HUGE

Hello

Getting started with BigQuery

September 9, 2017

Who am I?

William Gómez

Engineer at Huge

(Full stack wannabe web engineer at Huge)

Python Medellin meetups co-organizer

Google Cloud Certified Data Engineer

Freelancer

Machine learning lover

Data science FEM - Mentor

Agenda.

1. BigQuery
2. Features
3. Architecture
4. Use cases
5. Costs
6. Examples

1

BigQuery

What is it?

BigQuery is Google's fully managed, petabyte scale, low cost analytics data warehouse.

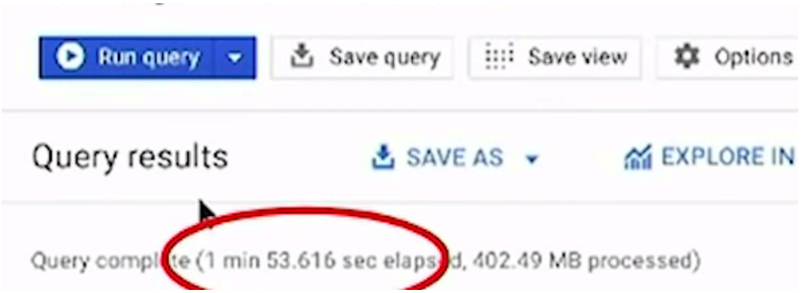
Remarkable features:

- NoOps.
- Pay-as-you-go model.
- Securely share insights.
- Streaming ingestion captures and real time analyses.
- Analyze up to 1 TB of data and store 10 GB of data for free each month.

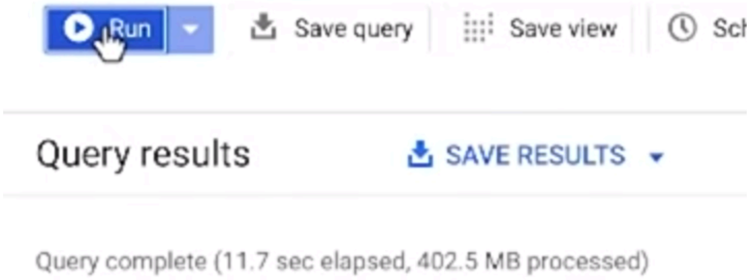
Performance evolution

Three years ago:

