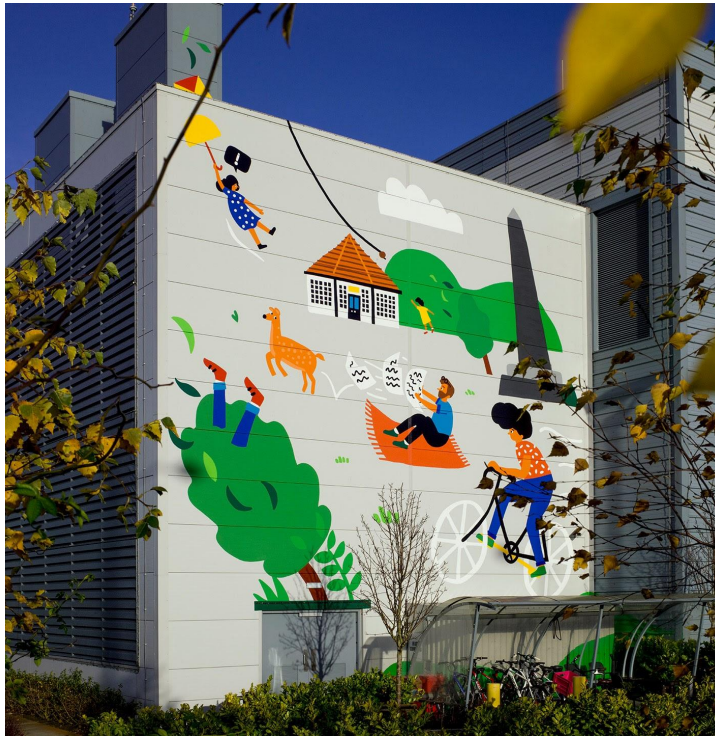
How is Google Cloud Platform Different

Bruno Basto
Customer Engineer, Data Analytics
basto@google.com

Google Cloud



Today's Agenda

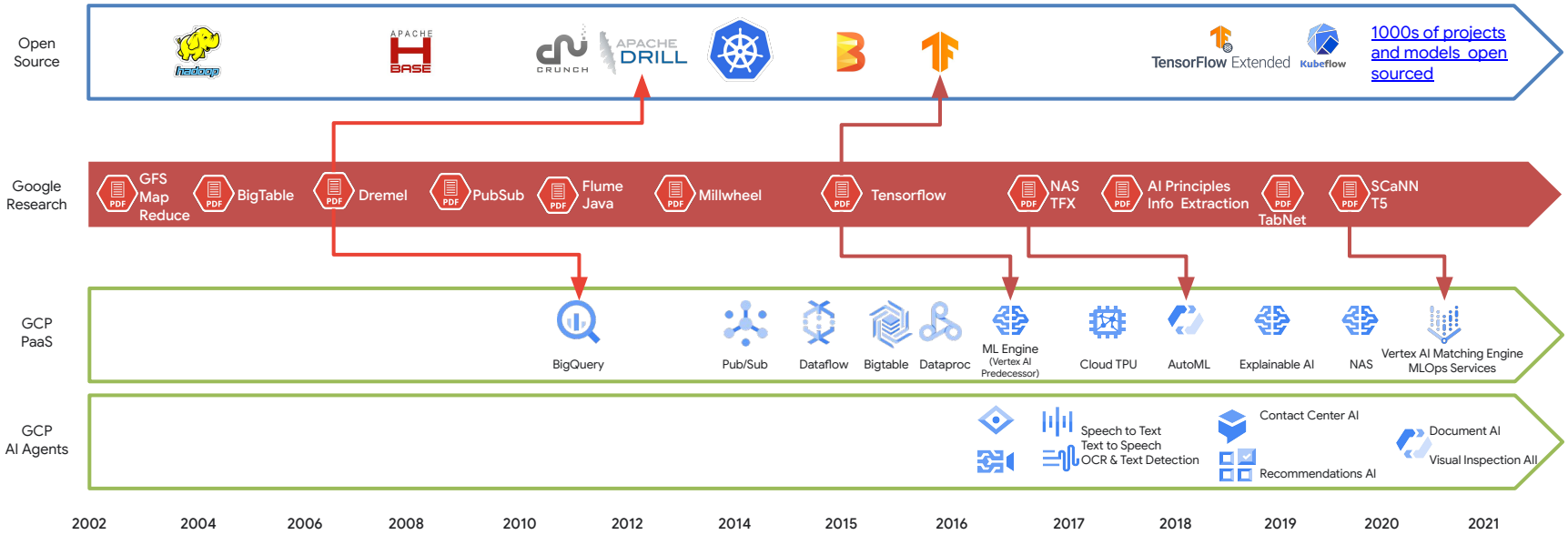
GCP DA
Databricks Timeline
Databricks Architecture
Databricks Deployment Model
@GCP Architecture Reference
DBX and BQ and
why hybrid make sense

State-of-the-Art AI

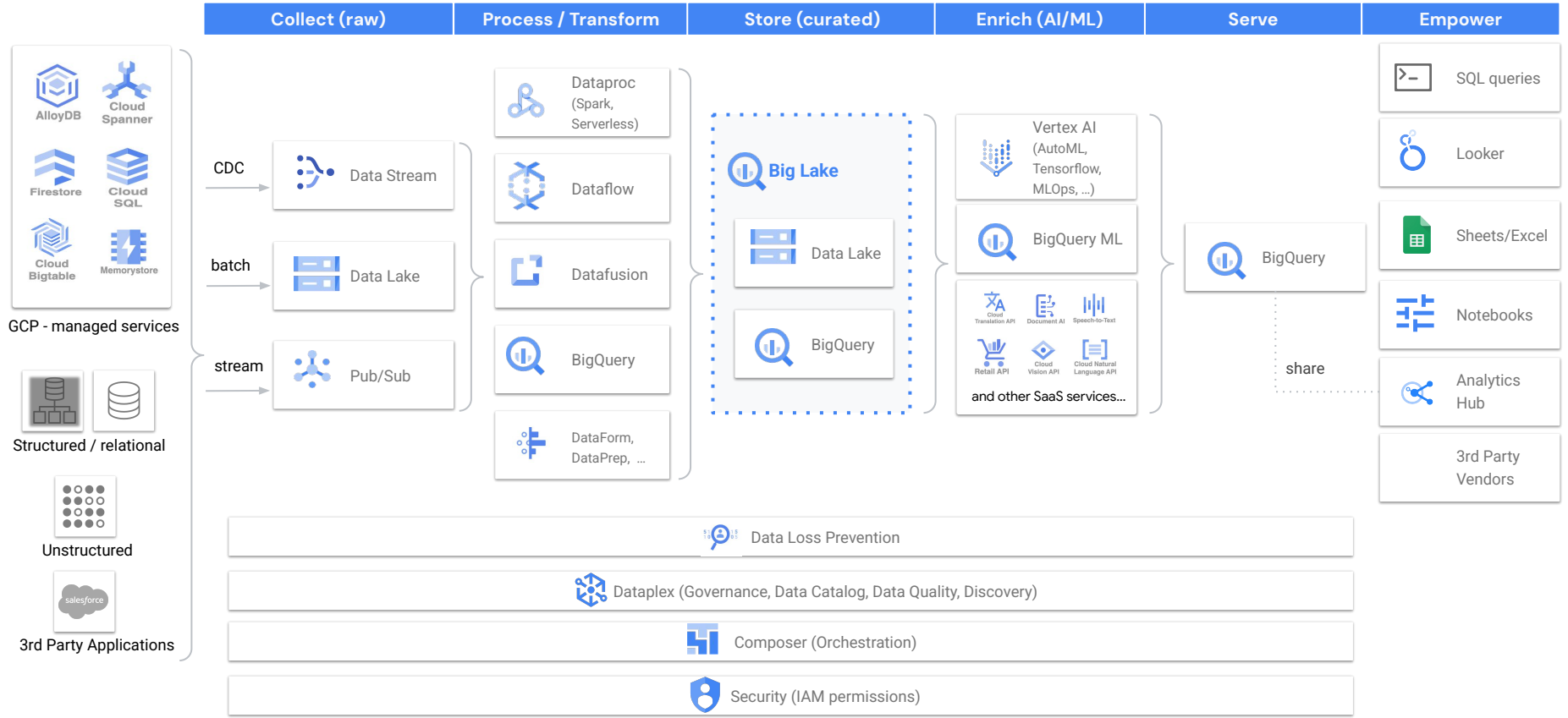
Building the next generation of Data to AI, E2E MLOps, and Scalable Infrastructure

3,000+ Google / Deepmind researchers and 7,000+ published publications drive a product pipeline

