從雲端代理商，看未來 AI 技術的應用發展

2022.05.06
余佑駿 Youjun
@ 台灣資料科學社群 TWDS

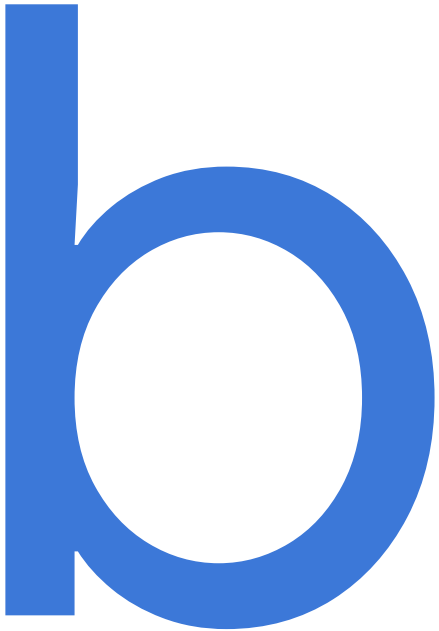
Medium 分享記錄　FB 活動頁面　FB 直播連結

Today's Agenda

Who am I	01
About Cloud Industry	02
How to become a solution architect	03

- 甚麼是雲端
- 雲端原廠(Cloud Provider) vs 雲端代理商
- 地盤怎麼分

Today's Agenda



AI/ML in Cloud 04

- Google Cloud
- Amazon

How to Build a AI/ML Kingdom 05

What's Next 06

Who am I

- 熱愛資料科學
- Coding 是興趣, 也是一輩子的事
- 現為雲端架構師

01

Google Cloud

余佑駿 Youjun
Cloud Architect

littlefish0331
you-jun-yu
littlefish0331
littlefish0331@gmail.com

經歷】

2021 宏庭股份有限公司 Microfusion Technology
GCP解決方案部 雲端架構師

2019 國家高速網路與計算中心 資料科學家

2017 國立政治大學 統計所

2012 國立清華大學 數學系輔修教育學程

競賽/Side-Project/演講】

2022 (GDSC-TKU)
Introduction to Data Visualization & Data Studio

