Data Engineering with Google Cloud

Simple. Limitless. Intelligent.

Google Cloud

68% of companies are unable to realize measurable value from data.

Accenture, Closing the Data Value Gap, 2019

Breaking down silos

- Across Products
- Across Clouds
- Across Teams

Predictive & real-time

- From the past to the future
- From analysis to action
- From passive to autonomous

Democratization of insights

- Access everywhere
- Access for everyone
- Access for any scenario

Intelligent and Unified Data Governance

"90% of employees say that their work is slowed by unreliable data sources."

Dimensional Research, 2020

"80% of analytics work is still descriptive."

MIT, 2020

"86% of analysts struggle with data that's out of date."

Dimensional Research, 2020

Google was born as a data company...

Organize the **world's information** — Data

and make it universally

accessible and useful — Intelligence

...and is a world leader in applying Machine Learning to real-world situations

Search
Search ranking
Speech recognition

Translate
Text, graphic and speech translations

Photos
Photos search

Gmail
Smart reply
Spam classification

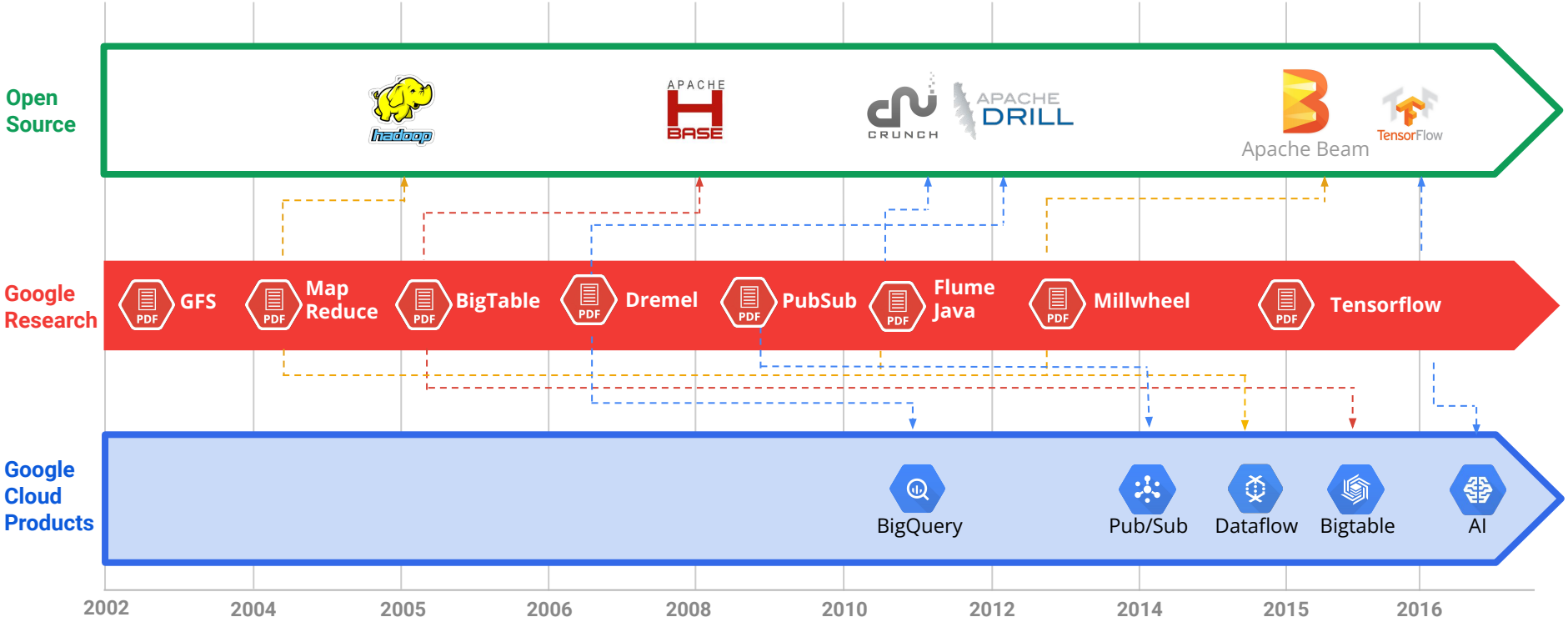
Self Driving Car
1.5MM miles driven

Data Center Power Usage
Reduced cooling energy 40%

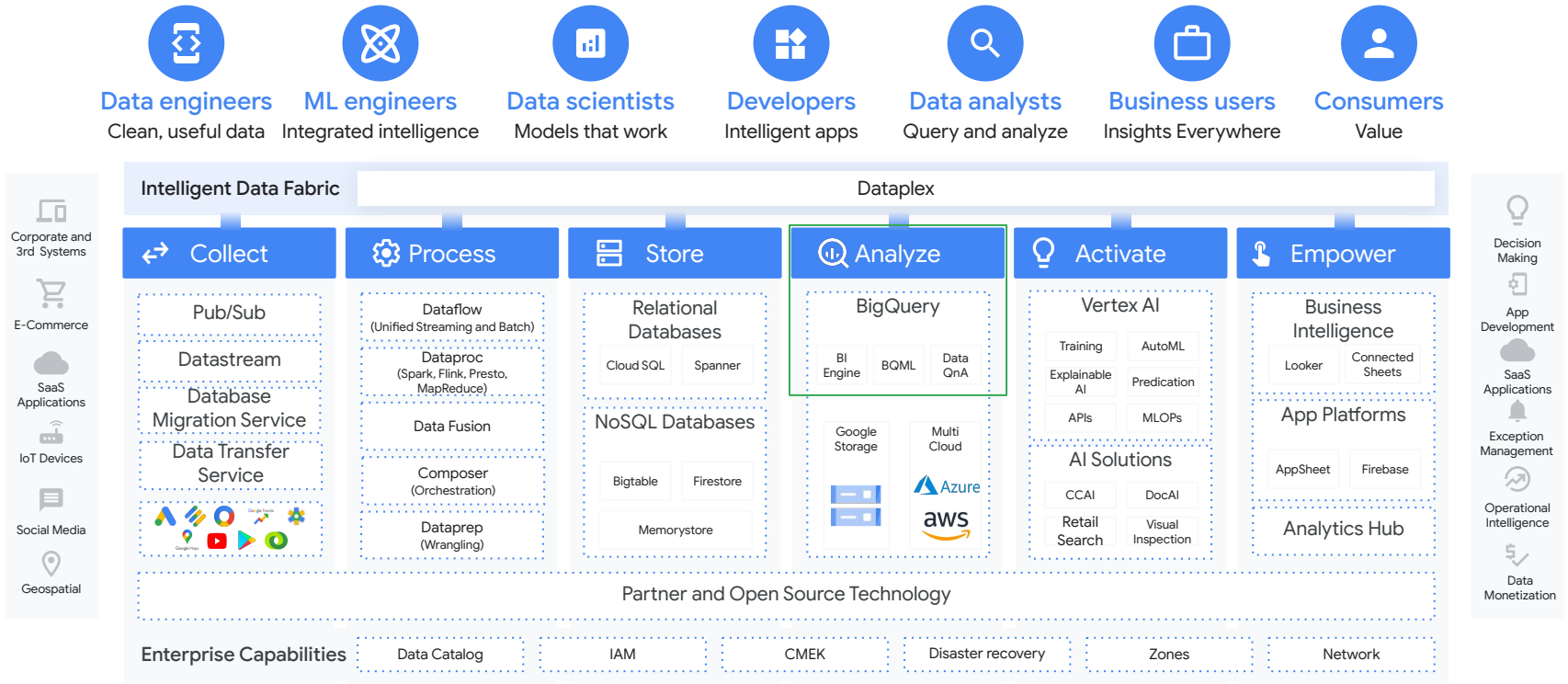
Maps
Street View image
Popular times

YouTube
Video recommendations
Better thumbnails

15+ years of solving data problems



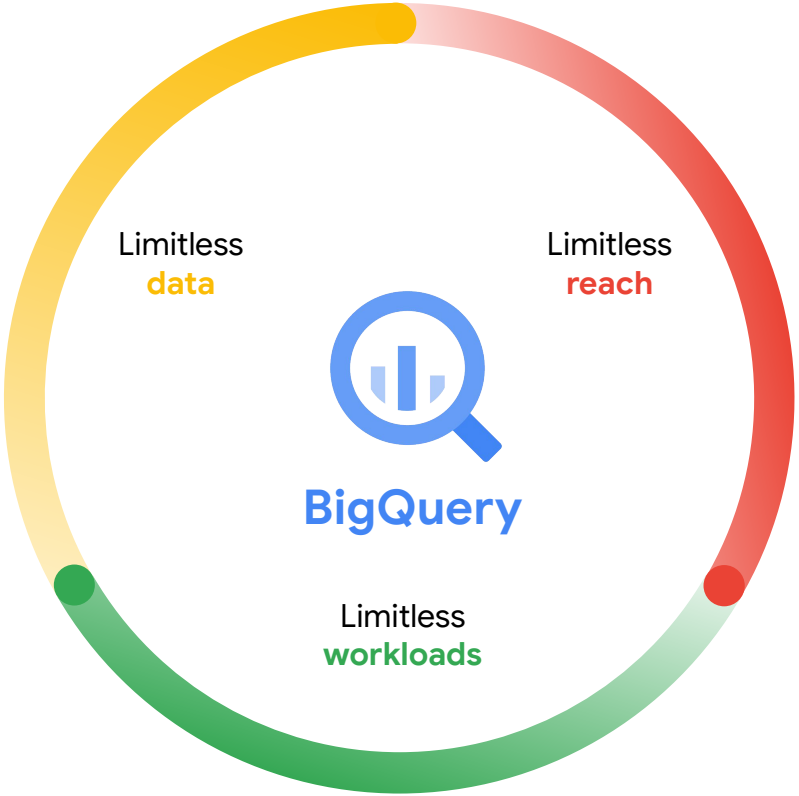
A Data Cloud that supports your needs



BigQuery

The core of Google's Data Cloud to power your data-driven innovation.

100k+ data professionals have started their Data Cloud journey using BigQuery with trials growing nearly **150% YoY in 2021**.



Google Cloud helps you do more

Limitless data

- Completely serverless
- All types of data
- Data exchanges

All workloads

- SQL, Spark, Search, Stream
- Built-in AI/ML
- Analytical + Transactional

For everyone

- Built-in BI
- Analytical applications
- Partner ecosystem

Cost Effective | Highly Productive | Always On | Easy to Secure | Clear Compliance | Open Extensions

Completely serverless

Why BigQuery?

