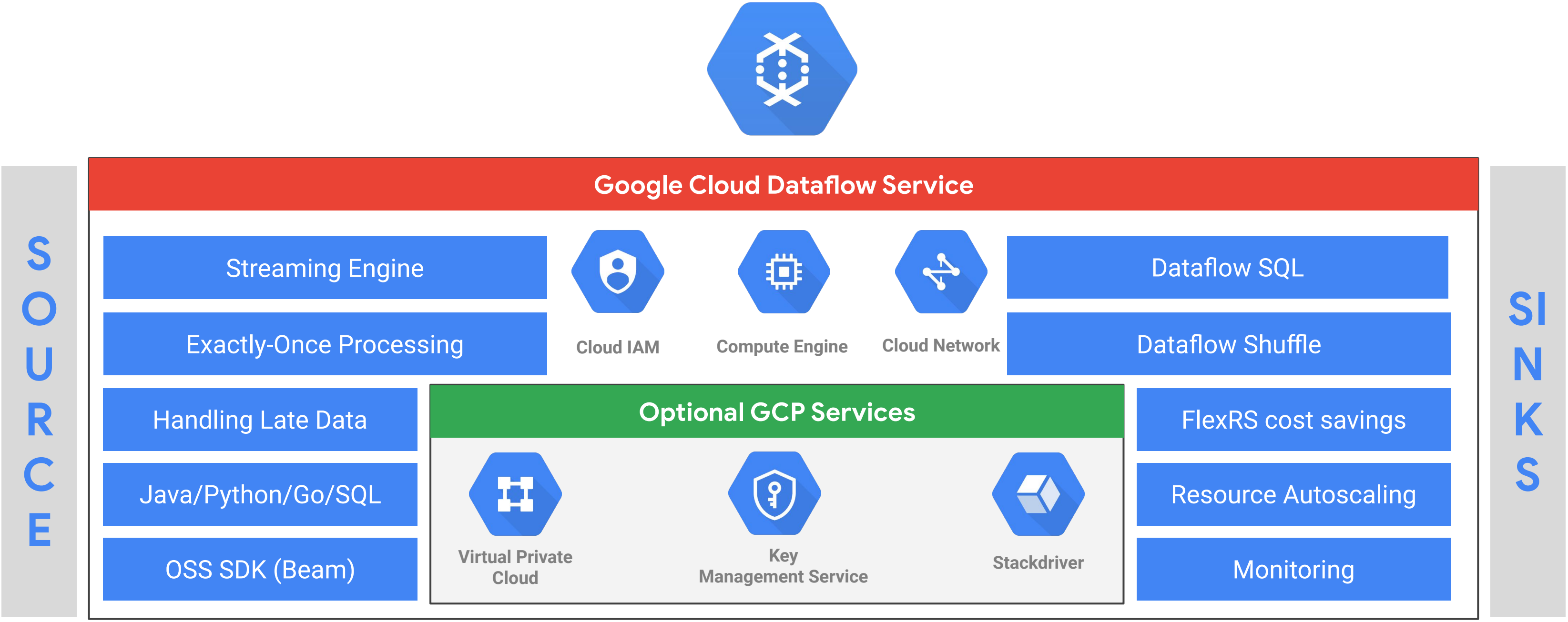
Job name	wordcount
Job ID	2017-05-08_11_09_32-10979476205999244379
Job status	✓ Succeeded
SDK version	Google Cloud Dataflow SDK for Java 1.10.0-INTERNAL
Job type	Batch
Start time	May 8, 2017, 11:09:33 AM
Elapsed time	4 min 6 sec

Autoscaling

Workers	0
Current state	Worker pool stopped.