2021 (GDG)
How to Migrate Your Google Drive to Cloud Storage

2021 (Tbrain) Tomofun 狗音辨識: 4th + 評審獎

2019 (Kaggle) LANL Earthquake Prediction: top 23%

塵世中一個迷途小書僮，從資料科學走向雲端科技。

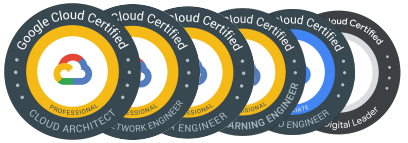
平時喜歡數據分析、機器學習、參與社群活動。

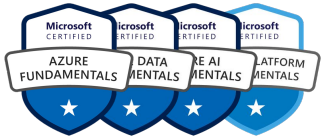
希望用科技創造更美好的世界。

Linktree
linktr.ee/youjun

履歷、專案
社群分享

可以交流
考證照經驗





About Cloud Industry

- 什麼是雲端
- 雲端原廠(Cloud Provider) vs 雲端代理商
- 地盤怎麼分

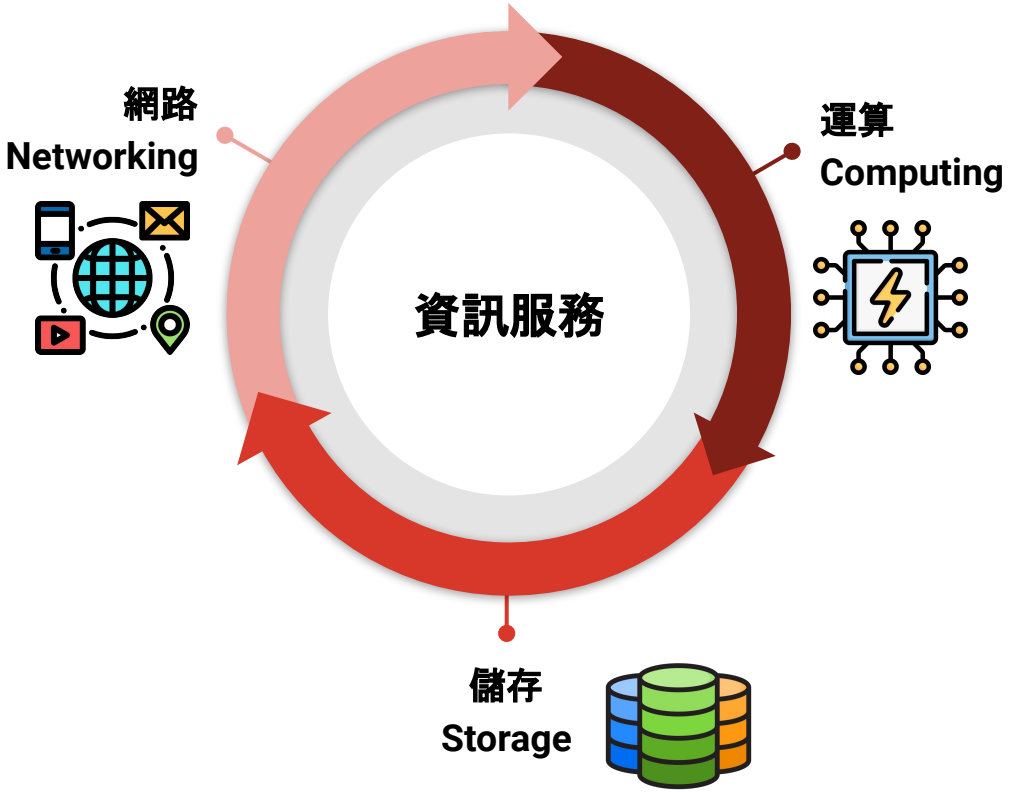
02

Google Cloud

什麼是雲端

資訊服務，本質不外乎
「運算、儲存、網路」

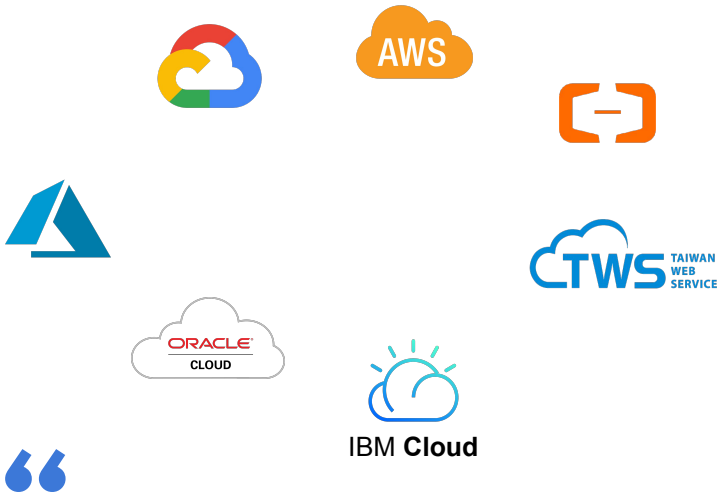
雲端服務也是，但雲端的崛起，
也歸功於科技的進步，並驅使各
行各業做數位轉型



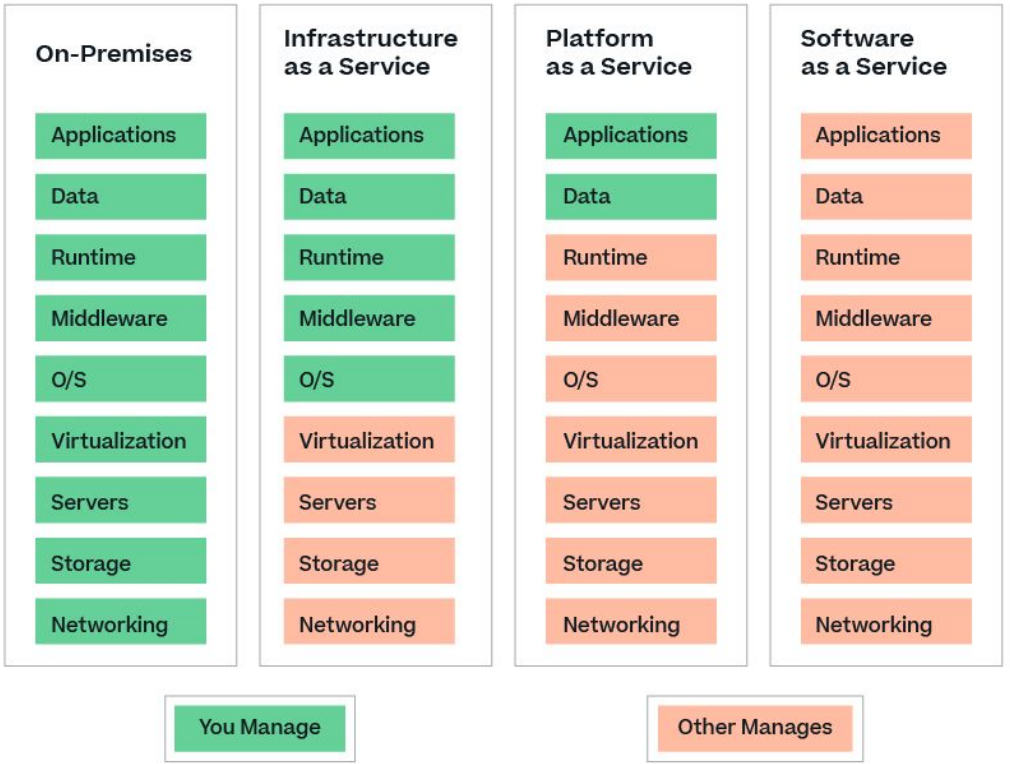
* 價格通常也分這三塊，有些要多考慮「版本授權」

什麼是雲端 (cont.)

「以 IaaS 為基礎，提供 PaaS、SaaS 等服務」



> 歲月靜好，
> 是有人替你負重前行



什麼是雲端 (cont.)