- Simplifies capacity management
- Dynamically adjusts to demand

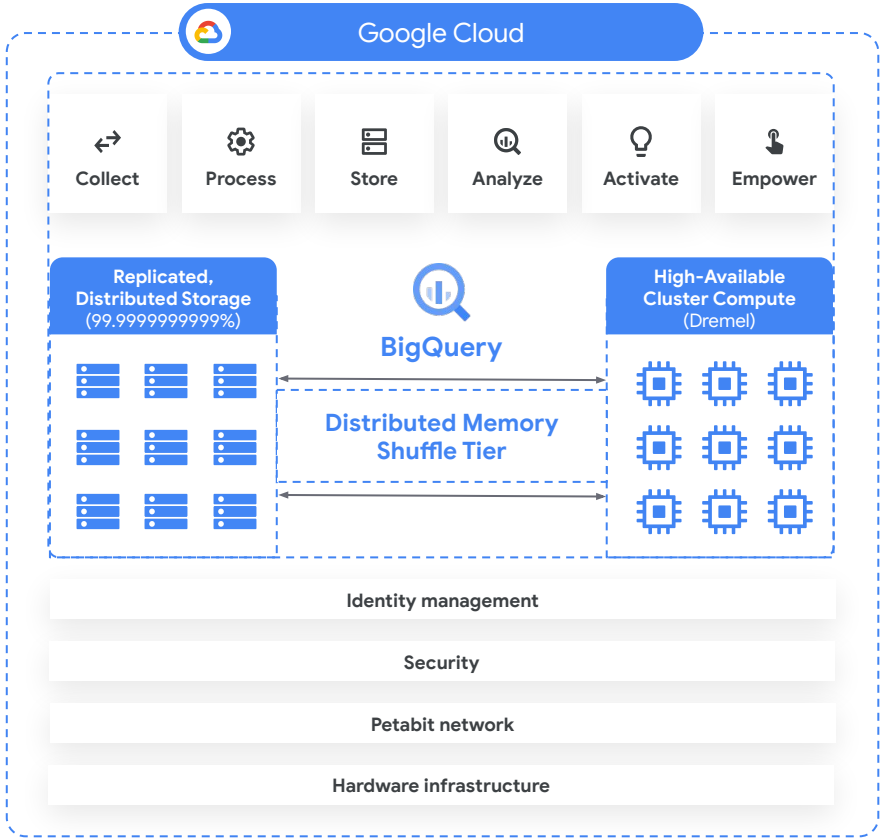
VS

- Plan, manage, pay VMs
- Limit use data due to capacity restrictions

Completely elastic

Distributed storage and compute with ultra-high bandwidth including distribute petabyte scale in-memory storage for temp data and state:

- Auto-start and auto-pause
- 0-Second warm up to get maximum performance
- Accelerate queries in flight
- No performance cliff due to local capacity saturation
- Immune to large-scale hardware failures



Limitless data

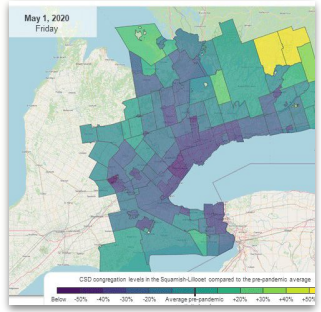
All types of data

Why BigQuery?

- Simplifies data type management with a unified ecosystem
- Provides unique data capabilities (geospatial)

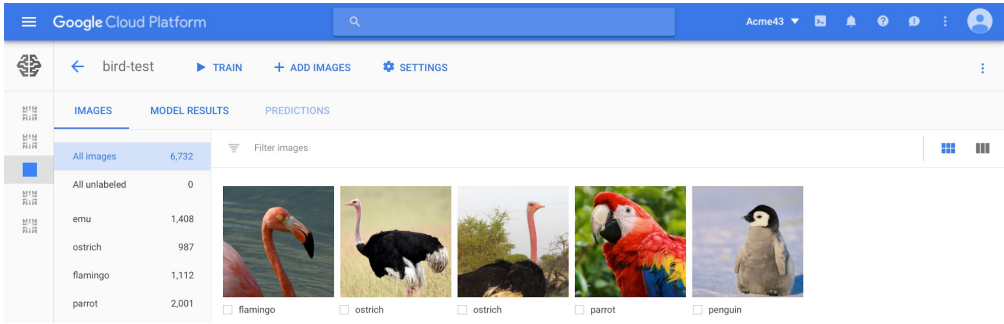
VS

- Manage pipelines and integrations
- Miss value from unsupported data types



All your data types in one platform

- Structured
- Semi-structured (JSON)
- Unstructured (text, images)
- Parquet
- JSON
- Nested Tables
- Geospatial



Analytical + Transactional

All workloads

Why BigQuery?

