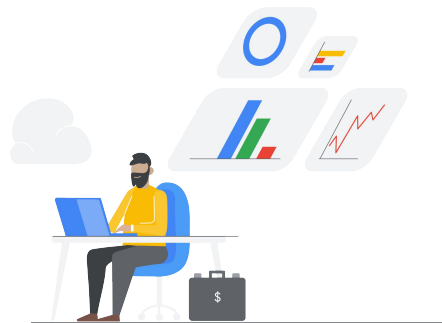Google Cloud

GCP & the self-sufficient quant
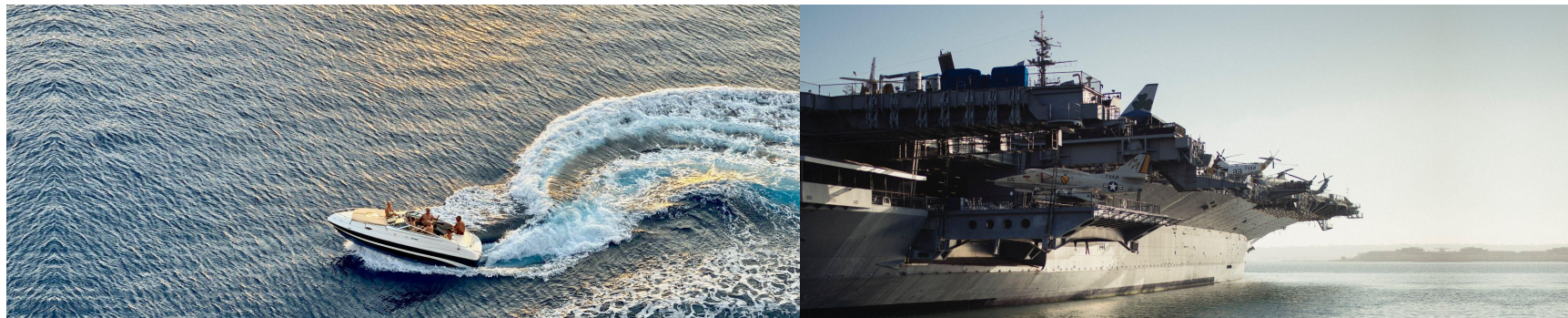
Ashish Majmundar

Director, Global Head of Capital Markets, Google Cloud

May 2021

Enterprise ML models in production

How an aircraft carrier can move at the speed and agility of a speedboat

Quant desktop

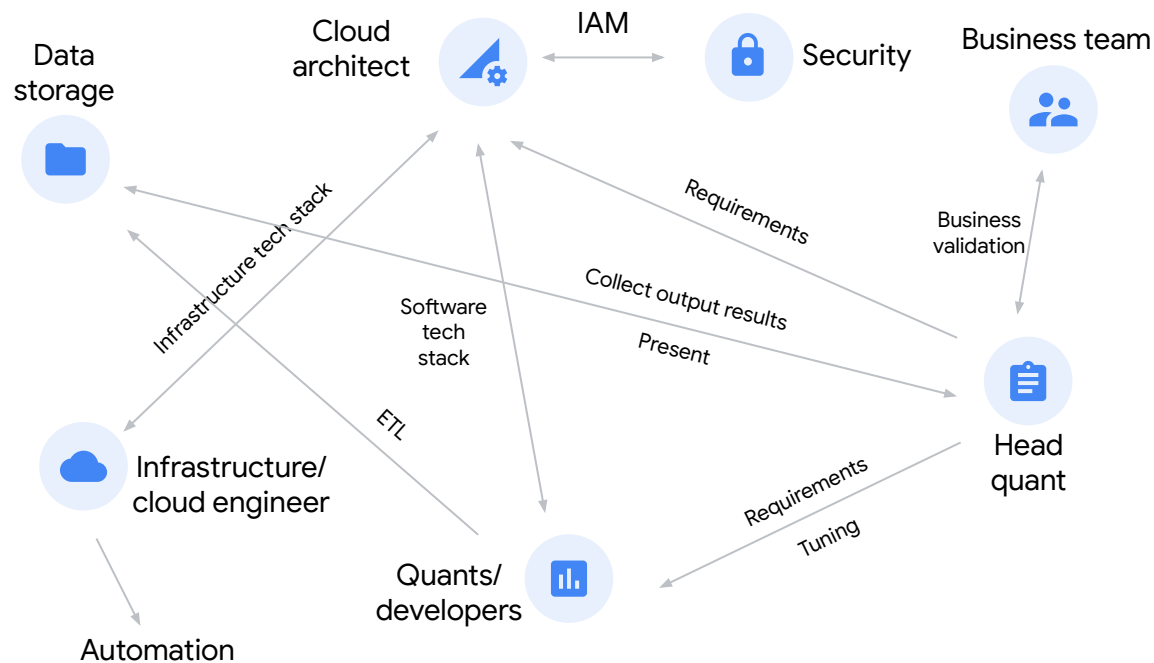| Desktop or few servers | Hundreds of nodes |
Sample data	Petabytes of data
Few instruments	Thousands of instruments
Single model	Universe of models
Single user	Multi-tenant
Low governance	High governance / Explainable AI
Approval of head quant	Enterprise approval