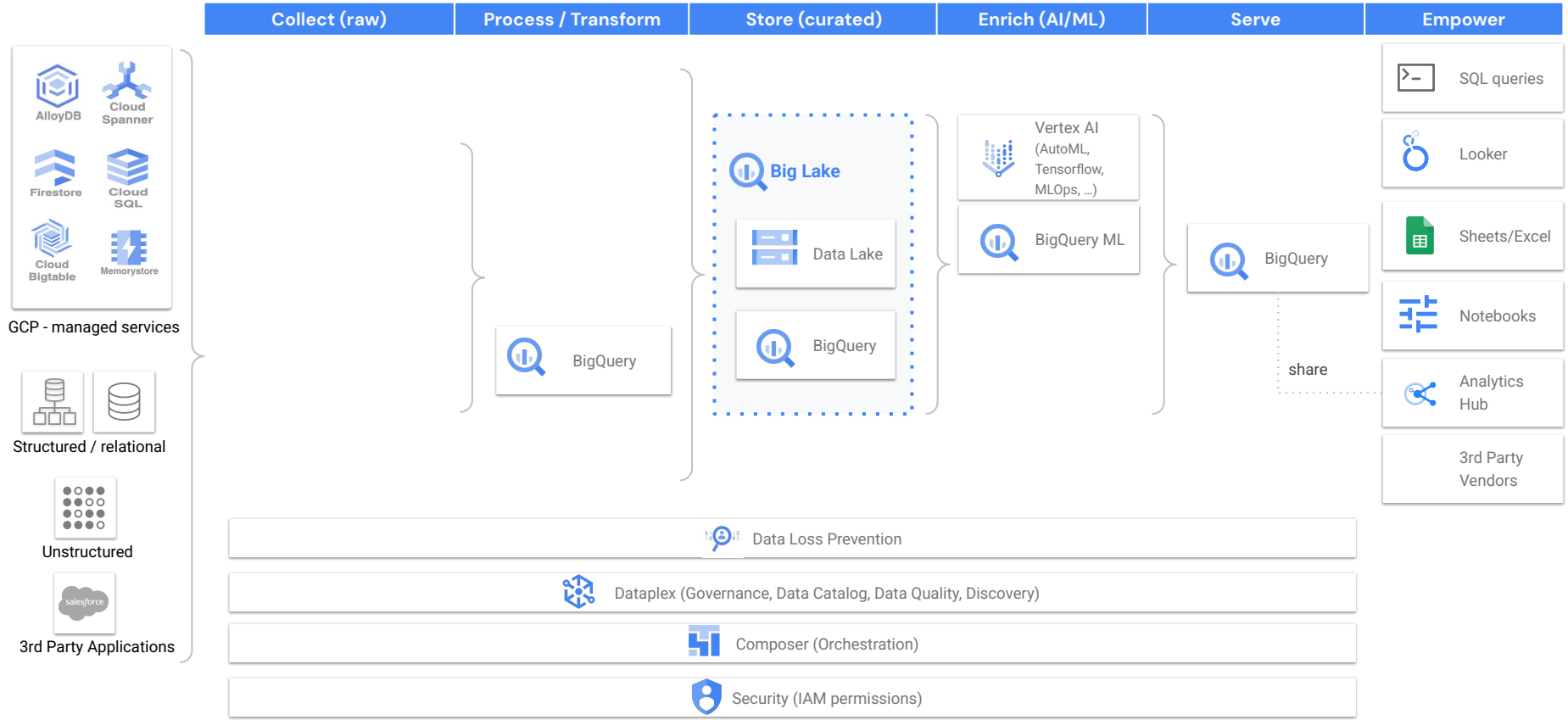
that results in **better product performance** and **lower costs** to reach that performance for Vertex AI customers.