Google Cloud

Dataflow: Stream Analytics as a managed service



SOURCE

Google Cloud Dataflow Service

Streaming Engine

Exactly-Once Processing

Cloud IAM

Compute Engine

Cloud Network

Dataflow SQL

Dataflow Shuffle

Handling Late Data

Java/Python/Go/SQL

OSS SDK (Beam)

Optional GCP Services

Virtual Private Cloud

Key Management Service

Stackdriver

FlexRS cost savings

Resource Autoscaling

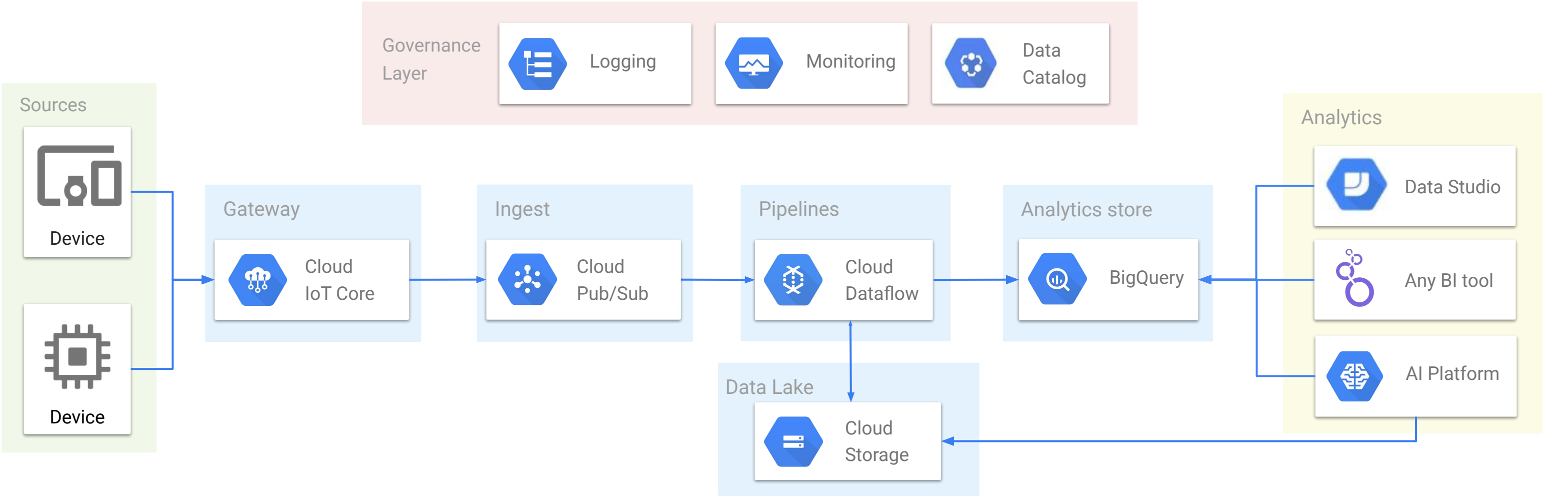
Monitoring

SINKS

Google Cloud

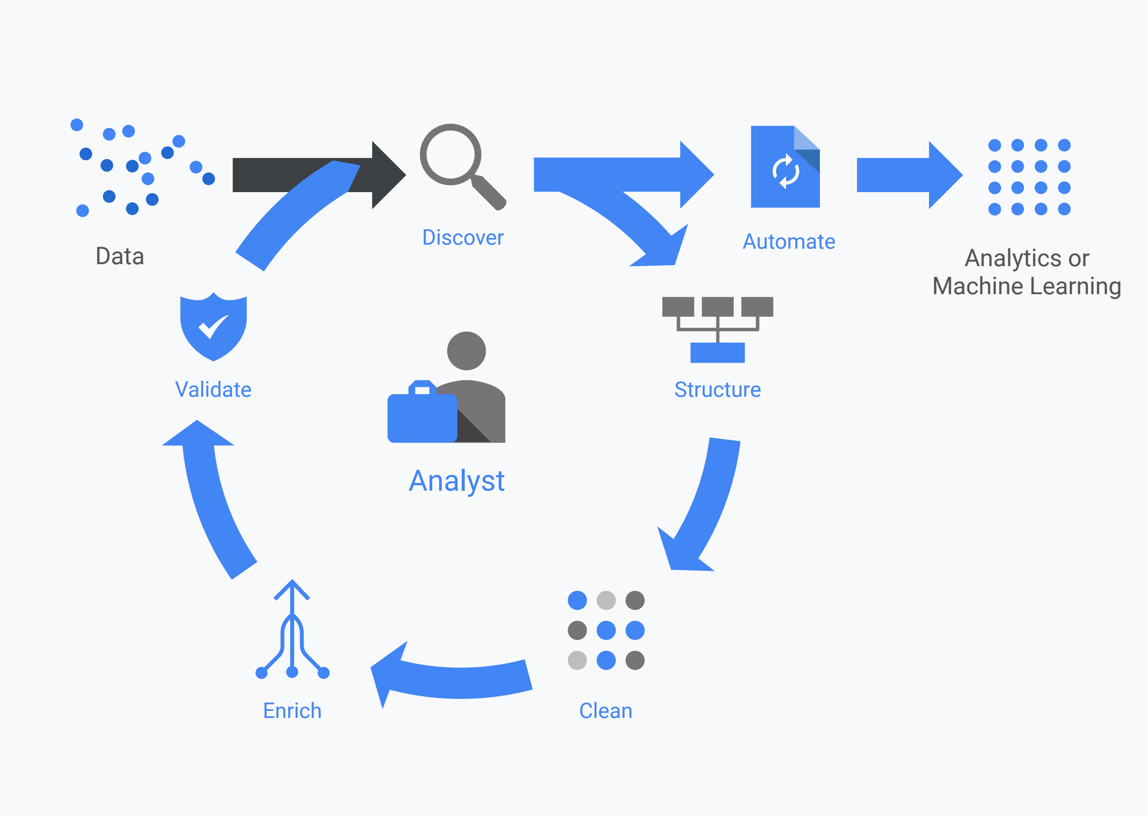
Demo: A simple Streaming reference architecture