Enterprise ML models in production

Google Cloud

Why is it so hard?

For every project:

- Collaboration between multiple parties
- Long-time delays
- Competing priorities
- Error prone
- Difficult to scale

Rinse & repeat

Data storage

Cloud architect

IAM

Security

Business team

Infrastructure/cloud engineer

Automation

Quants/developers

Head quant

Infrastructure tech stack

Software tech stack

ETL

Requirements

Collect output results

Present

Business validation

Requirements

Tuning

How do you scale?

1

Data management

Provide the quant team a centralized and consistent view of the data model, versus the legacy fragmented and scattered view across data stores

2

Model development environment

Enable a fleet of quants from various business teams to concurrently develop and collaborate to develop the best of the breed models for the organization

3

Model optimization & training environment

An automated factory model to enable parallel training of various models and provide a structured approach for model assessment

4

Model validation environment

Second level of model optimization, where expert quants add value to model building, resulting in a more perfected model

5

Model serving environment

An enterprise standard and centralized model execution platform created via specifications, versus multiple ad hoc systems stood up manually for various model types

6

Model performance assessment

A standard and structured approach for assessing models in production versus legacy ways of analyzing multiple reports, spreadsheets, and email exchanges

Google Cloud

How do you make it happen?

1

Data management

Provide the quant team a centralized and consistent view of the data model, versus the legacy fragmented and scattered view across data stores

2

Model development environment

Enable a fleet of quants from various business teams to concurrently develop and collaborate to develop the best of the breed models for the organization

3

Model optimization & training environment

An automated factory model to enable parallel training of various models and provide a structured approach for model assessment

Overlaying a governance and operating model

4

Model validation environment

Second level of model optimization, where expert quants add value to model building, resulting in a more perfected model