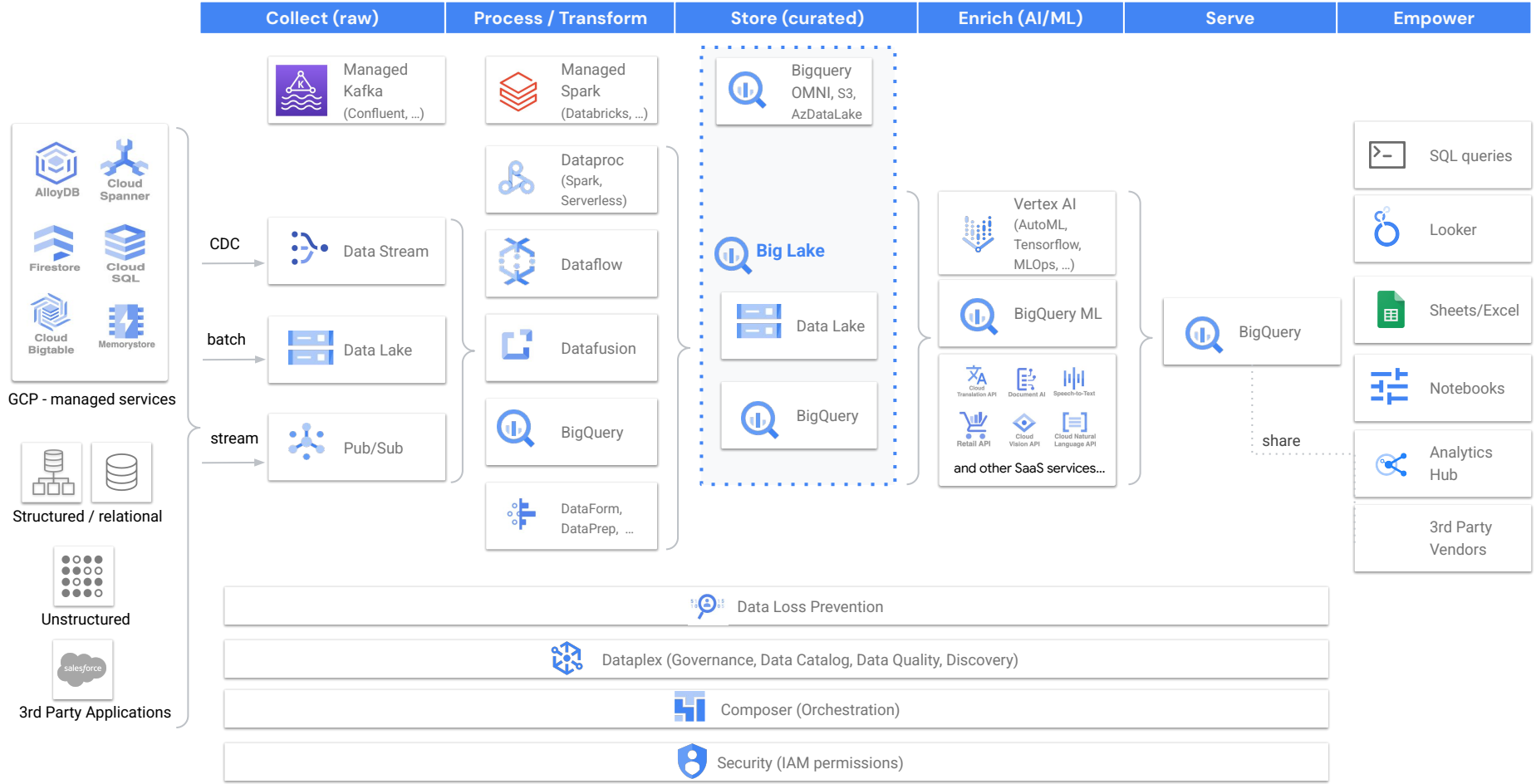
GCP main data services

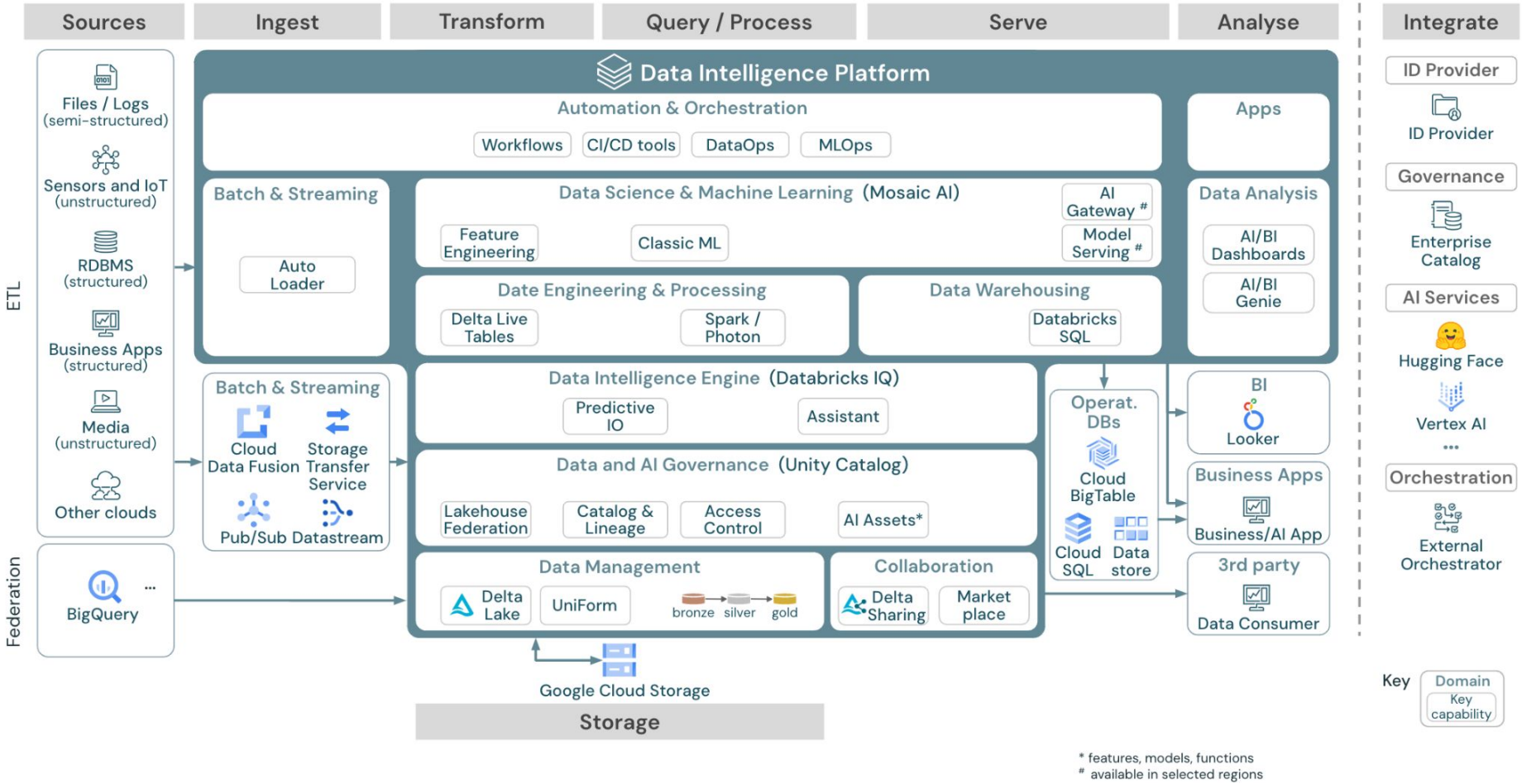


GCP main data services

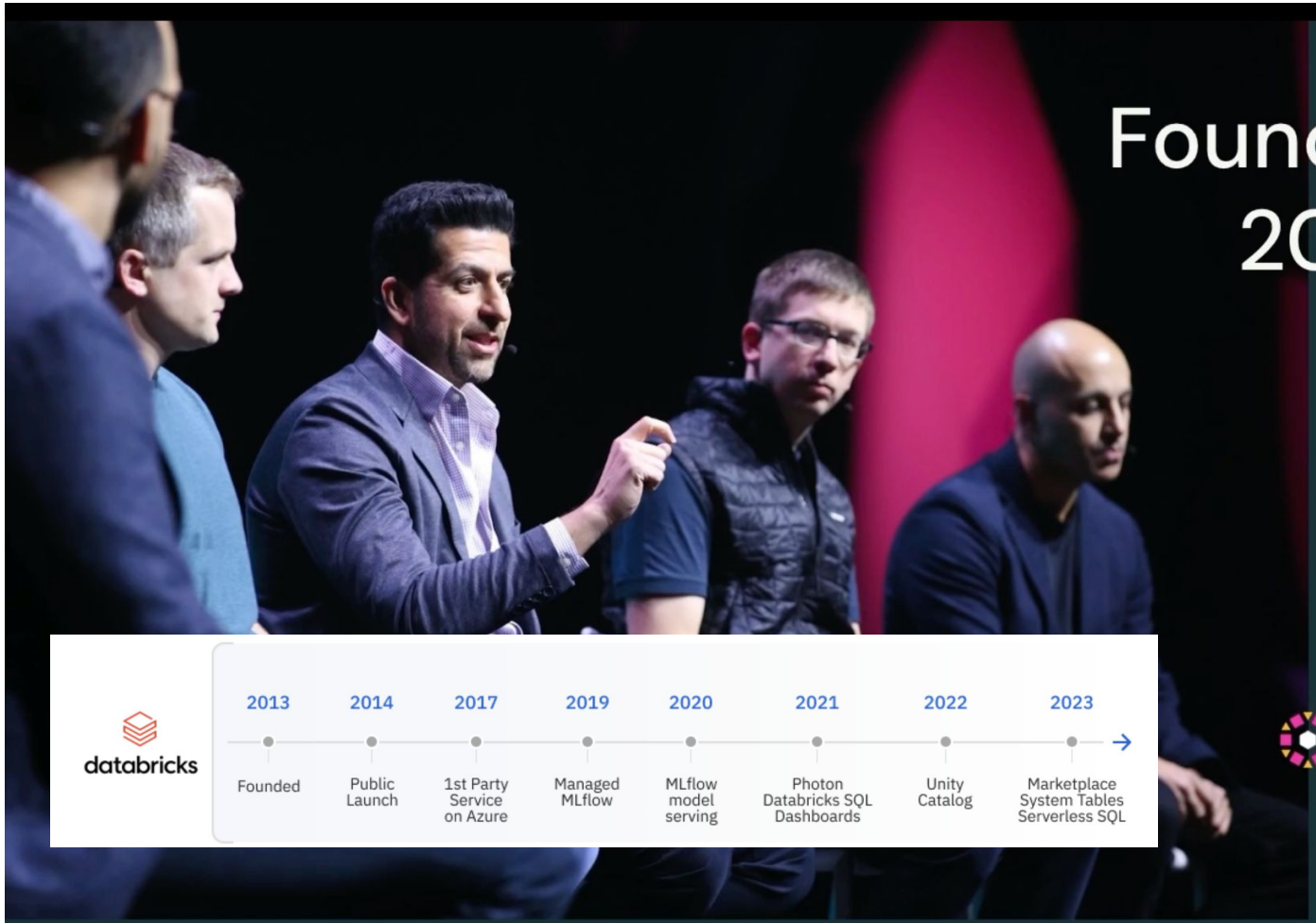


GCP - main data services + partner services (exemple)





[Databricks GCP Ref Arch]



Founded in 2013

DELTA LAKE

mlflow

APACHE Spark

Unity Catalog

DBRX

databricks

2013	2014	2017	2019	2020	2021	2022	2023
Founded	Public Launch	1st Party Service on Azure	Managed MLflow	MLflow model serving	Photon Databricks SQL Dashboards	Unity Catalog	Marketplace System Tables Serverless SQL

Strengths

- DA Centric
- Open Source Flavour
  - Delta (parquet+metadata)
    DLT are proprietary
  - MLFlow
- Developer Experience
- Single Copy of Data
- Multi-Cloud Support
- Integrated Platform

Main Competitors

- Snowflake
- Cloud Hyperscalers (Az/AWS)

Cons

- Lakehouse: don’t compete with DW
- Unity Catalog: is a must for many functionalities, management overhead, lockin
- DatabricksSQL: expensive and/or poor performant and/or hard to scale, complex to maintain and debug
- DLT is proprietary, data cannot be accessed outside o Dbx
- Migration between clouds is like between DataProc and Dbx
- …

Introducing DBRX: A New State-of-the-Art Open LLM

by The Mosaic Research Team

March 27, 2024 in Mosaic AI Research

Today, we are excited to introduce DBRX, an open, general-purpose LLM created by Databricks. Across a range of standard benchmarks, DBRX sets a new state-of-the-art for established open LLMs. Moreover, it provides the open community and enterprises building their own LLMs with capabilities that were previously limited to closed model APIs; according to our measurements, it surpasses GPT-3.5, and it is competitive with Gemini 1.0 Pro. It is an especially capable code model, surpassing specialized models like CodeLLaMA-70B on programming, in addition to its strength as a general-purpose LLM.

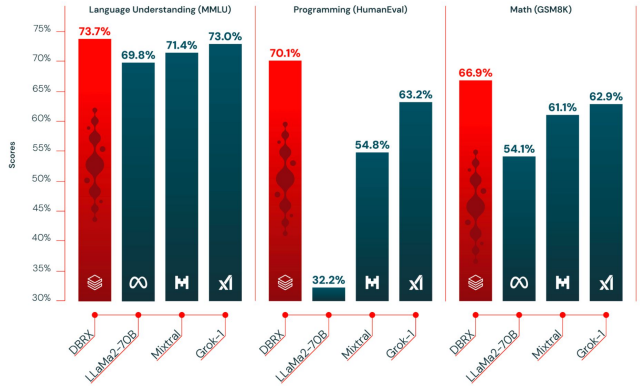


Figure 1: DBRX outperforms established open source models on language understanding (MMLU), Programming (HumanEval), and Math (GSM8K).

Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

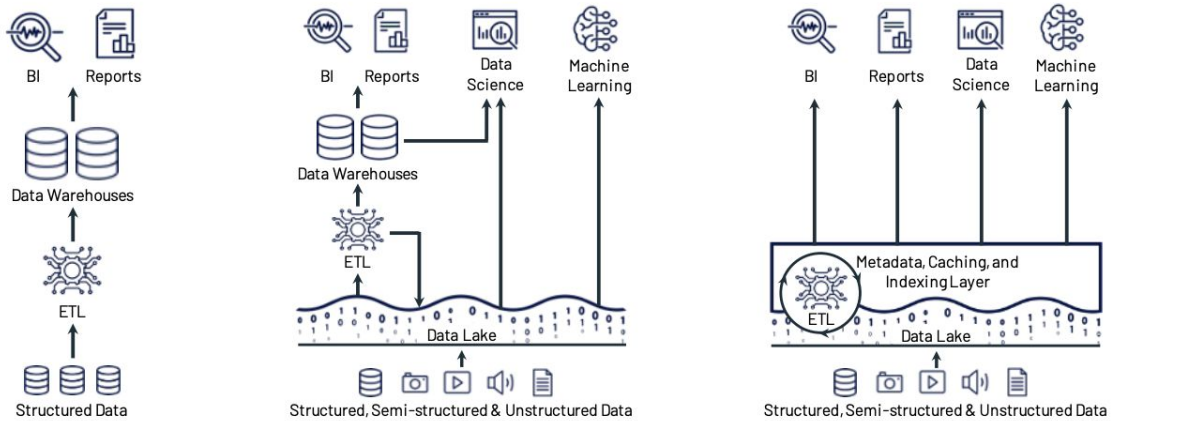
Michael Armbrust[1], Ali Ghodsi[1,2], Reynold Xin[1], Matei Zaharia[1,3]
[1]Databricks, [2]UC Berkeley, [3]Stanford University

Abstract

This paper argues that the data warehouse architecture as we know it today will wither in the coming years and be replaced by a new architectural pattern, the Lakehouse, which will (i) be based on open direct-access data formats, such as Apache Parquet, (ii) have first-

quality and governance downstream. In this architecture, a small subset of data in the lake would later be ETLed to a downstream data warehouse (such as Teradata) for the most important decision support and BI applications. The use of open formats also made data lake data directly accessible to a wide range of other analytics engines, such as machine learning systems [30, 37, 42].

From 2015 onwards, cloud data lakes, such as S3, ADLS and GCS,



(a) First-generation platforms. **(b) Current two-tier architectures.** **(c) Lakehouse platforms.**

Figure 1: Evolution of data platform architectures to today's two-tier model (a-b) and the new Lakehouse model (c).

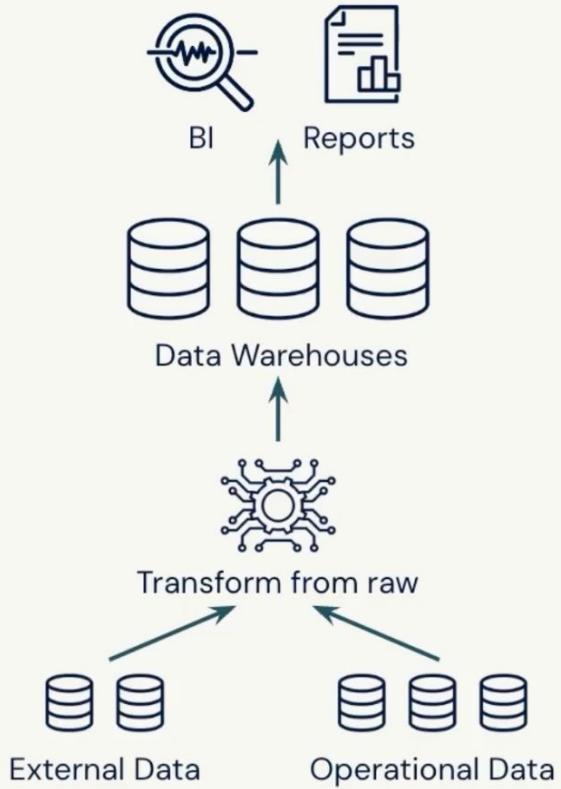
Data Warehouses

Strengths

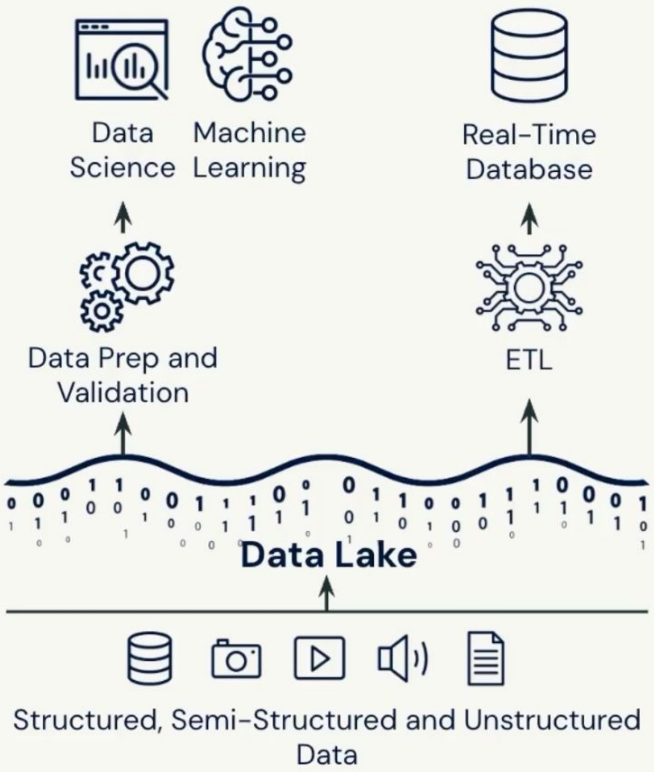
- Purpose-built for BI and reporting
- Meant to unify disparate systems
- Houses structured, clean data with a standardized schema