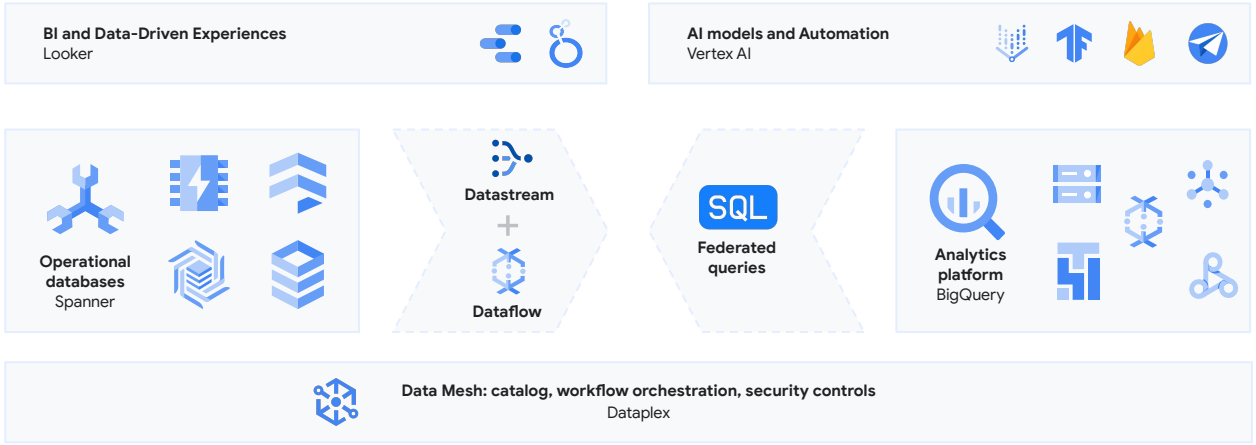
- Treat data as an asset regardless of where it resides (OLTP + OLAP)
- Simplify integrating transactional data stores

VS

- Work with technology that is solely focused on OLAP
- Manage data integration pipelines

Analytical & Transactional Workloads in One Place

Analyze Cloud Spanner and Cloud SQL data in real-time **without movement or copy**.



Built-in business intelligence

For everyone

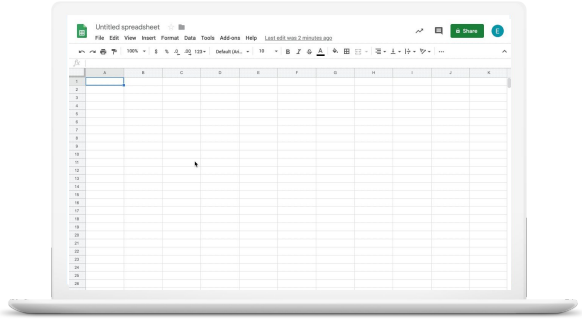
Why BigQuery?

- Simplify getting data into the hands of your decision makers

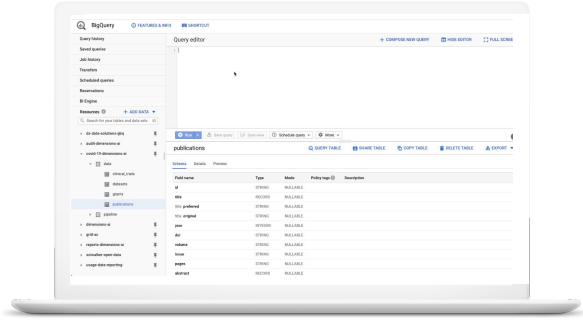
VS

- Run into data democratization bottlenecks

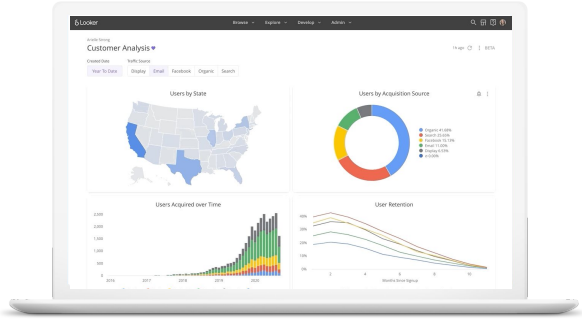
Intuitive analysis using Connected Sheets