Last year: Enabling clustering

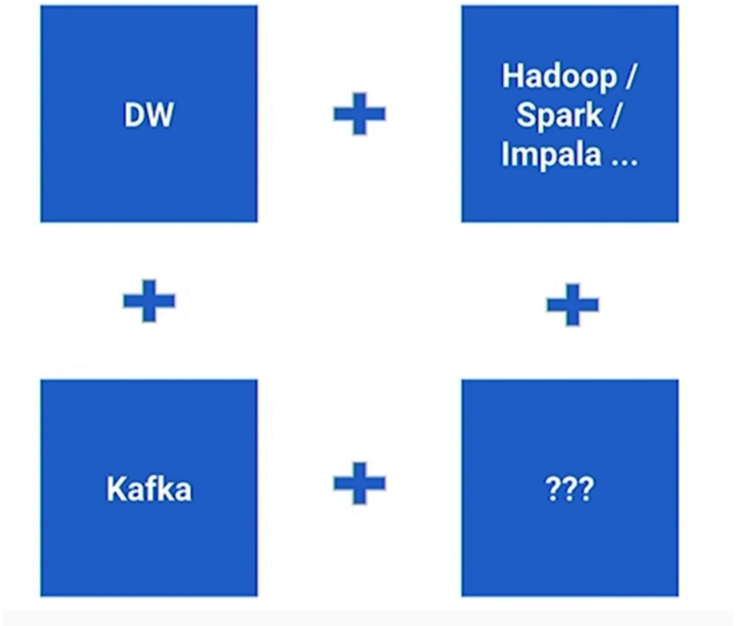


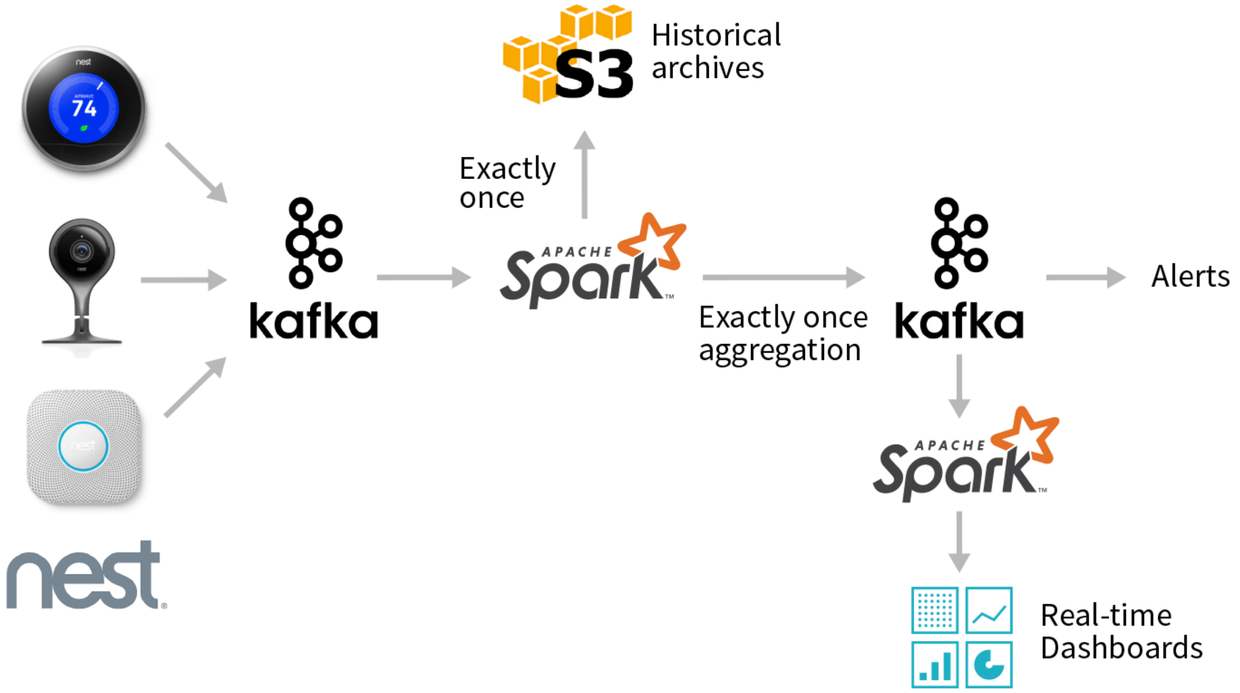
Google Next 19



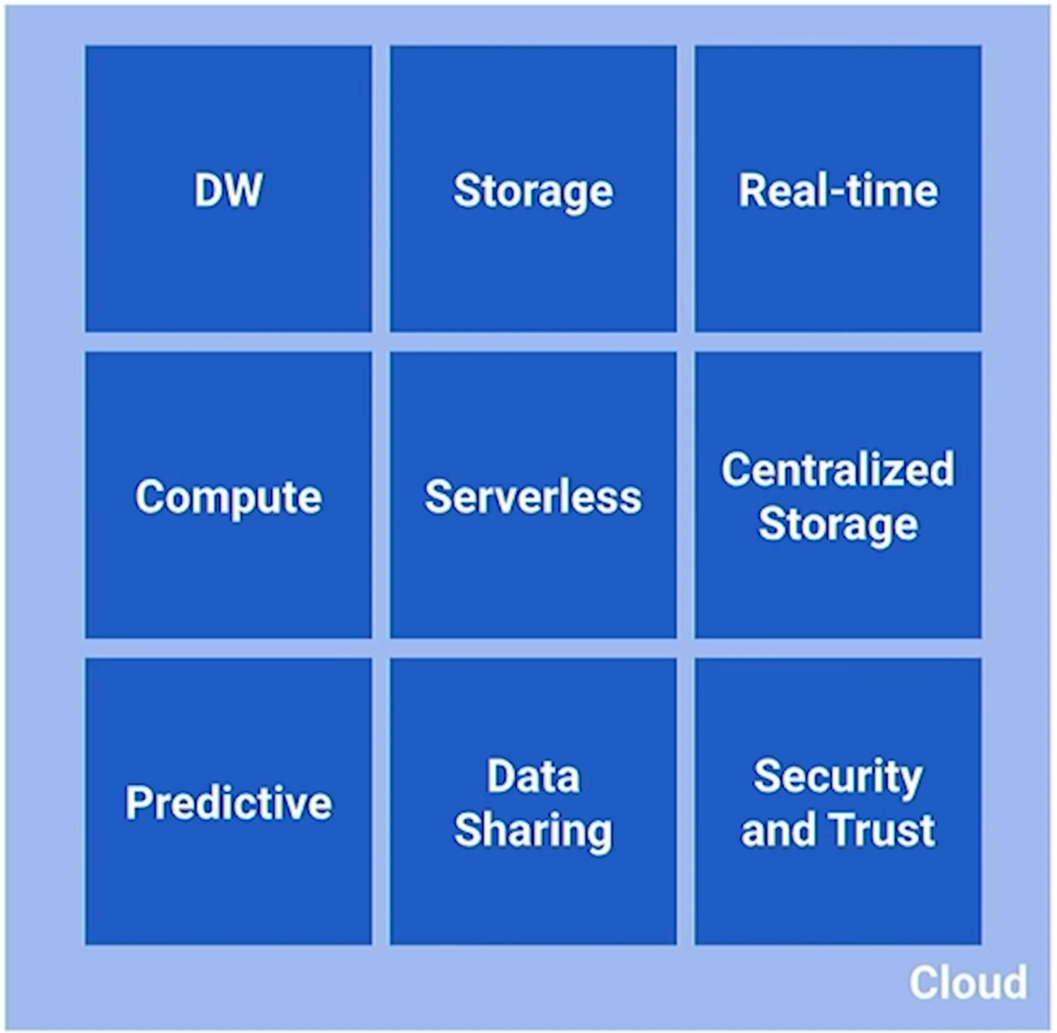
Youtube video: watch?v=eOQ3YJKgvHE

Traditional data warehousing



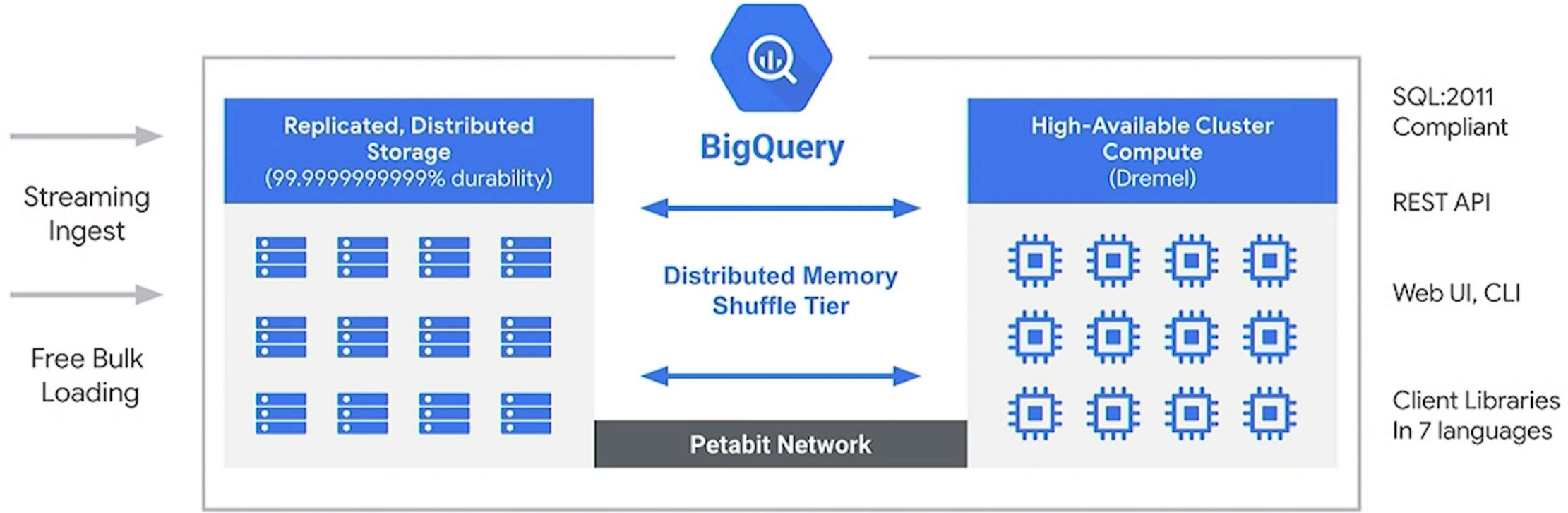


Modern data warehousing



BigQuery | Architecture

Decoupled storage and compute for maximum flexibility



Modern data warehousing