Downsides

- No support for semi or unstructured data
- Poor support for data science, AI and streaming use cases
- Uses closed and proprietary formats
- Expensive to scale



Data Lakes



Strengths

- Store any kind of data
- Inexpensive storage
- Good starting point
- Support for GenAI and streaming use cases

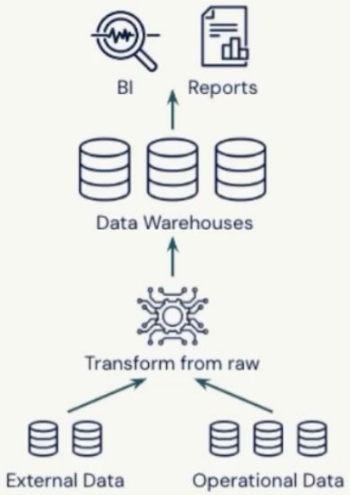
Downsides

- Complex to set up
- Poor BI performance
- Can become unreliable data swamps
- Governance concerns
- Warehouses still needed

Data Lakehouse

One platform to unify all your data, analytics, and AI workloads

Data Warehouses



Lakehouse

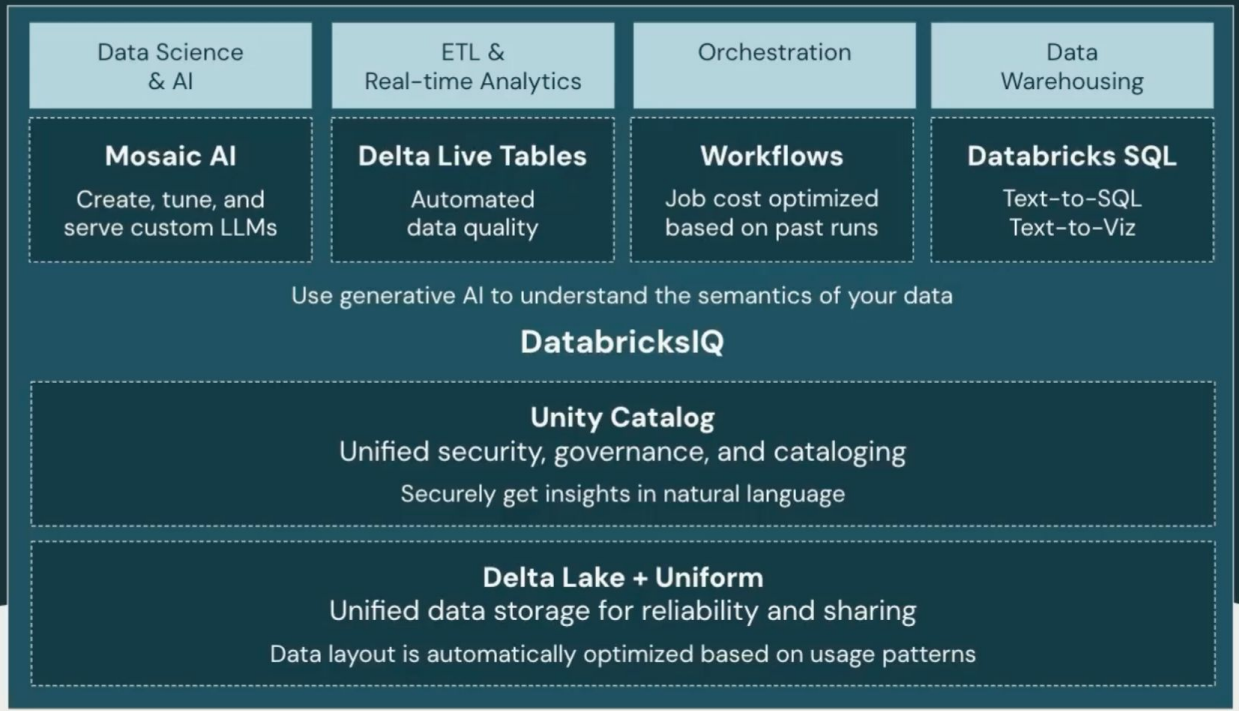
All machine learning, SQL, BI, and streaming use cases

One security and governance approach for all data assets on all clouds

An open and reliable data platform to efficiently handle all data types

Data Lakes

Databricks Data Intelligence Platform

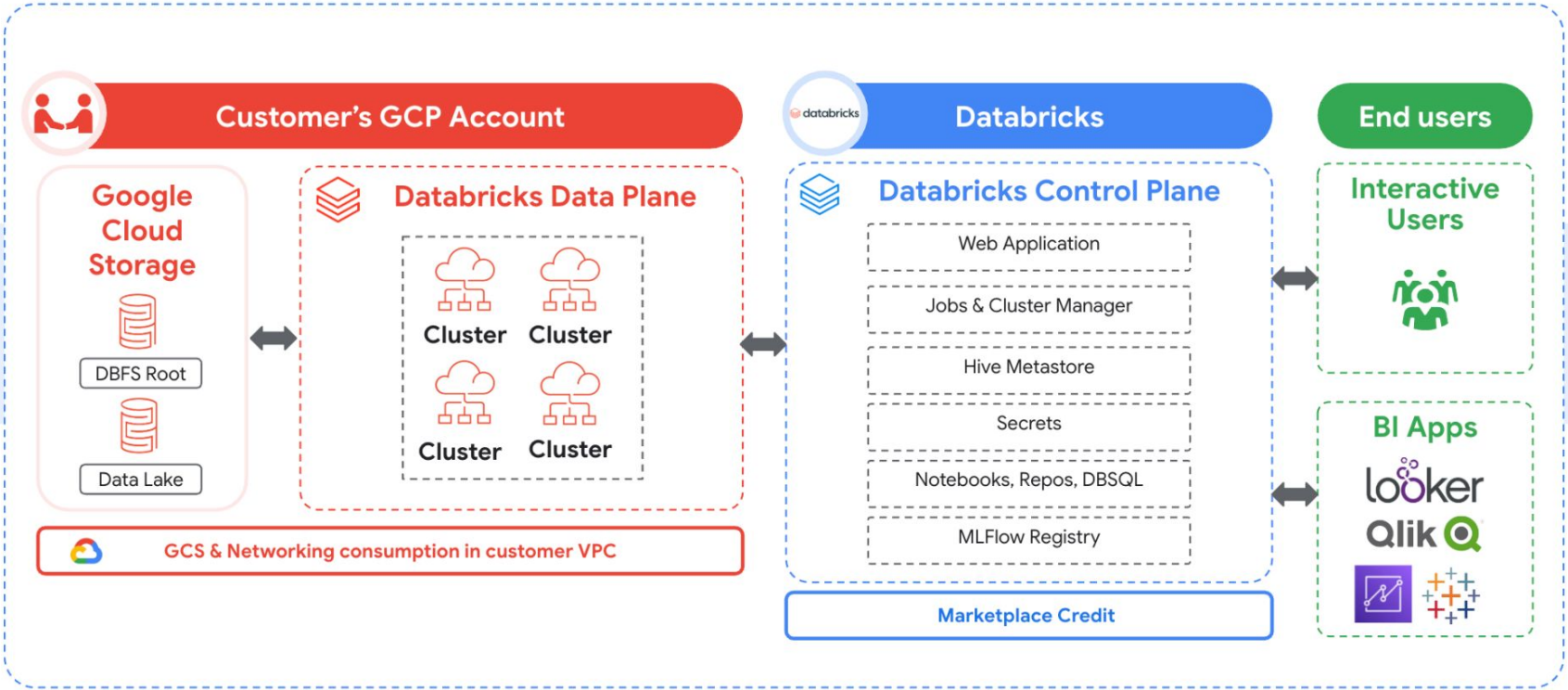


Open Data Lake

All Raw Data
(Logs, Texts, Audio, Video, Images)