5

Model serving environment

An enterprise standard and centralized model execution platform created via specifications, versus multiple ad-hoc systems stood up manually for various model types
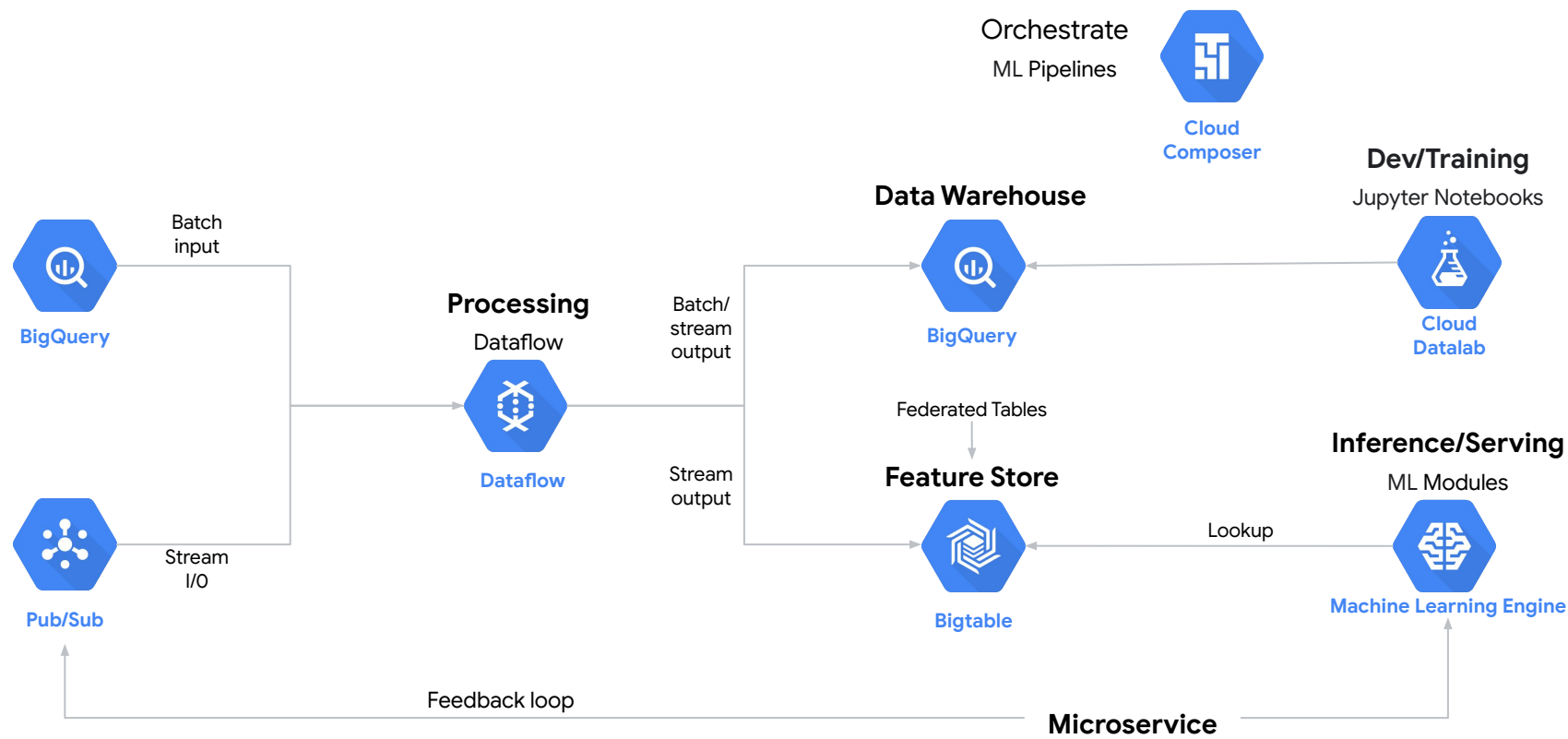
6

Model performance assessment

A standard and structured approach for assessing models in production versus legacy ways of analyzing multiple reports, spreadsheets, and email exchanges

Google Cloud

Speeding up the operating model

Operating model: RACI automation

	Business					Technology				
	Line of Business	Data Scientists / Developers	Head Quant	Quant Team	Model Validation Team	Enterprise Architecture	Technology Infrastructure	Platform Engineering	Cloud Team	Security
Data Management										
Centralized Data Storage Requirements	I	C	C	A	C	C	C	R	C	C
Review Solution Approaches	I	C	C	A	C	C	C	R	C	C
Evaluate an…										R
Build sa…										C
Data In…										C
Review …										C
Evaluate…										R
Build pr… LOB/ten…										C
Data Eg…										C
Review So…										C
Evaluate and Scorecard Solutions	I	C	C	A	C	R	R	R	R	R
Build prototype for chosen dataset of specific LOB/tenant	I	C	C	A	C	C	R	R	C	C
Data Catalog Requirements	I	C	C	A	C	C	C	R	C	C
Review Solution & Obtain Signoff	I	C	A	C	C	C	C	R	C	C
Evaluate and Scorecard Solutions	I	C	C	A	C	R	R	R	R	R
Build prototype for chosen dataset of specific LOB/tenant	I	C	C	A	C	C	R	R	C	C

Scalable ML architecture

Orchestrate
ML Pipelines

Cloud Composer

Dev/Training
Jupyter Notebooks

Cloud Datalab

Data Warehouse

BigQuery

Batch input

BigQuery

Processing
Dataflow

Dataflow

Batch/ stream output

Federated Tables

Inference/Serving
ML Modules

Stream output

Feature Store

Lookup

Bigtable

Machine Learning Engine

Stream I/O

Pub/Sub

Feedback loop

Microservice

Google Cloud

Let's discuss ML @ scale!

Ashish Majmundar

majmundar@google.com