BigQuery ML

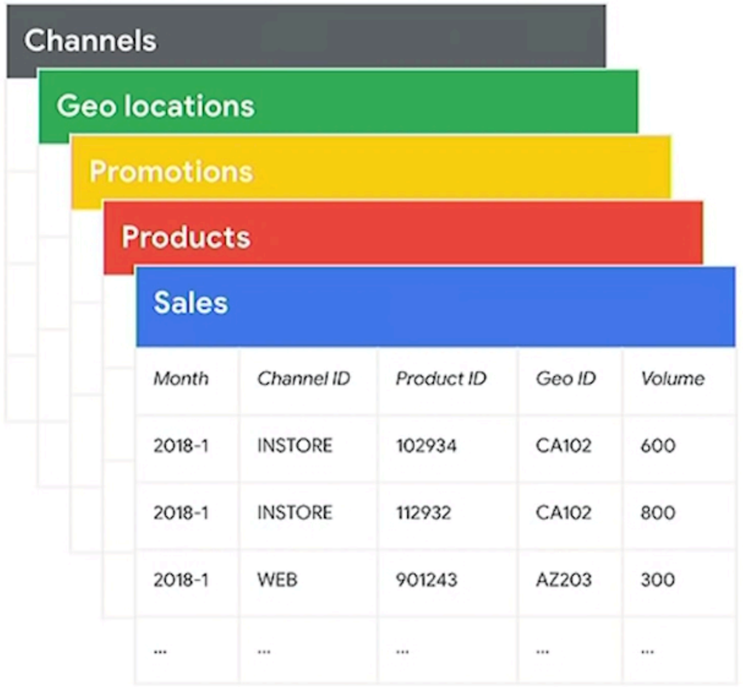
- **BigQuery ML (GA coming soon)** to build ML models (regressions)
- **K-means Clustering (Beta)** to build customer segmentations
- **Matrix Factorization (Alpha)** to build product recommendations
- **Import Tensorflow models (Alpha)** for predictions in BigQuery
- **Build Tensorflow DNN models (Alpha)** in BigQuery

1. **Execute** ML initiatives without moving data from BigQuery
2. **Iterate** on models in SQL in BigQuery to increase development speed
3. **Automate** common ML tasks and hyperparameter tuning

AutoML, what do I want to predict from a table?

AutoML Tables

Start with **raw tabular data**



Demand forecasting

Stock-out prediction

Price optimization

Customer lifetime value

Predict customer conversion / churn

Fraud prevention

and more...