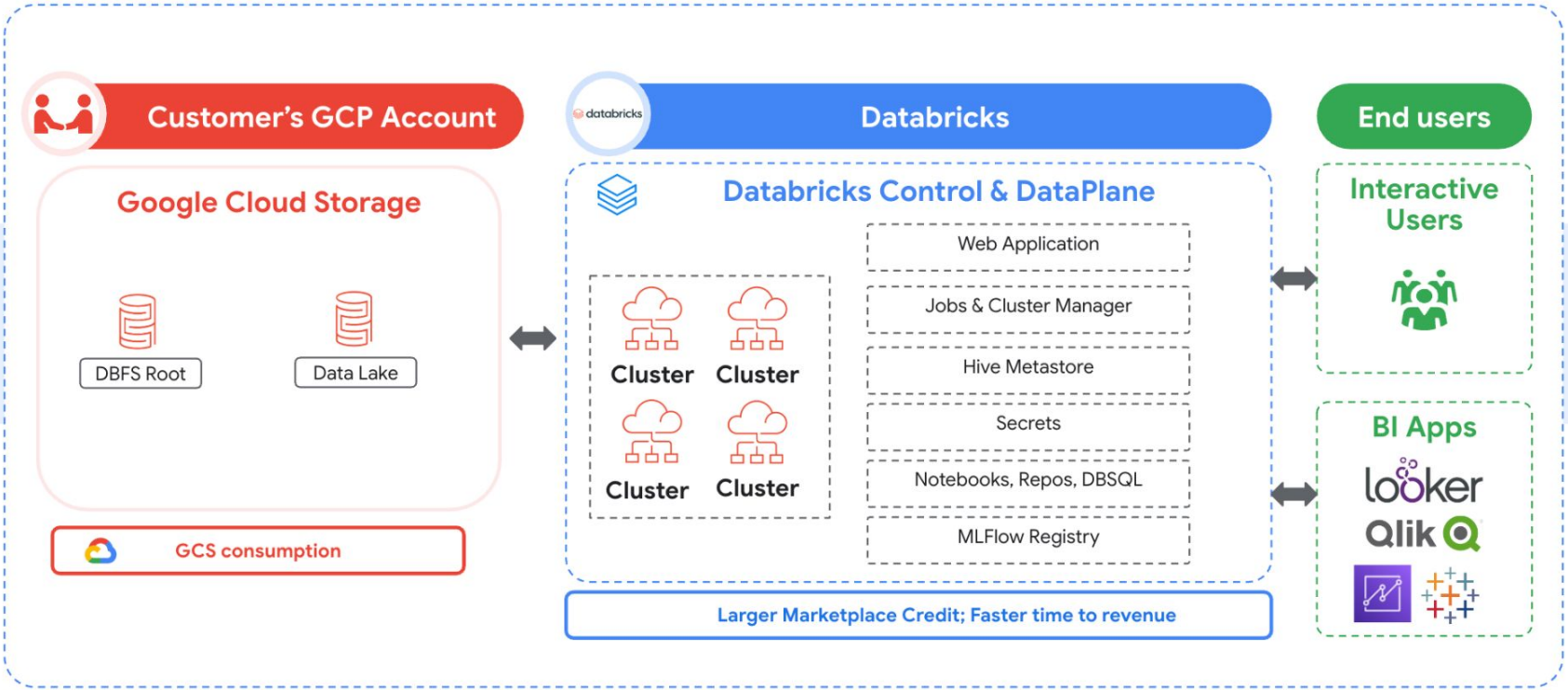
Databricks Deployment Model (Hybrid SaaS)

Databricks uses GCP compute and storage in Customer GCP projects



Databricks Deployment Model (Serverless)

Databricks uses storage in Customer GCP projects



Databricks Compute

Runtimes

Standard

Apache Spark and many other components and updates to provide an optimized big data analytics experiences.

Machine Learning

Adds popular machine learning libraries like TensorFlow, Keras, PyTorch, and XGBoost.

Specialized Compute

SQL Warehouses

Specifically designed for the optimization of SQL BI workloads with built in optimization for best price/performance.

Enhanced with Databricks Photon

Delta Lake Overview

Key Features


ACID Transactions


DML Operations

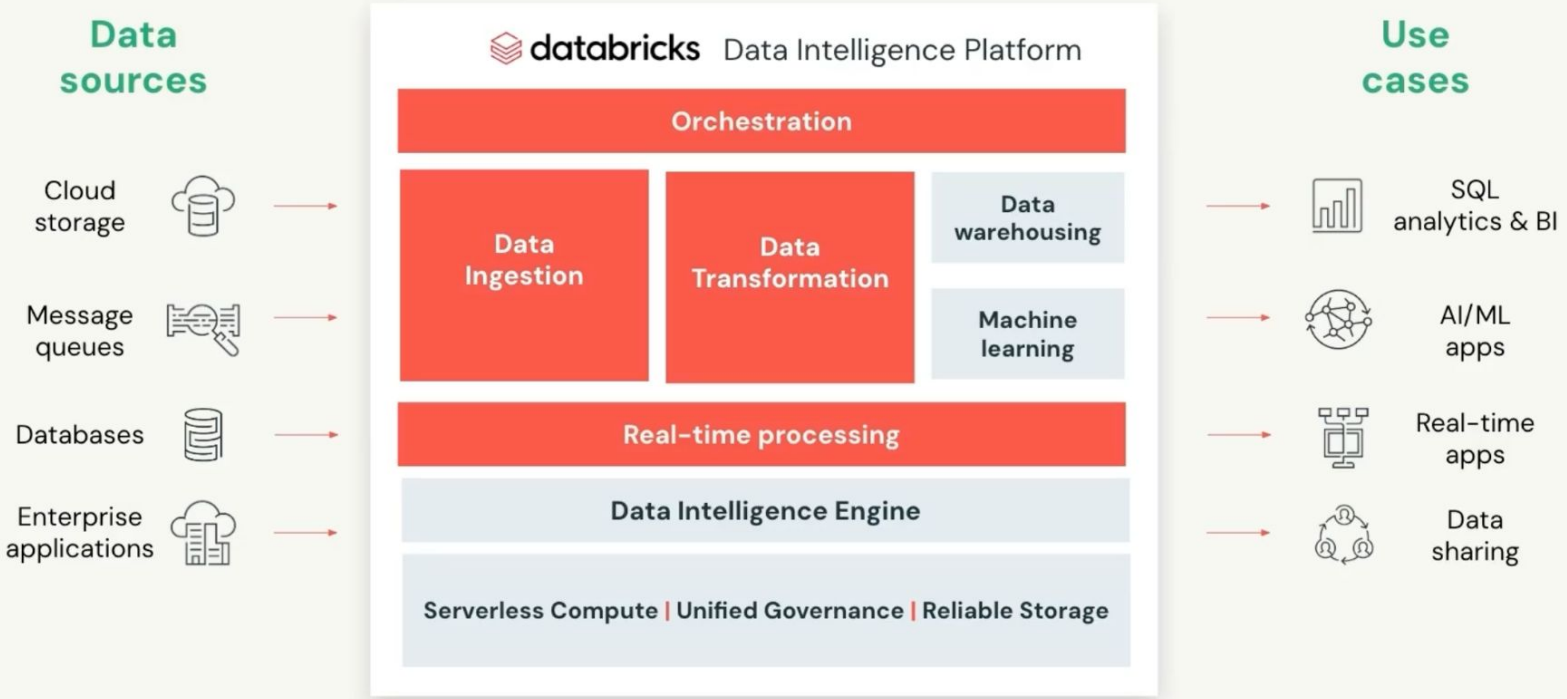

Time Travel


Schema Evolution and Enforcement


Many more!

- Unified Batch and Streaming
- Performance
- Scalable Metadata
- Optimization
- **Delta Lake is Open Source!**

Data Engineering on Databricks



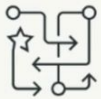
Delta Live Tables

The best way to do ETL on the lakehouse

```
CREATE STREAMING TABLE raw_data
AS SELECT *
FROM cloud_files ("/raw_data",
"json")

CREATE MATERIALIZED VIEW clean_data
AS SELECT ...
FROM LIVE.raw_data
```

Accelerate ETL development
Declare **SQL or Python** and DLT automatically orchestrates the DAG, handles retries, changing data



Automatically manage your infrastructure
Automates complex tedious activities like **recovery, auto-scaling, and performance optimization**

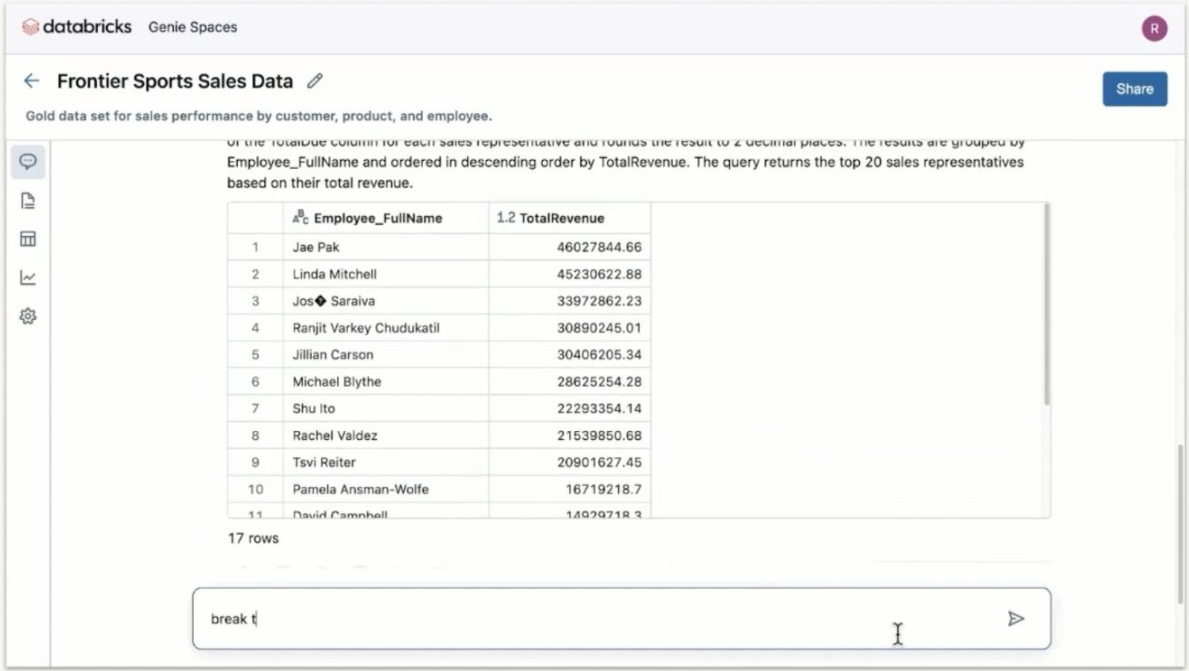
Ensure high data quality
Deliver reliable data with built-in **quality controls, testing, monitoring, and enforcement**

Unify batch and streaming
Get the simplicity of SQL with freshness of streaming with one **unified API**