Scales seamlessly to petabytes to let you focus on bringing actual value



Google Cloud

Simplify the data lifecycle with Cloud Dataprep



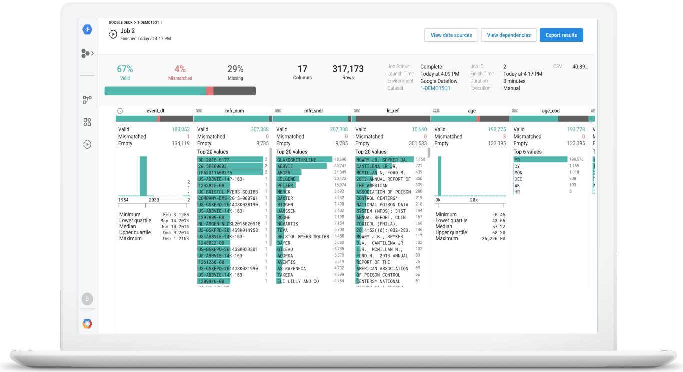
Serverless and cloud-native

Legacy data preparation

- ✗ Business users not empowered to transform data samples
- ✗ Must hire an IT/Data ops team and manage a Hadoop cluster
- ✗ Negotiate org-wide software licenses, arrange billing and manage seats
- ✗ Integrate application permissions with infrastructure permissions

Modern data preparation on Cloud Dataprep:

- ✓ Business users push the “Run Job” button to apply transformations to datasets of any size
- ✓ No need to create or manage infrastructure
- ✓ No need to provision software licenses
- ✓ Integrated, and highly scalable



GOOGLE DECK > 1-DEMO15Q1 >

Job 2
Finished Today at 4:17 PM

67% Valid | 4% Mismatched | 29% Missing | 17 Columns | 317,173 Rows

Job Status	Complete	Job ID	2	CSV	40.89...
Launch Time	Today at 4:09 PM	Finish Time	Today at 4:17 PM		
Environment	Google Dataflow	Duration	8 minutes		
Dataset	1-DEMO15Q1	Execution	Manual		

event_dt

Valid	183,053
Mismatched	1
Empty	134,119
Minimum	Feb 3 1955
Lower quartile	May 14 2013
Median	Jun 10 2014
Upper quartile	Dec 9 2014
Maximum	Dec 1 2103

mfr_num

Valid	307,388
Mismatched	0
Empty	9,785

Top 20 values

BD-2015-0177	2
2015FE00602	2
TPA2011A00275	2
US-ABBVIE-14P-163-	1
1232818-00	1
US-BRISTOL-MYERS SQUIBB	1
COMPANY-BMS-2015-000781	1
US-GSKPPD-2014GSK038190	1
US-ABBVIE-14K-163-	1
1297899-00	1
NL-AMGEN-NLDSL2015020918	1
US-GSKPPD-2014GSK014958	1
US-ABBVIE-14K-163-	1
1248022-00	1
US-GSKPPD-2014GSK023801	1
US-ABBVIE-14K-163-	1
1261266-00	1
US-GSKPPD-2014GSK021990	1
US-ABBVIE-14K-163-	1
1289916-00	1