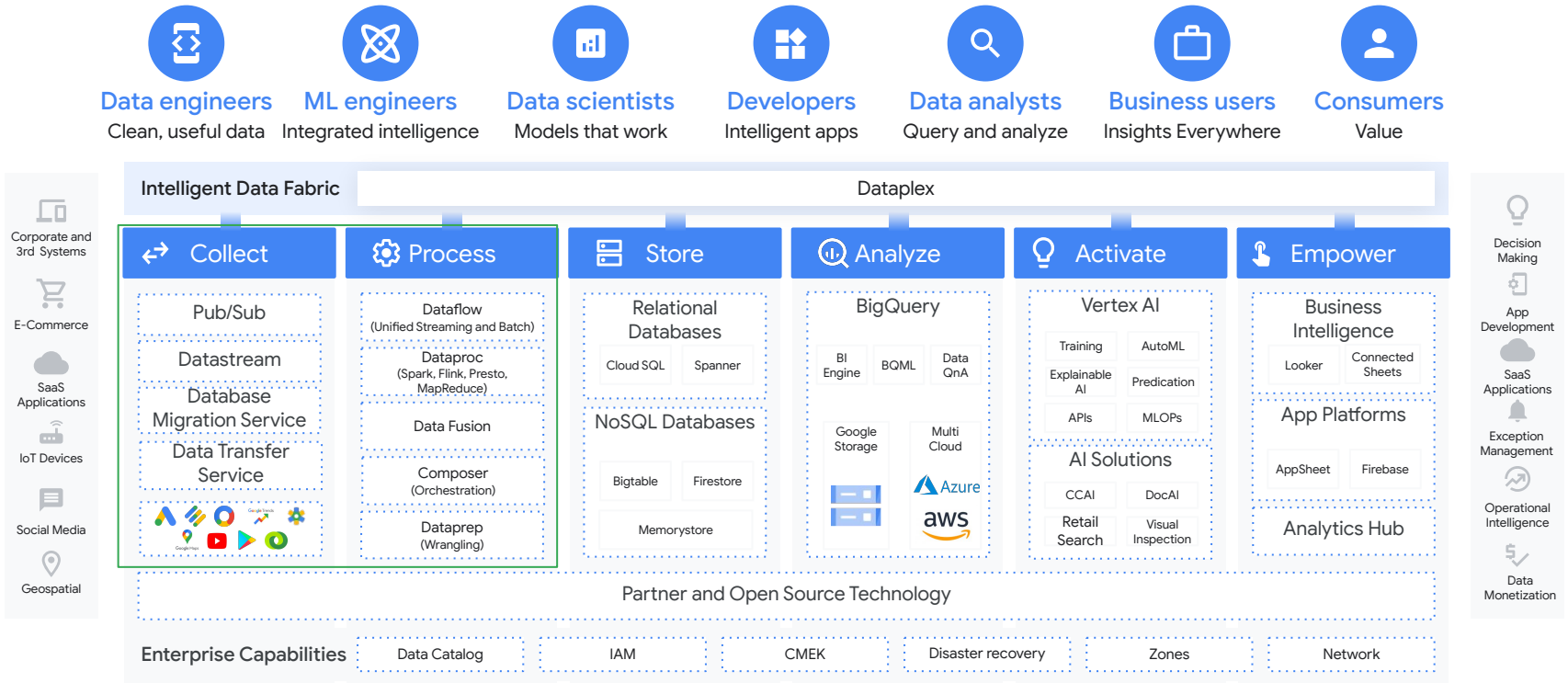
Built-in Analytics with Data Studio



Data Rich Experiences with Looker & LookML



A Data Cloud that supports your needs



Google Cloud and Data Movement patterns

Data Ingestion

Pure ingestion without transforms

Cloud-native unified batch and streaming ingestion platform, proven at scale with Pub/Sub and Dataflow

No-Code batch ingestion from 150+ sources with BigQuery Data Transfer Service

Data Replication

Replication and Change Data Capture

Cloud-native, serverless, performant, simple and flexible replication and change data capture with Datastream, with unmatched flexibility with Google Cloud services

Data Integration

Covering ETL and ELT patterns

Cloud-native, fully-managed and unified Data Integration platform - Dataflow

No-Code Data Integration platform for UI-driven development - Data Fusion

Open Source Data Integration for Hadoop/Spark workloads on GCP - Dataproc

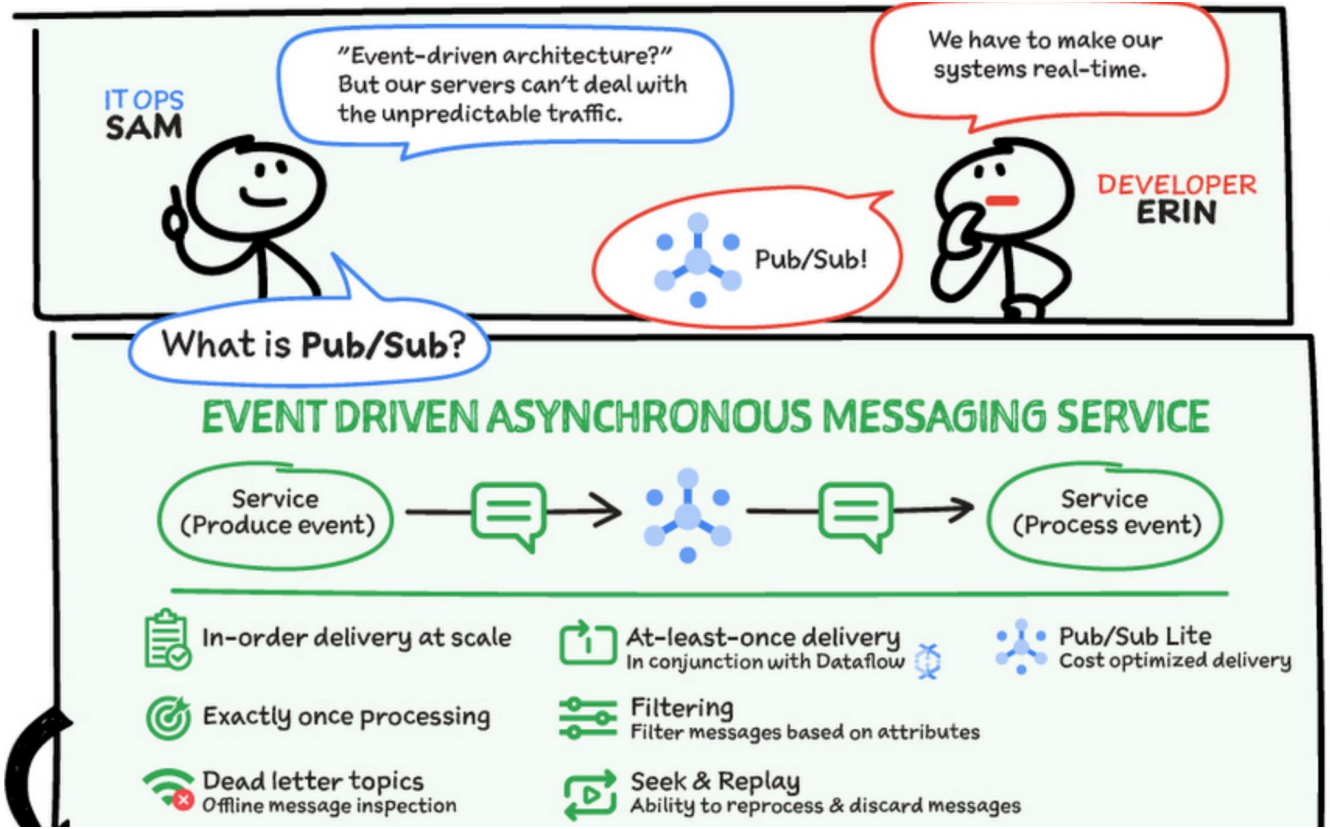
Data Ingestion

01

Pub/Sub

Events Data | SDKs | Streaming

Hyperscale and global async messaging service



Deeply Integrated
Scalable analytics with Dataflow, serverless actions with Functions, built in monitoring, audit logging, and compliance.