- Build state-of-the-art models automatically
- Enriched treatment for a wide range of data primitives (#s, text, etc.)
- Gracefully handle datasets at BigQuery scale (currently up to 10TB)
- Code-less graphical UI for the full ML lifecycle

AutoML Tables vs. BigQuery ML

These are complementary, not competing products

AutoML Tables	BigQuery ML
For problems that require best-in-class accuracy that is fully automated	For problems that require fast experimentation and development time, and explainability (e.g., simpler models like logistic regression, trees)
Discovers the best model for the problem	Supports a variety of models
Code-less graphical UI	SQL interface
Consistent experience for users that used other AutoMLs	Will support AutoML Tables as a model_type in the future

Customers


THE HOME DEPOT


HSBC


sky news


telegraphmediagroup


Heathrow


Spotify


The New York Times

2

Features

Features

Serverless: Google does all resource provisioning.

Real-time analytics: (Near real-time)

Automatic high availability: Replicated storage in multiple locations

Standard SQL: Reduces code rewrites (Columnar database)

Federated query and logical data warehousing: Several external sources

Storage and compute separation

- Every column is stored in a separated file.

Meant for immutable pretty large datasets. (Not a transactional DB)

Features

Automatic backup and easy restore: Seven-day history of changes.

Geospatial data types and functions: SQL support for GeoJSON and WKT.

Data transfer service

Big data ecosystem integration: Dataproc and Dataflow connectors

Petabyte scale

Flexible pricing models

Data governance and security: Security with fine-grained identity

Geoexpansion: Geographic data control.

Features

Foundation for AI

Foundation for BI

Flexible data ingestion

Programmatic interaction: REST API

Rich monitoring and logging with Stackdriver

Additional Features

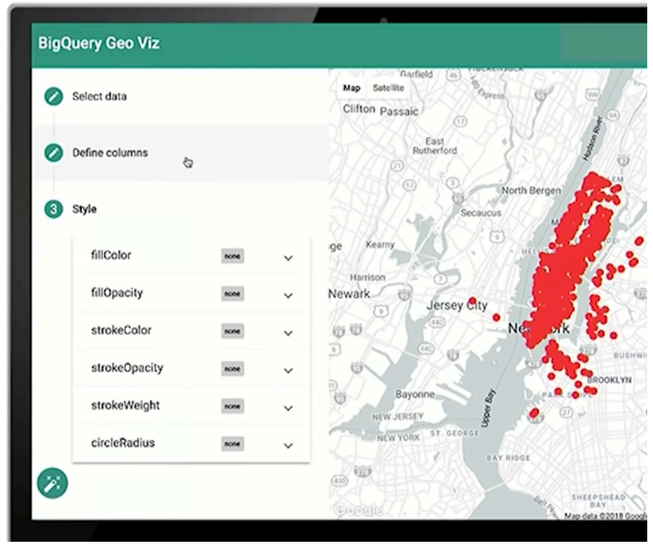
BigQuery ML (beta)