AI/BI Genie Spaces

Public Preview

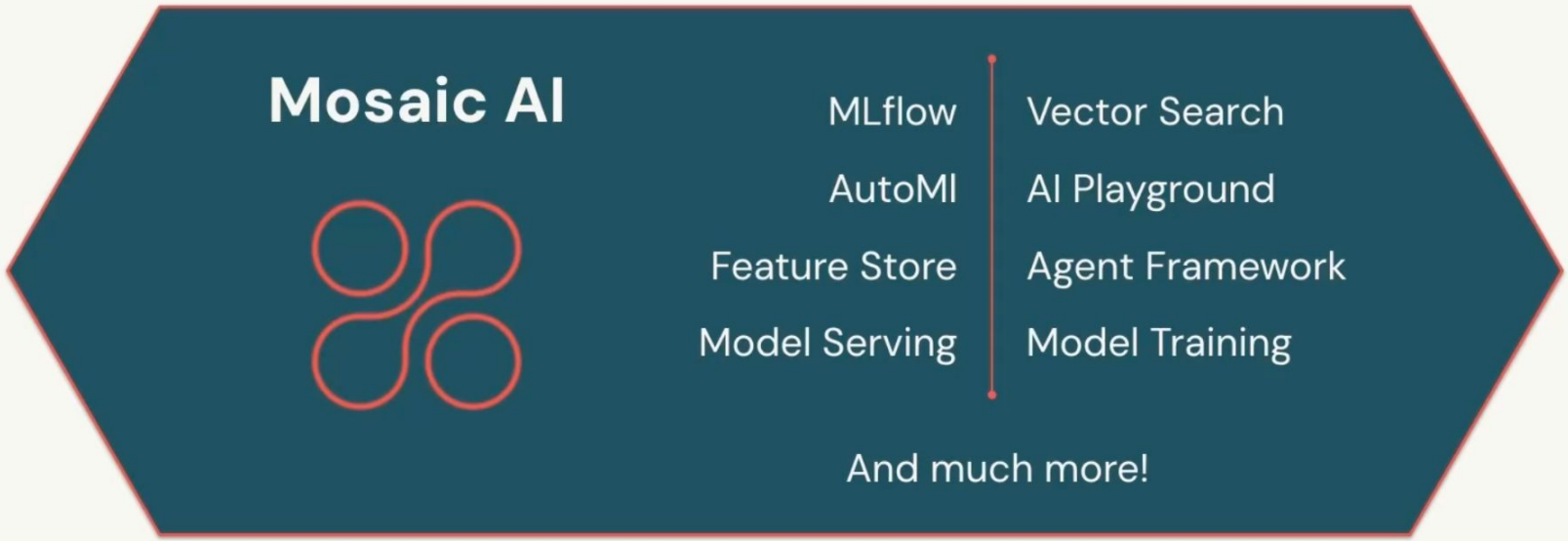
Ask questions of your data in natural language using your terms



- Supports non technical users with a no-code solution to gaining data insights using natural language.
- Provides a self-service access point to the data for follow-up questions without needing to involve a data analytics specialist.

Data Science, Machine Learning, and Gen AI

End-to-End Support for Generative and Traditional AI Application Development



Databricks on Google Cloud | Lakehouse Platform

Customers	Databricks on Google Cloud launched in 2021; 12,000+ global customers; 500+ 1M+; 1,000+ subscribers on GCP
Integrations	Built on Google Compute Engine; Integrated offering with Google BigQuery, Pub/Sub, Looker, Vertex AI, Cloud Storage

Deal Size: Small $200K, Medium $400K, Large $1M+ , Xtra-Large $3M+ TCV driving 1.5X GCP consumption in end customer tenant

Business Challenge

- Data silos drive high operational costs; Inconsistent policies reduce trust in the data
- Need for Insights and AI to support decisions about customer & products (Relevance & personalization, Data monetization, Real-time data for decision making)
- Disparate tools slow down cross-team productivity

Why Databricks?

- Data Forward: Insights and AI support decisions about customers and products
- Data, analytics, and AI working in concert (Customer lifecycle, Product development, Workforce productivity, Business Operations)
- Data powers every aspect of the business

databricks

Google Cloud
TECHNOLOGY
Partner of the Year
Data — AI & ML
2024

Sales Play

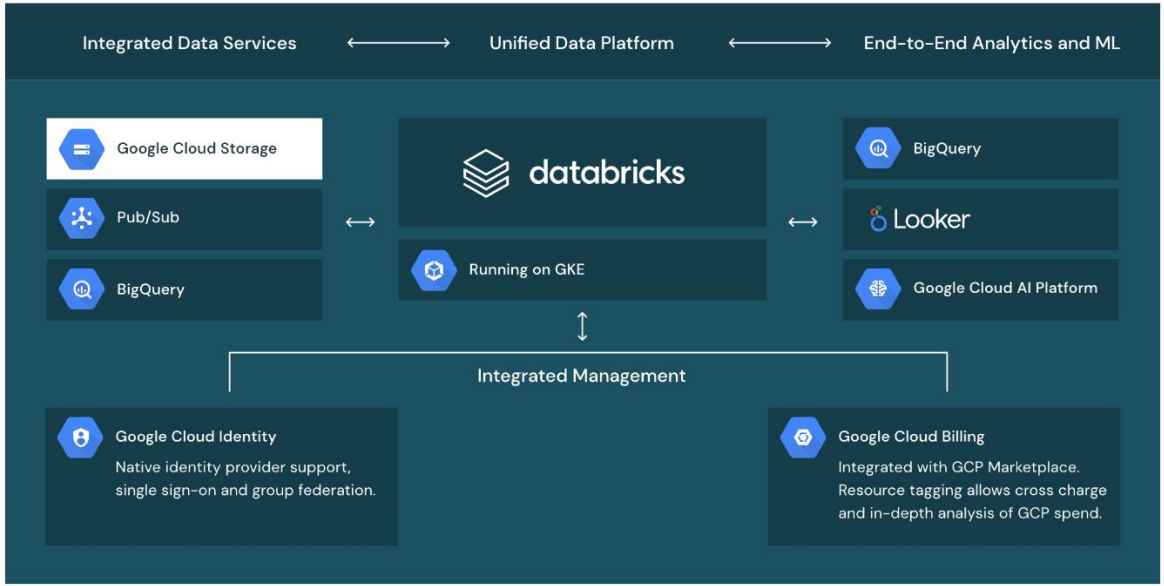
- **Migration:** Databricks Migration from Azure/AWS to Google Cloud; Migrating on-premise data (Hadoop) to build lakehouse
- **Multi-Cloud:** Drive Google Cloud consumption in existing Databricks Multi-Cloud customer
- **Data Engineering:** Data Lake, Delta Lake, Lakehouse (Bronze-> Silver -> Gold)
- **Data Science:** Data Prep, Exploration, Model Development & Training, Model Serving

Google + Databricks (Better Together)

Platform: Unified platform for all data use cases leveraging Google Cloud's superior performance, scale, and security from Google Cloud Identity, Google Cloud Storage and dedicated global network
Interoperability: Seamless interoperability across GCP Data & Analytics portfolio giving customers choice & flexibility

Streamlined integration with Google Cloud

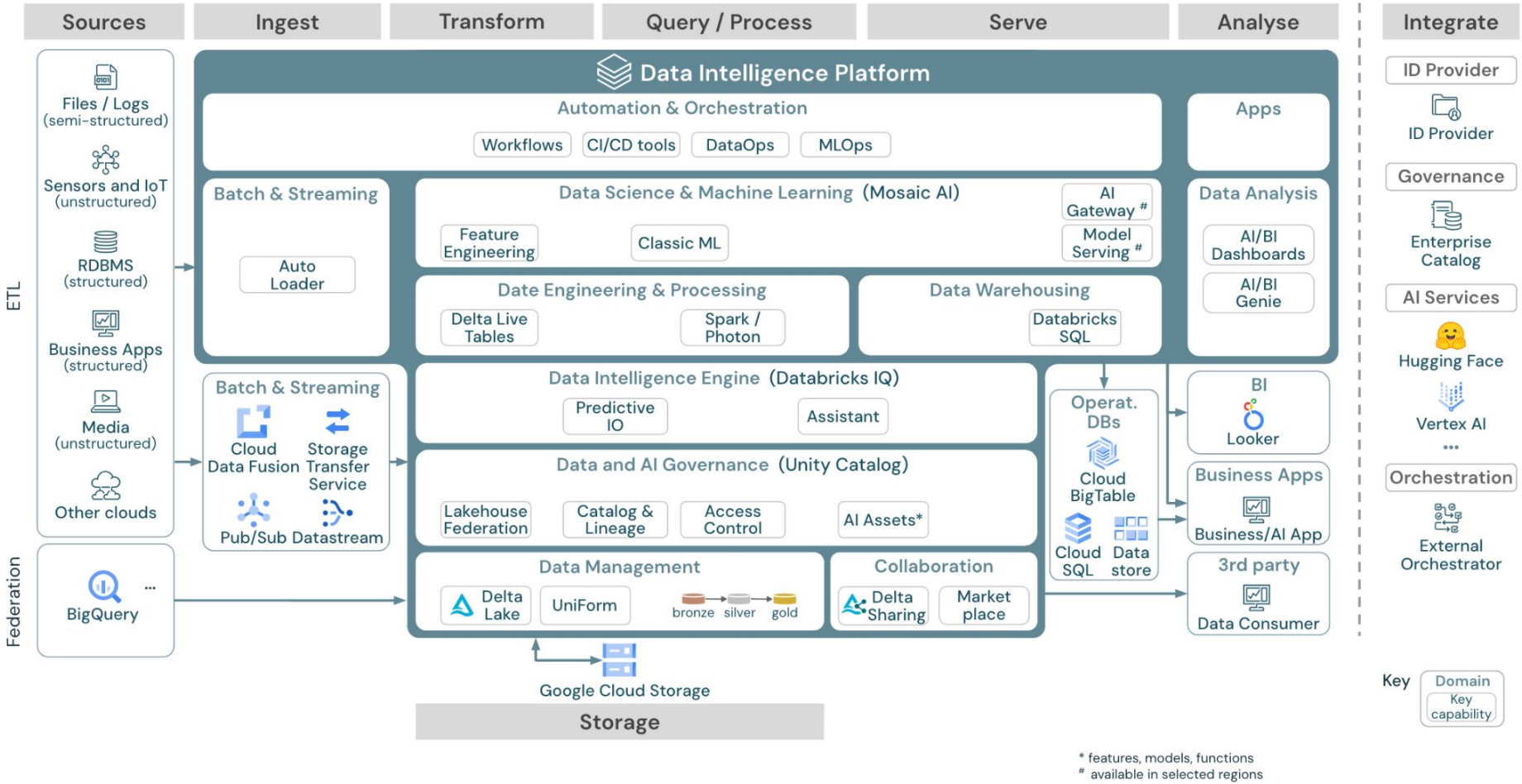
- Google Cloud Storage
- Google Kubernetes Engine
- Google Pub/Sub
- Google Big Query
- Google Cloud Identity
- Google Cloud AI Platform
- Google Cloud Billing
- Looker



Google Cloud Storage

Enable seamless read/write access for data in Google Cloud Storage (GCS) and leverage the Delta Lake open format to add powerful reliability and performance capabilities within Databricks.