Dataflow Templates

No Code | Batch & Streaming

Serverless, data processing service for both streaming and batch data

The power of Google Cloud Native, the ease of a no-code solution for all users

Turn-key, click to deploy experience for the most common data ingestion tasks, powered by Google Cloud Dataflow

Preferred platform for data ingestion and movement tasks for both Google products and for partners

Extensible templates for user-defined functions for each customer's unique needs

40+

Google & Partner provided Templates

Cloud Database Example Templates

- Cloud Spanner <-> GCS
- Cloud BigTable <-> GCS / Cassandra
- Cloud Datastore <-> GCS
- Cloud Datastream to GCS
- Cloud Datastream to BQ

Analytics Example Templates

- Pub/Sub to BigQuery
- Pub/Sub to Pub/Sub
- Pub/Sub to GCS
- GCS to BigQuery
- BigQuery to Parquet

SaaS Services and Third Party Example Templates

- BQ/GCS <-> Splunk
- BQ/GCS/PubSub <-> Elastic Cloud
- Salesforce to BQ (via JDBC)
- SAP to BQ (via JDBC)
- Apache Cassandra to Bigtable
- Apache Hive to BigQuery

Public Clouds Data | No Code | Batch Ingestion

BigQuery Data transfer service

No-code, fully automated loads into BigQuery

Cloud storage

Ingest from Google Cloud Storage

Automatically ingest files in Google Cloud Storage. Maintains awareness of files that have already been ingested to simplify file management.

Amazon Redshift & S3

Transfer and ingest from Amazon Redshift & S3

Handles transfers of data from the Amazon ecosystem into BigQuery.

Scheduled query

Generate new tables from queries

Invoke queries on existing BigQuery data to materialize new data. Simplify ETL/ELT style workloads.

100+ SaaS apps...

Connectors available via Partners

Native Connectors

Google Cloud

Data Replication

02

SaaS & DBs | No Code | Streaming

Datastream

Performant, Simple, Flexible Change Data Capture

Datastream

Serverless CDC and data replication

Replication

Cloud SQL | Spanner | BigQuery | Memorystore | DMS | Elasticsearch | Extensibility

Change Streams

Pub/Sub (Q1 2022) | Cloud Storage | Kafka

Integrations

Cloud Data Fusion | Dataflow | Cloud Functions

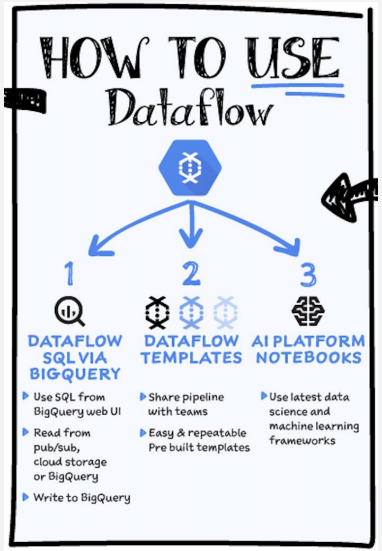
Data Integration

03

Dataflow

Apache Beam | SDKs | Batch and Streaming

Serverless, data processing service for both streaming and batch data



Ultra simplicity

Serverless, auto-provisioning, and self healing; automated performance, work balancing, unified batch and streaming under one lane

Accessible ML

Out of the box support for MLOps (TFX, Kubeflow), ML Inferencing, GPU and large scale data processing

Best of OSS and optimized platform

Innovate with Apache Beam (open source unified programming model) SDK; create pipelines in your language of choice (Java, Python, Go, SQL); optimize workloads with built in monitoring and observability

Built to scale

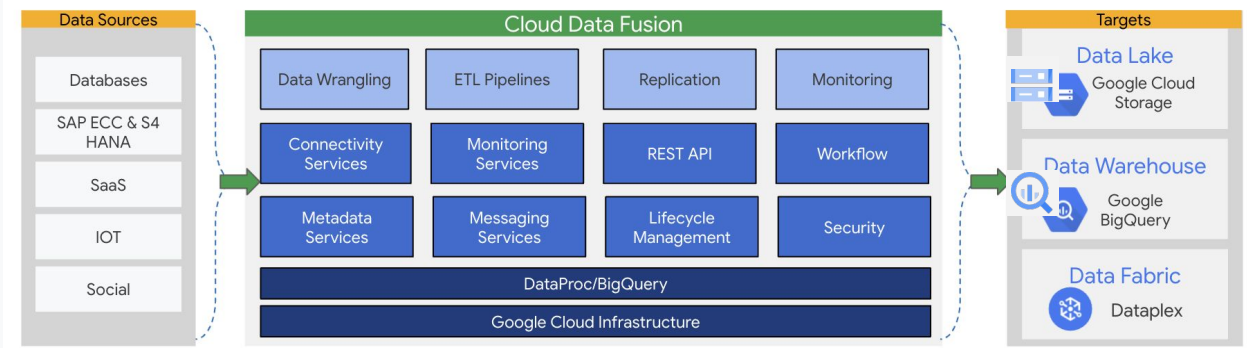
Designed and built for horizontal and vertical scaling (first in industry). Optimize pipeline performance at any scale; maximize utilization to save costs