「以 IaaS 為基礎，提供 PaaS、SaaS 等服務」

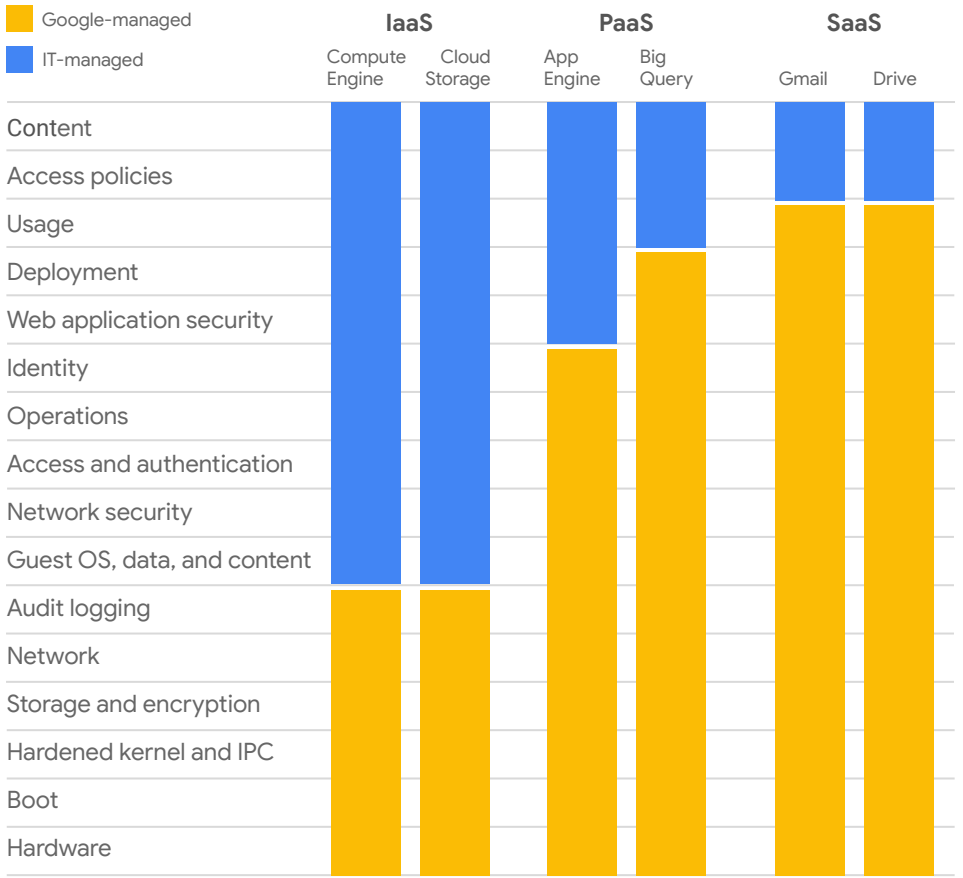
btw, 雲端的安全由供應商和客 戶共同承擔

Cloud security is a
“shared responsibility”

We are responsible for securing
the infrastructure

You are responsible for securing
your data

We help you with best practices,
templates, products, and solutions



原廠 vs 代理商

雲端原廠/供應商的功能

- 產品服務的創新
- 技術的研發
- 資源的生產、製造

第一時間參與科技的開發

與國外團隊合作，經驗 up

更多跨國、大規模的數位轉型情境

雲端代理商的功能

- 幫忙推廣產品
- 協助客戶打造客製化的雲端架構
- 「研發、生產、製造」以外的項目

還是可以開發自有 產品

享受最新的技術、解決問題

貼近客戶真實場景，應用落地

原廠 vs 代理商

雲端原廠/供應商的功能

- 產品服務的創新
- 技術的研發
- 資源的生產、製造

第一時間參與科技的開發

與國外團隊合作，經驗 up

更多跨國、大規模的數位轉型情境

雲端代理商的功能

- 幫忙推廣產品
- 協助客戶打造客製化的雲端架構
- 「研發、生產、製造」以外的項目

還是可以開發自有 產品

享受最新的技術、解決問題

貼近客戶真實場景，應用落地

“

Consulting

原廠 vs 代理商

雲端原廠/供應商的功能

- 產品服務的創新
- 技術的研發
- 資源的生產、製造

第一時間參與科技的開發

與國外團隊合作，經驗 up

更多跨國、大規模的數位轉型情境

Consulting 的 Consulting

雲端代理商的功能

- 幫忙推廣產品
- 協助客戶打造客製化的雲端架構