mfr_sndr

Valid	307,388
Mismatched	0
Empty	9,785

Top 20 values

GLAXOSMITHKLINE	48,690
ABBVIE	43,747
AMGEN	21,849
CELGENE	20,123
PFIZER	16,974
MERCK	8,692
BAXTER	8,232
BIOGEN	7,498
JANSSEN	7,402
ROCHE	7,198
NOVARTIS	7,154
TEVA	6,700
BRISTOL MYERS SQUIBB	6,458
BAYER	6,060
GILEAD	5,735
ACORDA	5,570
AVENTIS	5,519
ASTRAZENECA	4,732
TAKEDA	4,399
ELI LILLY AND CO	4,284

lit_ref

Valid	15,640
Mismatched	0
Empty	301,533

Top 20 values

MOWRY JB, SPYKER DA,	1,158
CANTILENA LR JR,	721
MCMILLAN N, FORD M.	439
2013 ANNUAL REPORT OF	350
THE AMERICAN	309
ASSOCIATION OF POISON	280
CONTROL CENTERS^	219
NATIONAL POISON DATA	218
SYSTEM (NPDS): 31ST	194
ANNUAL REPORT. CLIN	167
TOXICOL (PHILA).	166
2014;52(10):1032-283.	146
MOWRY J.B., SPYKER	117
D.A., CANTILENA JR	102
L.R., MCMILLAN N.,	102
FORD M.. 2013 ANNUAL	83
REPORT OF THE	75
AMERICAN ASSOCIATION	69
OF POISON CONTROL	66
CENTERS^ NATIONAL	61

age

Valid	193,775
Mismatched	3
Empty	123,395
Minimum	-0.45
Lower quartile	43.65
Median	57.22
Upper quartile	68.20
Maximum	36,226.00

age_cod

Valid	193,778
Mismatched	0
Empty	123,395

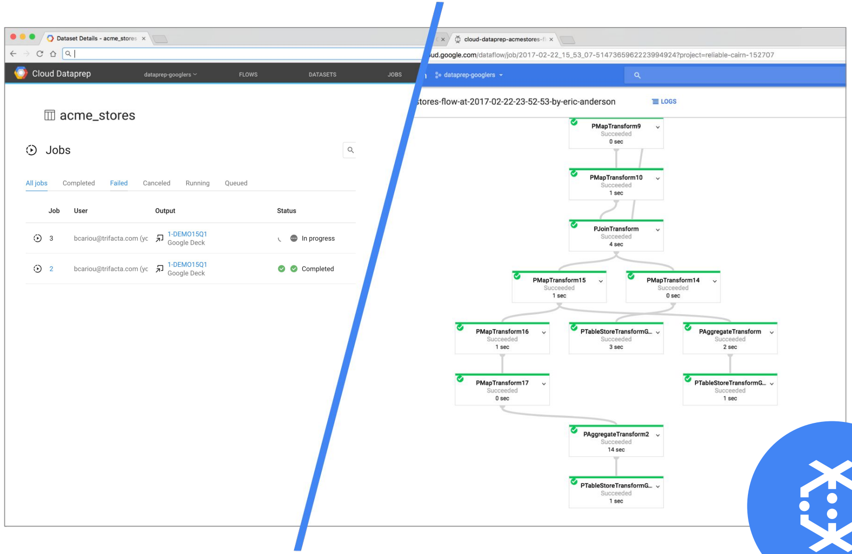
Top 6 values

YR	190,576
DY	1,165
MON	1,018
DEC	908
WK	103
HR	8

Serverless Simplicity	Fast Exploration	Easy Preparation

Powerful & easy processing with Cloud Dataflow under the hood

- ✓ Process diverse datasets - structured or unstructured
- ✓ Prepare datasets of any size, PB or MB, with equal ease
- ✓ Leverages Cloud Dataflow without needing to write any scripts
- ✓ Auto-scalable and can easily handle processing massive data sets