OSS CDAP | No Code | Batch and Streaming

Data Fusion

No-Code Data Integration on Google Cloud

Google Enterprise Data Integration Service that can provide enterprise grade, simple to use, cloud native, open and comprehensive capabilities for can provide enterprise data ingestion and data integration



OSS Ecosystem | SDKs | Batch and Streaming

Cloud Dataproc

Open Source with Google Scale

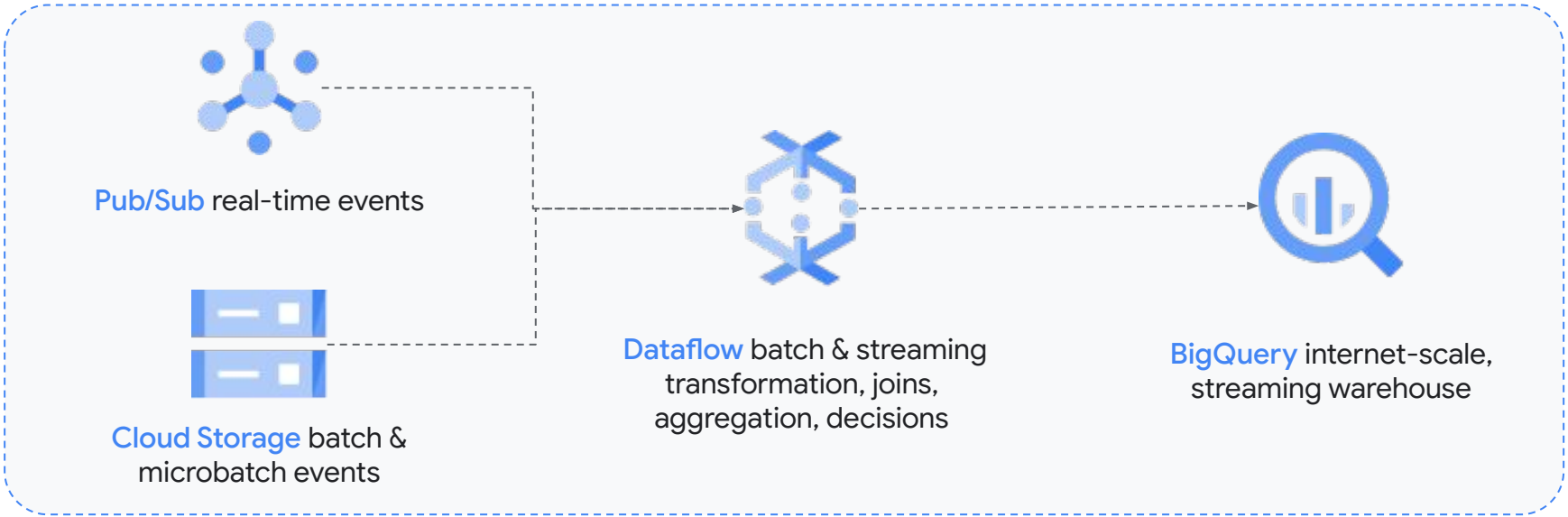
- Fast
- Secure
- Managed
- Cost-effective
- Serverless Option

...and more!

Google Cloud

Build Data Pipelines Like Industry Leaders

After 15 years powering billion-user Google products, we help run digital leaders & transform enterprises worldwide.



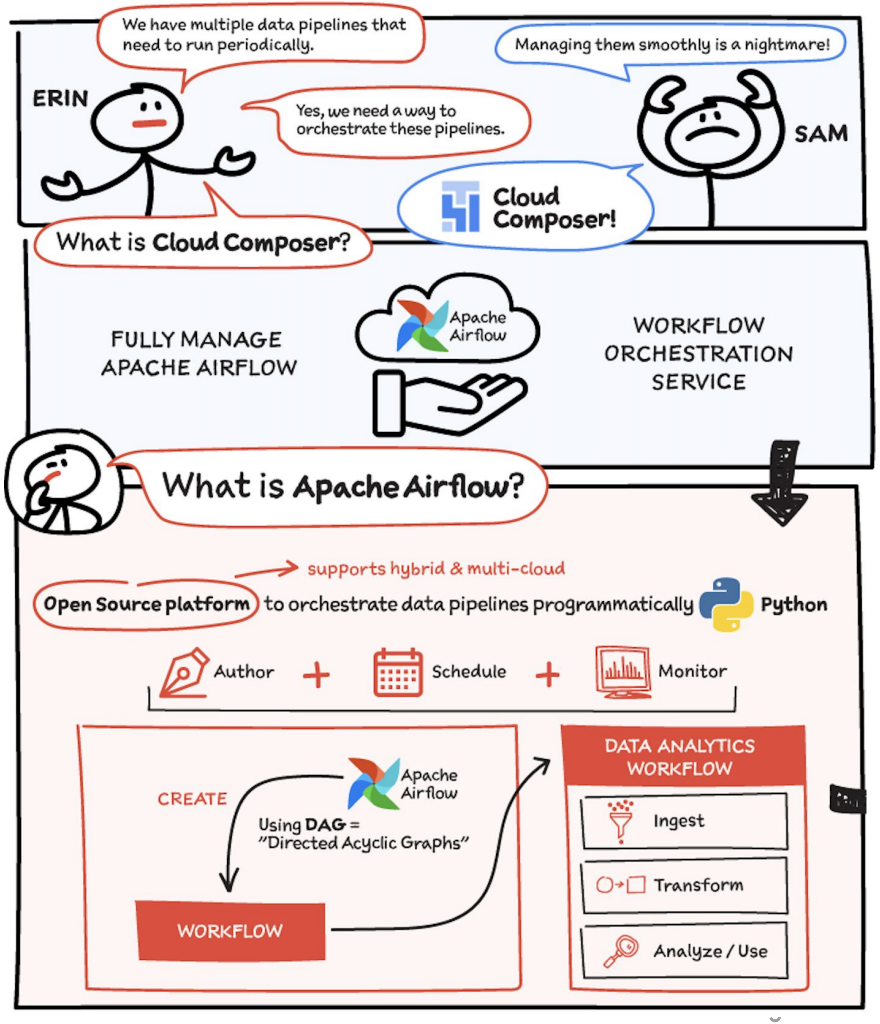
Proven at Google scale, scales with usage for ease of use, and pervasive ML