[Databricks Site]



[Databricks GCP Ref Arch]

Business Drivers for Databricks on Google Cloud

Infrastructure Cost Savings

Databricks on GCP processes workloads with **improved performance** reducing infrastructure costs

Use of GCS's **multiple storage tiers** and **auto-management options** to optimize cost efficiency.

Performance & Scalability

Databricks benefits from GCS's **gRPC API**, which offers **fast and reliable data routing**

Effortlessly transfer petabytes of data with **event-driven transfers** for GCS, supporting **scalable and efficient** data movement.

Streamlined GCP Integrations & Security

Built-in connectors for Cloud Storage, BigQuery, Pub/Sub, Dataflow, Looker

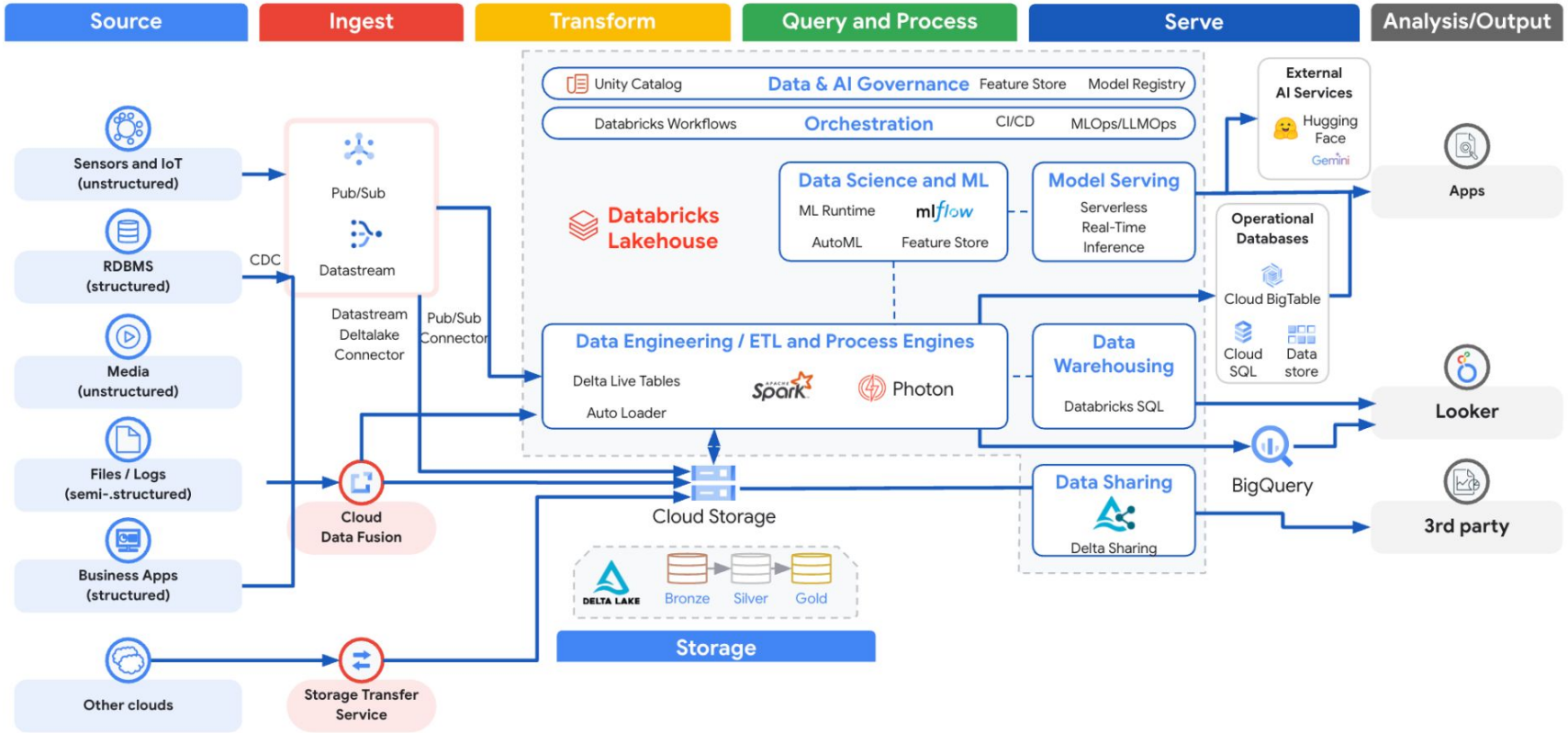
GCS provides **flexible location options** and **Managed Folders** for detailed access control, along with features like **Object Lock and Soft Delete** for robust data protection.

Access Google Cloud AI / Gen AI capabilities

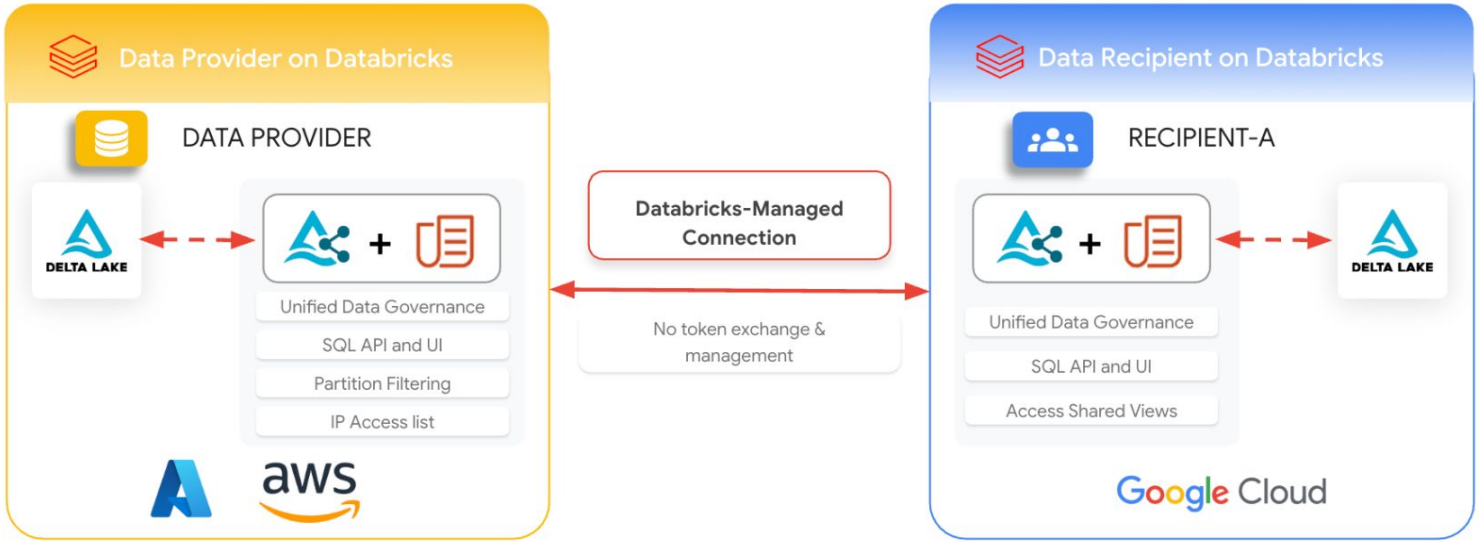
Access to Google's **state-of-the-art Vertex AI suite** with access to multi-modal GenAI capabilities plus **150+** models accelerating business outcomes

Seamless AI at Scale with **End-to-End MLOps** and **scalable** AI infrastructure to successfully deploy AI

Databricks on Google Cloud Reference Architecture



Databricks-to-Databricks Sharing



BigQuery adds native support for Delta Lake

- **Seamless Integration**: Query Delta Lake tables directly in BigQuery from Google Cloud Storage and AWS S3 without data movement.
- **Unified Data Governance**: Advanced security with row and column-level controls, plus dynamic data masking on Delta Lake tables.
- **End-to-End Analytics**: Easily leverage Delta Lake data for ML, BI, and reporting with BigQuery ML and Vertex AI.
- **Streamlined Setup**: Create and query Delta Lake tables using familiar GoogleSQL syntax, similar to other BigQuery tables.

```
CREATE EXTERNAL TABLE `PROJECT_ID.DATASET.DELTALAKE_TABLE_NAME`
WITH CONNECTION `PROJECT_ID.REGION.CONNECTION_ID`
OPTIONS (
  format ="DELTA_LAKE",
  uris=['DELTA_TABLE_GCS_BASE_PATH']);
```

https://www.tpc.org/information/benchmarks5.asp

Click to go forward, hold to see history

gt-weather edp-ai atpix tdid agentspace m gs gt grh da gsfdld edp-ai solar

TPC®

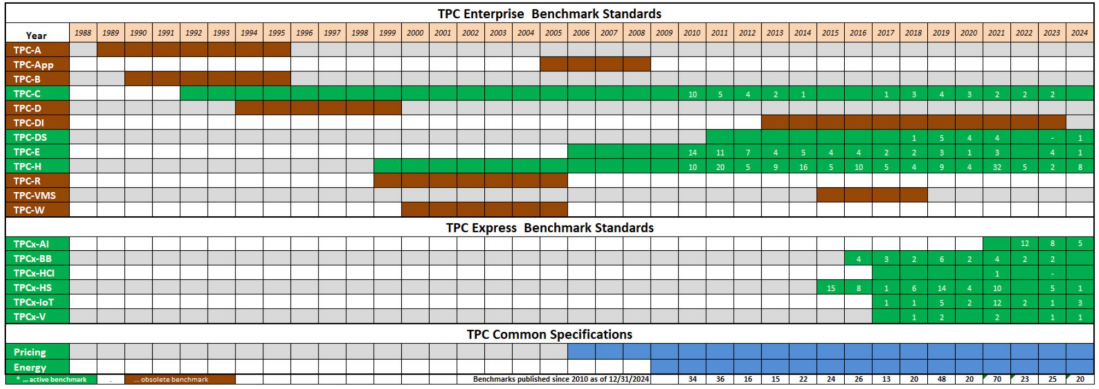
Alibaba.com AMD 怕睿数据 cisco DELL Technologies Hewlett Packard Enterprise HITACHI

HUAWEI IEIT SYSTEMS 浪潮云 intel Lenovo Microsoft NUTANIX

NVIDIA. ORACLE Red Hat timecho TRANSWARP TTA vmware

Home | About the TPC | Benchmarks/Results | Downloads | TPCTC | Miscellaneous | Search | Newsletter | Member Login

TPC Benchmarks Overview



TPC-DS

TPC-DS is the de-facto industry standard benchmark for measuring the performance of decision support solutions including, but not limited to, Big Data systems. The current version is v2. It models several generally applicable aspects of a decision support system, including queries and data maintenance. Although the underlying business model of TPC-DS is a retail product supplier, the database schema, data population, queries, data maintenance model and implementation rules have been designed to be broadly representative of modern decision support systems.