BigQuery GIS

Analyze GIS data in BigQuery with familiar SQL

BigQuery GIS

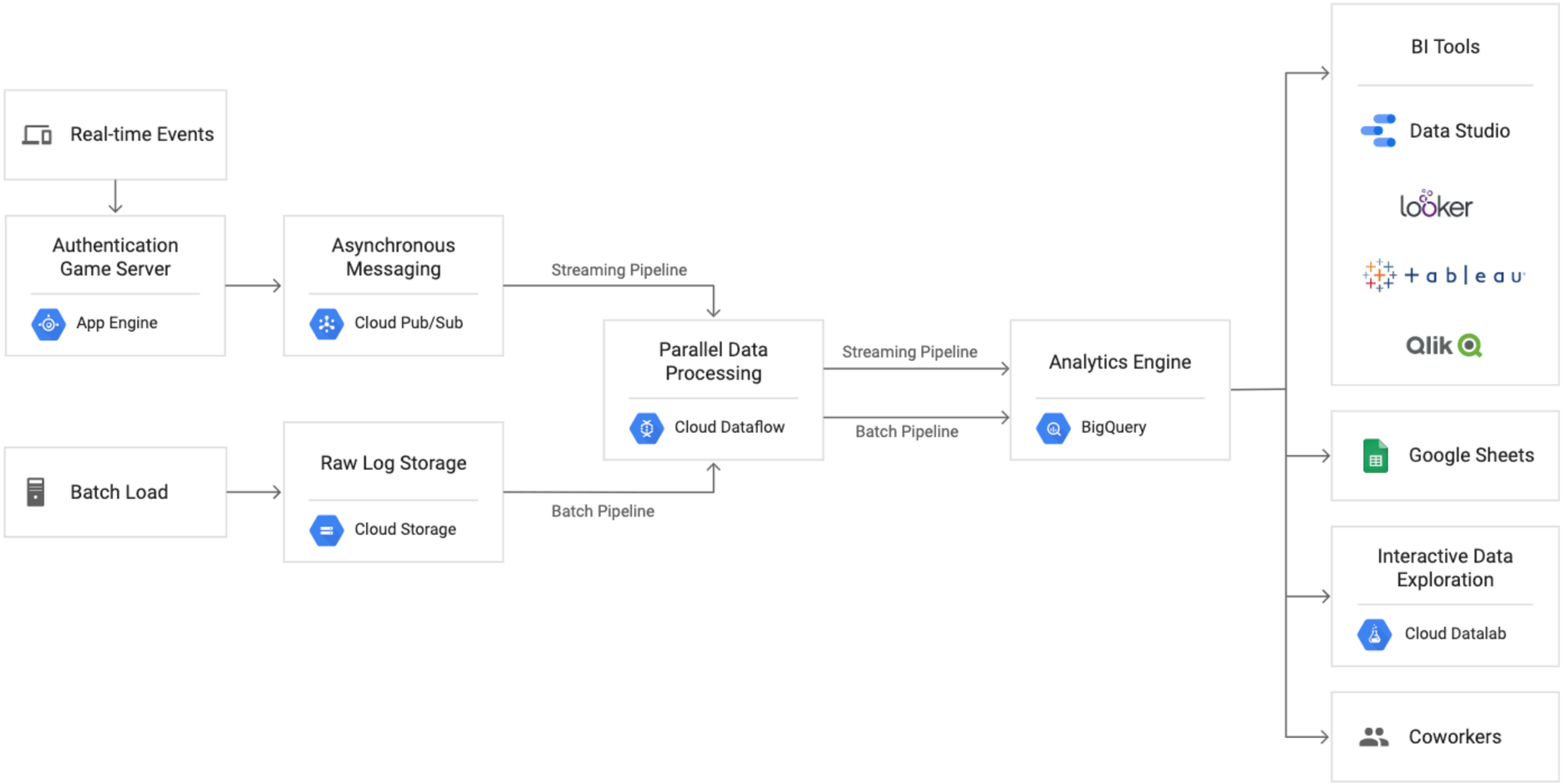
- Accurate spatial analyses with Geography data type over GeoJSON and WKT formats
- Support for core GIS functions – measurements, transforms, constructors, etc. – using familiar SQL