Data Orchestration

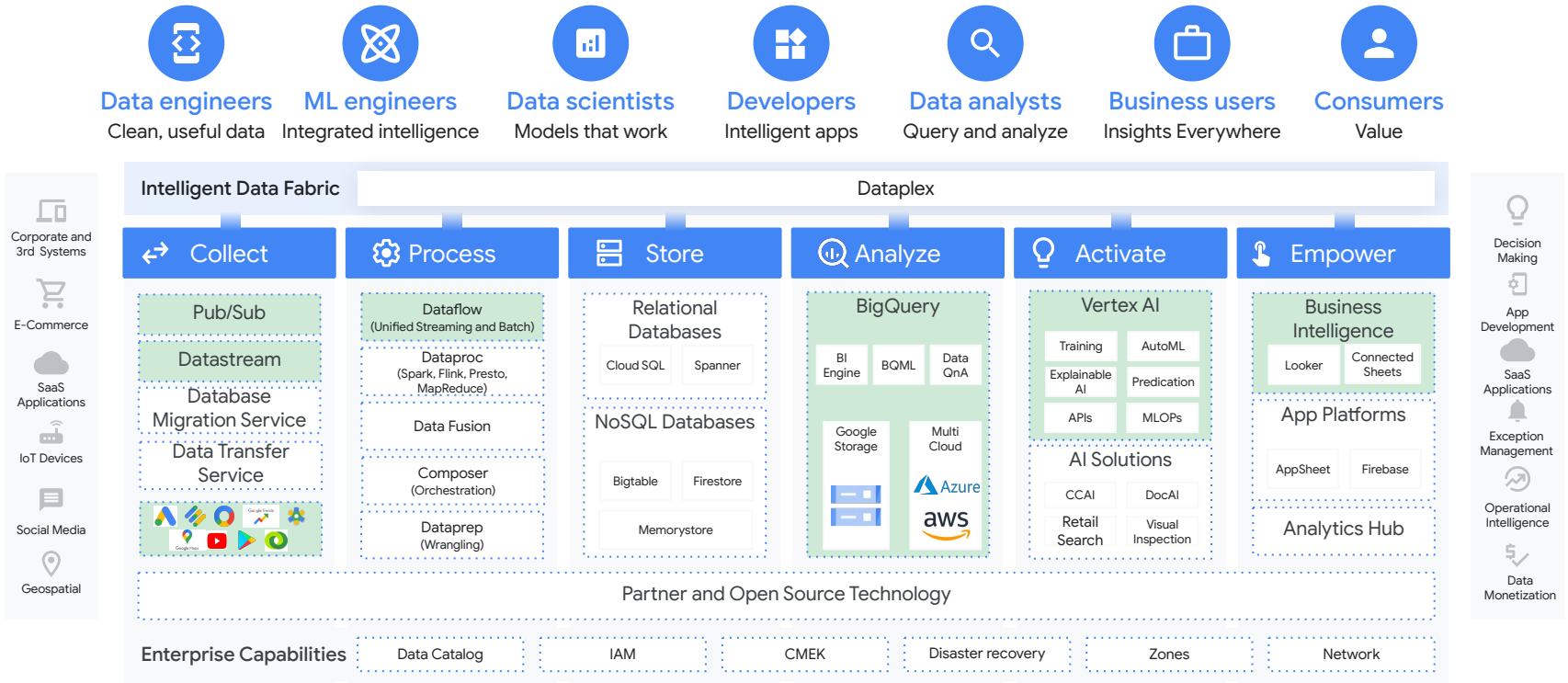
04

Cloud Composer

- **Workflow Orchestration service**
- **Fully Managed Apache Airflow**
- Based on Airflow, a very popular open-source orchestration pipeline
- Designed to enable users to programmatically author, schedule and monitor workflows
- Workflows are authored as directed acyclic graphs (DAGs), and can be configured as code - using Python



At Hackfast, you will have a chance to explore ...



QUESTIONS?

Thank you

Google Cloud