- 「研發、生產、製造」以外的項目

還是可以開發自有 產品

享受最新的技術、解決問題

貼近客戶真實場景，應用落地

Consulting

原廠 vs 代理商 (cont.)

風光產業下的酸「?」苦辣

- 「甜」咧?
- 夾在中間

客戶的需求能否滿足？

Google 沒有這個服務怎麼辦？

我可以推其他原廠的服務嗎?

原廠 vs 代理商 (cont.)

風光產業下的酸「?」苦辣

- 「甜」咧?
- 夾在中間

客戶的需求能否滿足？

Google 沒有這個服務怎麼辦？

我可以推其他原廠的服務嗎?

> **不打壞合作關係的情況下，客戶至上**

地盤怎麼分

- AWS 是雲端市占率的龍頭
- Microsoft 第二；Alibaba 第三
- Alibaba 亞洲第一
- Google 的市占成長率最高

Table 1. Worldwide IaaS Public Cloud Services Market Share, 2019-2020 (Millions of U.S. Dollars)

Company	2020 Revenue	2020 Market Share (%)	2019 Revenue	2019 Market Share (%)	2019-2020 Growth (%)
Amazon	26,201	40.8	20,365	44.6	28.7
Microsoft	12,658	19.7	7,950	17.4	59.2
Alibaba	6,117	9.5	4,004	8.8	52.8
Google	3,932	6.1	2,367	5.2	66.1
Huawei	2,672	4.2	882	1.9	202.8
Others	12,706	19.8	10,115	22.1	25.6
Total	**64,286**	**100.0**	**45,684**	**100.0**	**40.7**

Source: Gartner (June 2021)

地盤怎麼分 (cont.)