Dataset Details - acme_stores
cloud-dataprep-acmestores-fl
ud.google.com/dataflow/job/2017-02-22_15_53_07-5147365962223994924?project=reliable-cairn-152707
Cloud Dataprep
dataprep-googlers
FLOWS
DATASETS
JOBS
dataprep-googlers
LOGS

acme_stores

Jobs

All jobs · Completed · Failed · Canceled · Running · Queued

Job	User	Output	Status
3	bcariou@trifacta.com (yc	1-DEMO15Q1 Google Deck	In progress
2	bcariou@trifacta.com (yc	1-DEMO15Q1 Google Deck	✓ Completed

stores-flow-at-2017-02-22-23-52-53-by-eric-anderson

Serverless Simplicity	Fast Exploration	**Easy Preparation**

Proprietary + Confidential

Supports common data types of any size

Sources

BigQuery tables

Cloud Storage or local upload using common file formats:

- CSV
- JSON
- TXT
- LOG
- GZIP
- BZIP

Targets

BigQuery tables

Cloud Storage:

- CSV (compressed or not)
- JSON (compressed or not)
- Avro

Serverless Simplicity	Fast Exploration	**Easy Preparation**

CRUNCH
druid
Flink
hadoop
APACHE HBASE
HIVE
jupyter
oozie
presto
APACHE Spark

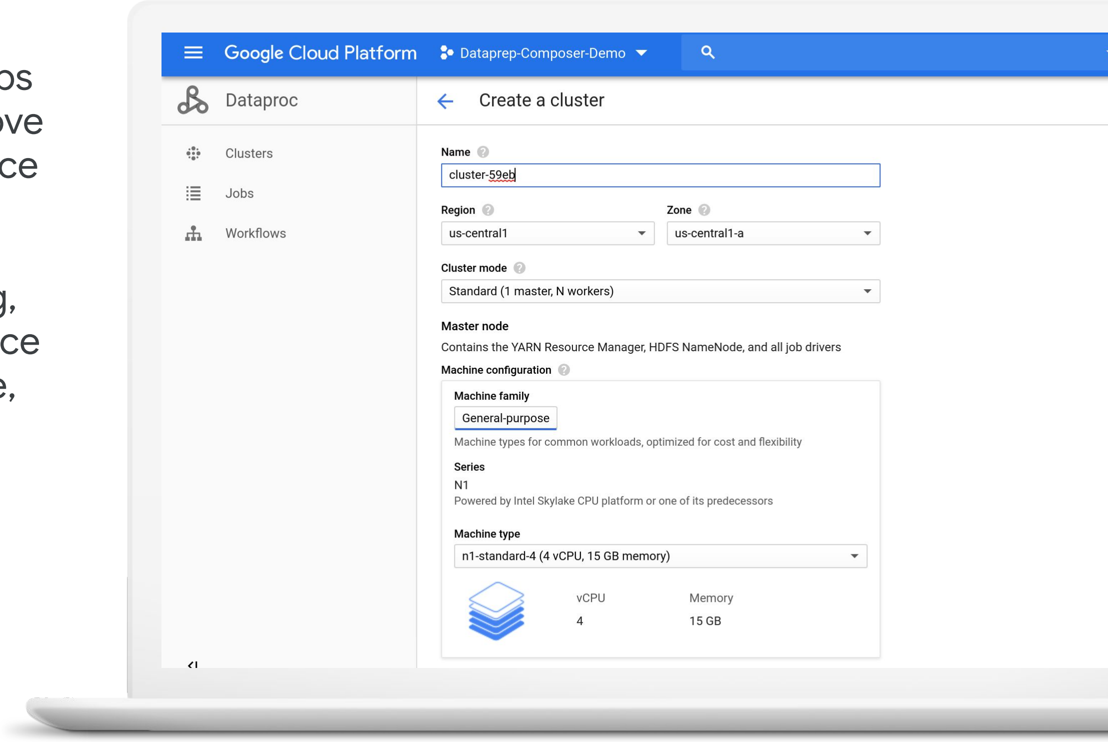
Cloud Dataproc

Combining the best of open source and cloud.

Open source data and analytics processing at scale on Cloud Dataproc

Build data and analytics processing jobs using the open source software you love with the scale, security, and governance of the cloud.