This benchmark illustrates decision support systems that:

- Examine large volumes of data
- Give answers to real-world business questions
- Execute queries of various operational requirements and complexities (e.g., ad-hoc, reporting, iterative OLAP, data mining)
- Are characterized by high CPU and IO load
- Are periodically synchronized with source OLTP databases through database maintenance functions
- Run on “Big Data” solutions, such as RDBMS as well as Hadoop/Spark based systems

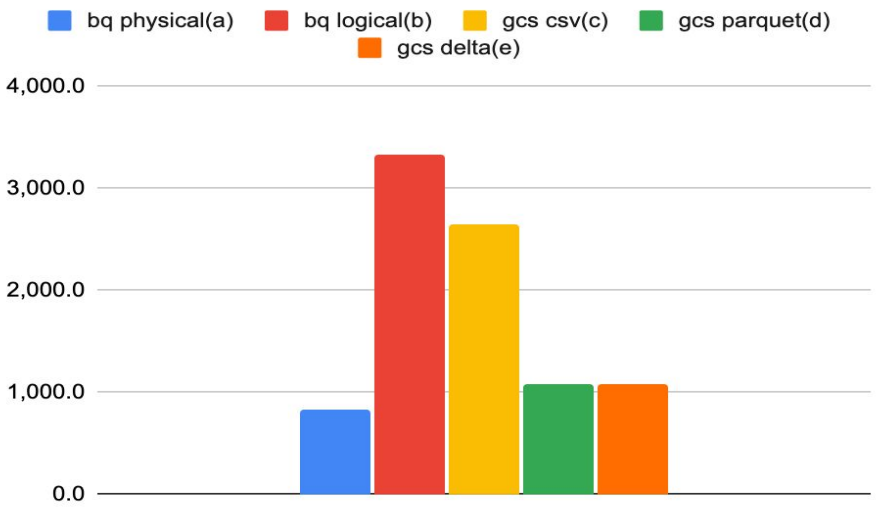
DEMO

why GCP - Storage

Using Industry tests for DW: TPC-DS

Storage - Comparing compression ratio

- for TPC-DS, default write in spark to parquet bq physical size is 76% of parquet size



TPC-DS 2 TB tables (in MB)									
table	**bq physical(a**	**vs bq log**	**bq logical(b)**	vs csv	**gcs csv(c)**	vs parquet	**gcs parquet(**	vs delta	**gcs delta(e)**
Total	**831,009.8**	**25.00%**	**3,323,802.4**	**31.44%**	**2,643,298.0**	**76.79%**	**1,082,204.5**	**76.79%**	**1,082,192.7**
store_sales	342,067.5	23.21%	1,473,547.6	29.50%	1,159,662.81	72.9%	469,288.0	72.89%	469,264.5
catalog_sales	282,048.9	26.29%	1,072,712.5	32.72%	862,064.06	81.2%	347,405.9	81.19%	347,411.2
web_sales	141,838.3	26.34%	538,472.1	32.89%	431,192.21	84.1%	168,686.3	84.08%	168,690.7
store_returns	33,023.8	26.77%	123,348.4	33.25%	99,316.53	65.3%	50,542.9	65.34%	50,543.9
catalog_returns	21,302.9	27.40%	77,736.3	35.38%	60,215.12	69.3%	30,747.2	69.28%	30,747.7
web_returns	10,043.4	28.91%	34,740.4	35.43%	28,348.78	67.8%	14,816.0	67.79%	14,816.3
customer	448.7	32.75%	1,370.3	38.28%	1,172.24	90.8%	493.9	90.84%	494.0
inventory	144.4	12.75%	1,132.6	20.64%	699.48	121.8%	118.5	121.75%	118.6
customer_addre	76.0	13.05%	582.5	14.57%	521.74	90.0%	84.4	90.03%	84.4
customer_demo	9.0	7.38%	122.0	11.99%	75.09	76.8%	11.7	76.77%	11.7
item	3.6	23.66%	15.2	26.61%	13.53	66.2%	5.4	66.08%	5.5
date_dim	1.7	12.69%	13.6	17.65%	9.77	82.6%	2.1	82.11%	2.1
time_dim	0.9	13.49%	6.8	19.01%	4.79	79.3%	1.1	78.82%	1.2
call_center	0.0	0.00%	0.0	326.42%	0.00	79.4%	0.0	36.04%	0.0

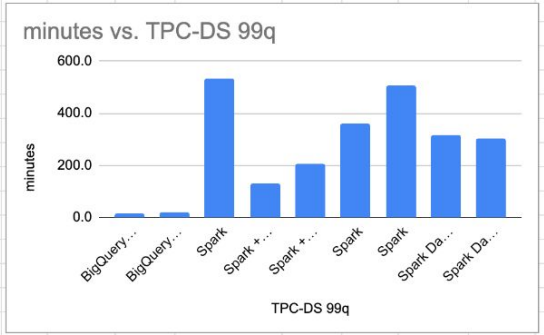
why GCP - Compute

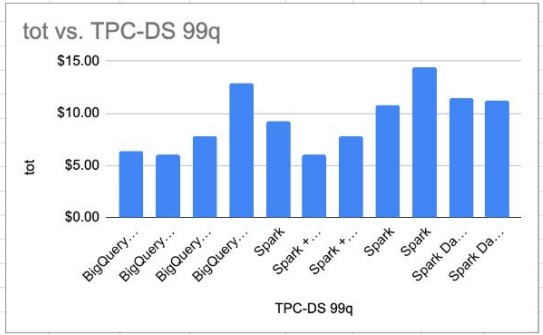
TPC-DS 99q 2TB	minutes	tot	m/$	Data	BQ ba	BQ max a	BQ$	compute	vcpu/w	workers	cpu	ram	DBU	DataProc	Pricing avg/slot	GB Processe	OperA	OperB	BQ Com	DBK	Compute	Operations
BigQuery On-Demand	19.4	$110.97	0.2	raw	onD	-	-	-	-	-	-	-	-	-	1,043.8	22,726.0	-	-	110.967	-	-	-
BigQuery On-Demand	14.3	$104.22	0.1	(a)	onD	-	-	-	-	-	-	-	-	-	766.8	21,345.2	-	-	104.224	-	-	-
BigQuery Standard	95.9	$6.73	14.2	(a)	std	100	$0.040	-	-	-	-	-	-	-	105.4	21,345.1	-	-	6.734	-	-	-
BigQuery Standard	18.7	$6.31	3.0	(a)	std	800	$0.040	-	-	-	-	-	-	-	506.8	21,345.1	-	-	6.306	-	-	-
BigQuery Standard	13.1	$6.03	2.2	(a)	std	1600	$0.040	-	-	-	-	-	-	-	690.2	21,345.1	-	-	6.033	-	-	-
BigQuery Entp	16.8	$7.81	2.2	(a)	ent	1600	$0.036	-	-	-	-	-	-	-	774.3	21,345.1	-	-	7.808	-	-	-
BigQuery Entp	21.5	$12.92	1.7	external	std	1600	$0.036	-	-	-	-	-	-	-	999.9	33,617.9	302	412,629	12.919	-	-	$0.1666
Spark	533.6	$9.28	57.5	external	-	-	-	n2-standard-4	4	4	20	75	3.60	-	-	-	418,523	944,381	-	$4.937	$1.874	$2.4704
Spark + photon	129.7	$6.00	21.6	external	-	-	-	n2-standard-4	4	4	20	75	9.00	-	-	-	384,345	819,750	-	$3.255	$0.494	$2.2496
Spark + photon + cache:false	205.3	$7.77	26.4	external	-	-	-	n2-standard-4	4	4	20	75	9.00	-	-	-	213,505	2,724,890	-	$4.957	$0.654	$2.1575
Spark	361.9	$10.76	33.6	external	-	-	-	n1-standard-4	4	8	36	270	6.39	-	-	-	471,532	920,119	-	$6.021	$2.015	$2.7257
Spark	505.6	$14.45	35.0	external	-	-	-	n1-standard-8	8	4	20	75	7.05	-	-	-	373,415	789,032	-	$9.175	$3.093	$2.1827
Spark Dataproc	315.9	$11.48	27.5	external	-	-	-	n1-standard-8	8	8	72	270	-	$3.7908	-	-	1,234,524	4,819,254	-	-	$3.378	$8.1003
Spark Dataproc	301.5	$11.19	27.0	external	-	-	-	n1-standard-4	4	16	68	255	-	$3.4170	-	-	1,217,734	4,752,216	-	-	$3.197	$7.9896

databricks

- (a) a few tables clustered + small tables w/partition (<1MB)
- external = parquet dump from bq
- no tunning
- only a few corrections on queries to run with SparkSQL

	pvp			
Databricks DBU Premiur	$0.1500	h		
BQ on Demand	$5.0000	TB		
BQ Editions Std	$0.0400	h		
BQ Editions Entp 3y	$0.0360	h	$0.06	(-40% 3y reservation)
n2-standard-4	$0.2137	h	$0.041	(spot)
n1-standard-4	$0.1562	h	$0.036	(spot)
n1-standard-8	$0.4184	h	$0.071	(spot)
OperA	$0.0050	p/1000	list	





Thank You!

and reach your
local PE or CE

free training - Databricks gcp training

How is Google Cloud Platform Different

Bruno Basto
Customer Engineer, Data Analytics
basto@google.com

Google Cloud

Thank You!

and reach your

local PE or CE