AWS

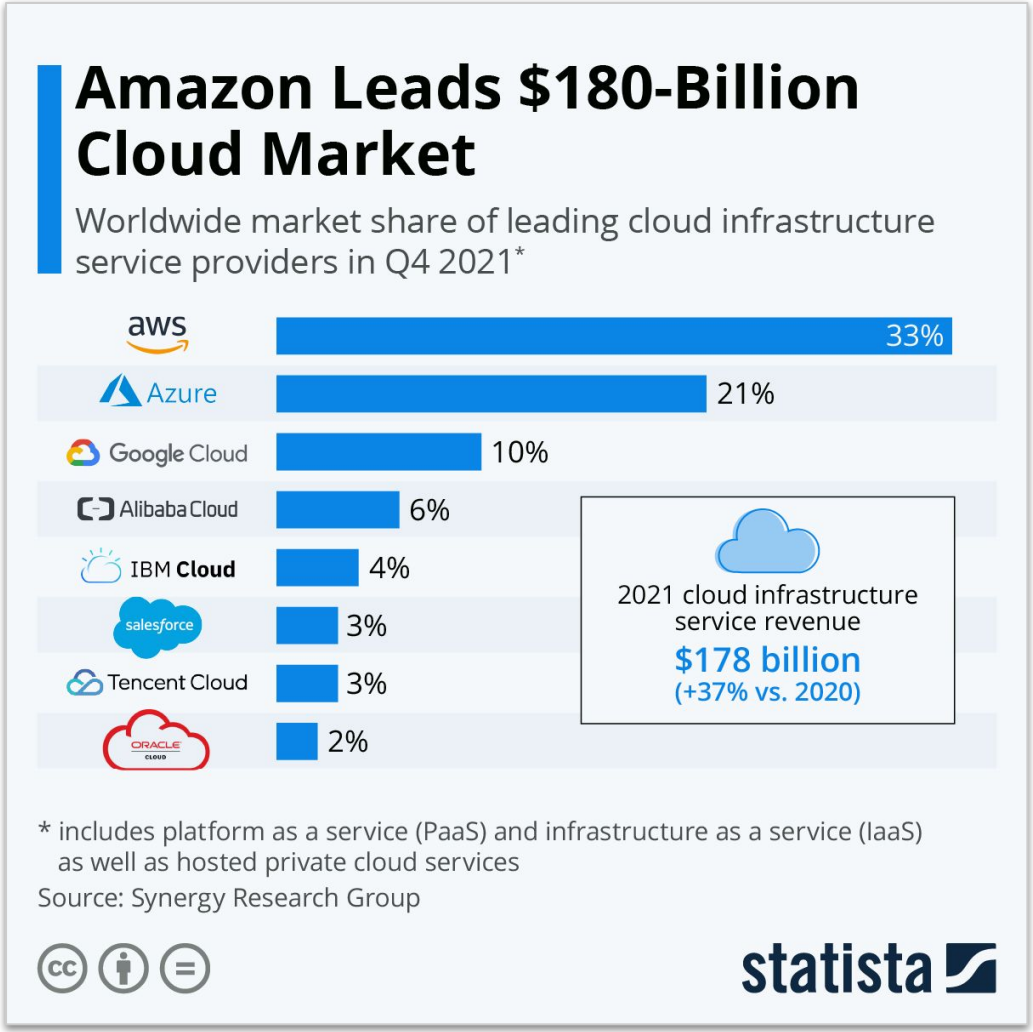
- 2006年推出第一個服務 S3，同年推出 EC2
- 主打中小企業，提供解決方案
- 關注於公有雲

Azure

- 2010年正式推出，首要服務有VM、SQL、Storage
- Windows 深植人心
- 專注與既有的數據中心混合使用

GCP

- 2008年推出第一個服務 App Engine
- 技術與開發能力著稱
- 專注於多雲的管理



地盤怎麼分 (cont.)