3

Architecture

Data warehousing architecture



4

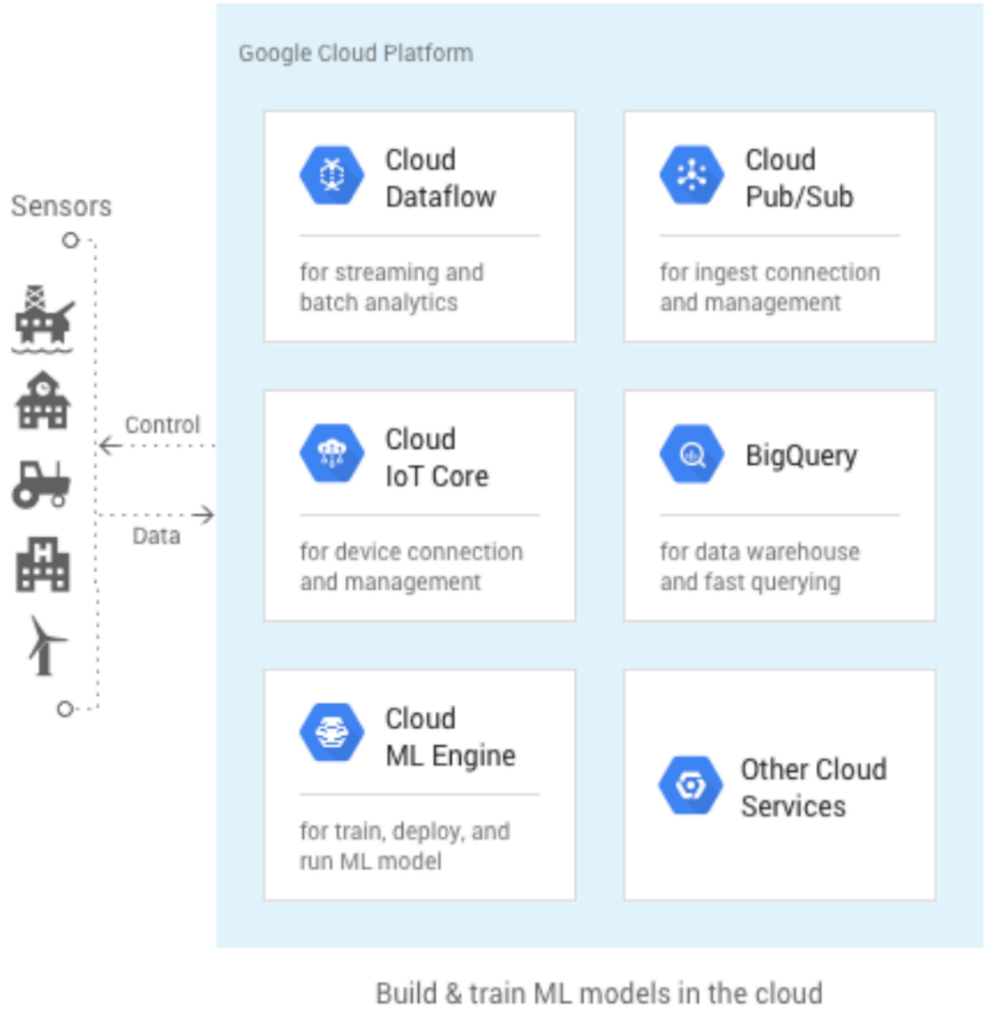
Use cases

Internet of Things

Google Cloud IoT is a complete set of tools to connect, process, store, and analyze data both at the edge and in the cloud.

Use cases:

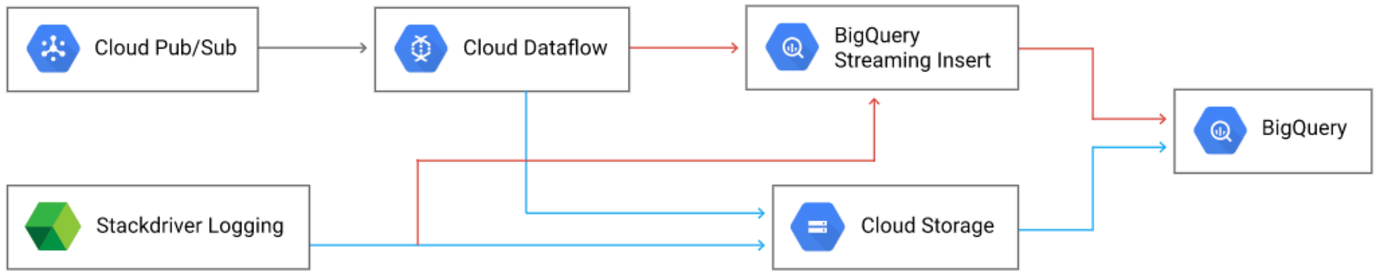
1. Predictive maintenance
2. Real-time asset tracking
3. Logistics & supply chain management
4. Smart Cities & Buildings



Optimizing Large-Scale Ingestion of Analytics Events and Logs

Benefits:

1. Log integrity. No logs are lost due to streaming quota limits or sampling.
2. Cost reduction. Logs inserted in Cloud Storage using batch jobs.
3. Reserved query resources. Moving lower-priority logs to batch loading.