- Autoscale SQL, batch, streaming, and machine learning open source processing (Apache MapReduce, Apache Spark, Presto, etc.)
- Lower TCO of running OSS
- Build Spark jobs on Kubernetes



Name: cluster-59eb

Region: us-central1

Zone: us-central1-a

Cluster mode: Standard (1 master, N workers)

Master node

Contains the YARN Resource Manager, HDFS NameNode, and all job drivers

Machine configuration

Machine family: General-purpose

Machine types for common workloads, optimized for cost and flexibility

Series: N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type: n1-standard-4 (4 vCPU, 15 GB memory)

vCPU	Memory
4	15 GB

Google Cloud

The benefits of Hadoop/Spark on Cloud

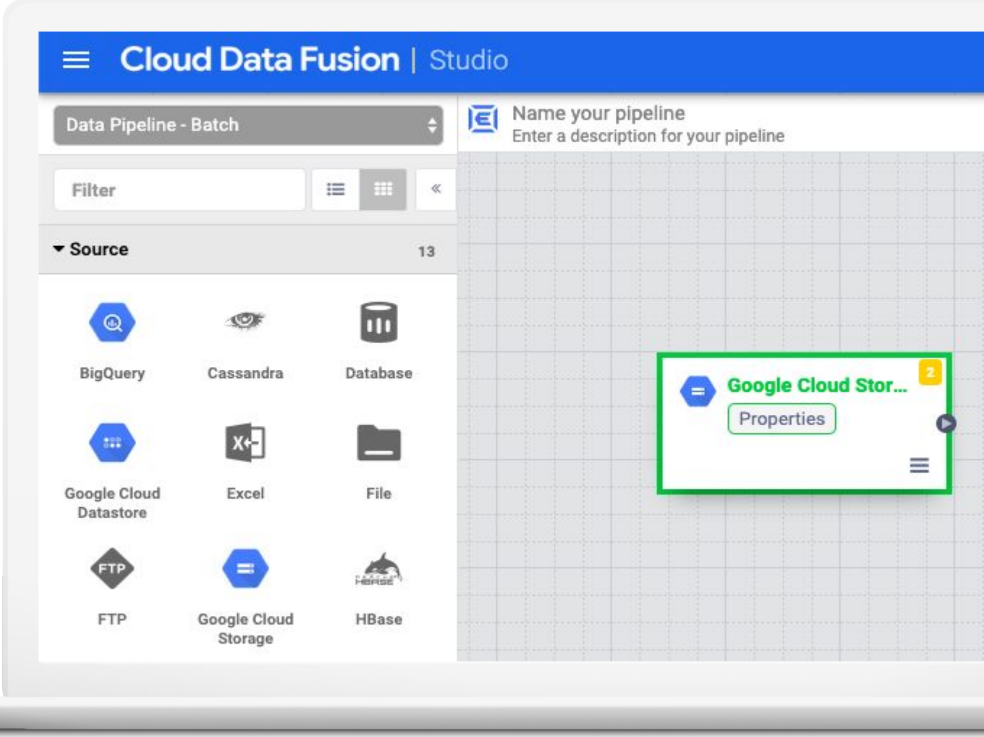
On premises	On compute engine	Cloud Dataproc
🟩 Custom code	🟩 Custom code	🟩 Custom code
🟩 Monitoring/Health	🟩 Monitoring/Health	🟦 Monitoring/Health
🟩 Dev integration	🟩 Dev integration	🟦 Dev integration
🟩 Scaling	🟩 Scaling	🟦 Scaling
🟩 Job submission	🟩 Job submission	🟦 Job submission
🟩 GCP connectivity	🟦 GCP connectivity	🟦 GCP connectivity
🟩 Deployment	🟦 Deployment	🟦 Deployment
🟩 Creation	🟦 Creation	🟦 Creation

🟩 Self-managed 🟦 Google managed

Build code free data pipelines with Data Fusion

Cloud Data Fusion is a fully managed, cloud-native data integration service that helps users efficiently build and manage ETL/ELT data pipelines.

- Use pre build open source library of connectors
- Execute data pipelines in Apache Spark
- Metadata and lineage integrations
- Build Apache Kakfa pipelines



Google Cloud

That’s a wrap.

pascalr@google.com

Google Cloud