Cold path:

Batch process.

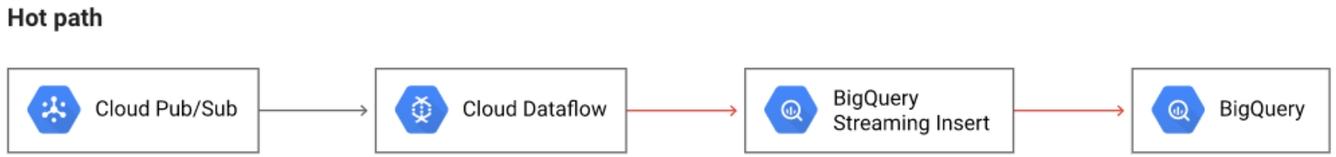
Hot path:

Streaming input.

Large-scale events and log analytics

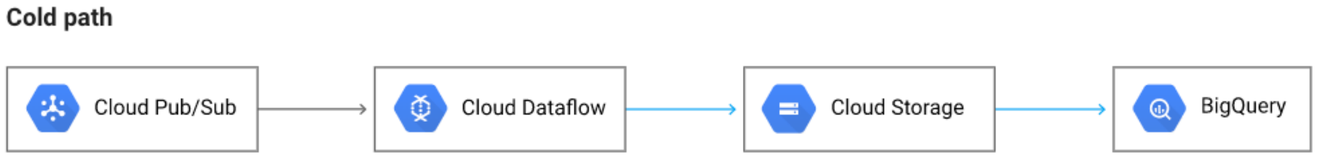
Hot path:

Immediate analysis, e.g., an event might indicate undesired client behavior or bad actors



Cold path:

Events that need to be tracked and analyzed on an hourly or daily basis



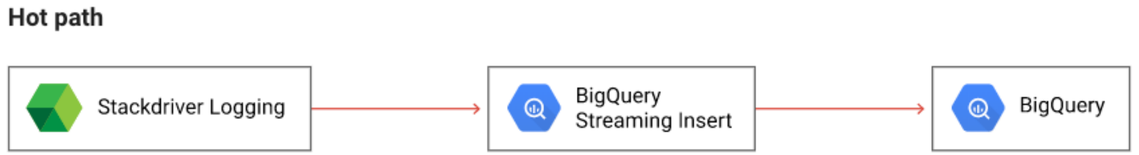
Large-scale events and log analytics

Hot path:

Critical logs.

Cold path:

Don't require near real-time analysis



5

Datasets

Commercial Datasets

Commercial data providers are accelerating your time to insight by hosting their data offerings directly in BigQuery, Cloud Storage and Cloud Pub/Sub.

Examples:

- AccuWeather: min/max temperatures, precipitation, and snowfall
- Dow Jones: historical News Snapshots
- HouseCanary: historical home price
- Xignite: financial market data

Public datasets

A public dataset is any dataset that is stored in BigQuery and made available to the general public through the Google Cloud Public Dataset Program.

Weather and climate

Understand how weather impacts your business.

Cryptocurrency

Start analyzing cryptocurrency blockchains.

Healthcare and life sciences

Improve patient care and accelerate discovery of new treatments.

Transportation

Understand how transportation affects retail and food distribution.

6

Costs

7

Examples

Medium: williamegomezo

williamegomezo.me

HUGE

Done.

Getting started with BigQuery

September 11, 2017