地盤怎麼分 (cont.)

地盤怎麼分 (cont.)

地盤怎麼分 (cont.)

地盤怎麼分 (cont.)

Google 怎麼做

- 和系統整合商合作
  (打不贏就買、欣賞就收購)
- 人工智慧是賣點
- 技術開源持續領先
  (K8s, istio, Tensorflow)
- 著眼混合雲(Hybrid Cloud)的解決方案(Anthos)

Table 1. Worldwide IaaS Public Cloud Services Market Share, 2019-2020 (Millions of U.S. Dollars)

Company	2020 Revenue	2020 Market Share (%)	2019 Revenue	2019 Market Share (%)	2019-2020 Growth (%)
Amazon	26,201	40.8	20,365	44.6	28.7
Microsoft	12,658	19.7	7,950	17.4	59.2
Alibaba	6,117	9.5	4,004	8.8	52.8
Google	3,932	6.1	2,367	5.2	66.1
Huawei	2,672	4.2	882	1.9	202.8
Others	12,706	19.8	10,115	22.1	25.6
Total	**64,286**	**100.0**	**45,684**	**100.0**	**40.7**

Source: Gartner (June 2021)

地盤怎麼分 (cont.)

以ML為切入，看看官網

AWS

- 以需求出發
- 鎖定特定領域提供 Solution，先解決 80% 的問題。
- 開發平台 SageMaker[1]，偏向開發者；但近期有推出 Canvas[2]

產業

ML Solutions Lab 已經成功為世界各地各行各業的客戶提供協助，包括生產製造、醫療保健和生命科學、金融服務、體育、公共部門和汽車，以建立採用機器學習技術的全新解決方案。

生產製造

ML Solutions Lab 專精於異常偵測、預測等方面的專業知識，可以協助製造公司改善其核心生產、維護、安全和品質，以及研發和供應鏈功能。例如，ML Solutions Lab 協助 Formosa Plastics 運用機器學習更準確地偵測缺陷，並將員工花費在手動檢查上的時間減少了一半。

醫療保健與生命科學

ML Solutions Lab 已使用 ML 協助醫療保健與生命科學領域的客戶降低成本並改善患者照護。例如，ML Solutions Lab 與 Cerner 合作建立解決方案，以便研究人員分析匿名患者資料，從而開發能在臨床表現之前最長 15 個月預測出充血性心力衰竭的演算法。

金融服務

ML Solutions Lab 與銀行、投資組織、保險公司和抵押公司等金融機構合作，以改善預測，讓監視系統能夠標記新的或正在出現的威脅，針對金融產品產生個人化建議，自動化文件處理，以及改善客戶使用聊天機器人和對話界面的體驗。

運動

ML Solutions Lab 擁有與 Formula 1、National Football League 和 Seattle Seahawks 等體育組織合作的豐富經驗，從而改善球迷的體驗並提高比賽品質。在 AWS 與 NFL 的合作中，ML Solutions Lab 確定並建立了一種全新的方式，讓球迷可以透過 Next Stats 參與運動賽事，並且目前正在運用 ML 來支援 NFL 的球員健康和安全計劃。

環境與社會影響

ML Solutions Lab 密切配合解決了一些全球最大的挑戰，從人口販運到飢荒，再到使用機器學習更充分地了解我們的世界。例如，ML Solutions Lab 與 Maxar 合作，透過機器學習更有效地分析衛星資料，讓組織能夠更快地採取行動以應對野火等災難，並向非洲村莊提供疫苗。

汽車

在汽車產業，ML Solutions Lab 與客戶在各領域密切合作，包括供應鏈最佳化、車載娛樂體驗和自動駕駛。這包括透過準確的道路情境感知和進階分析，來提高駕駛員和行人的安全。

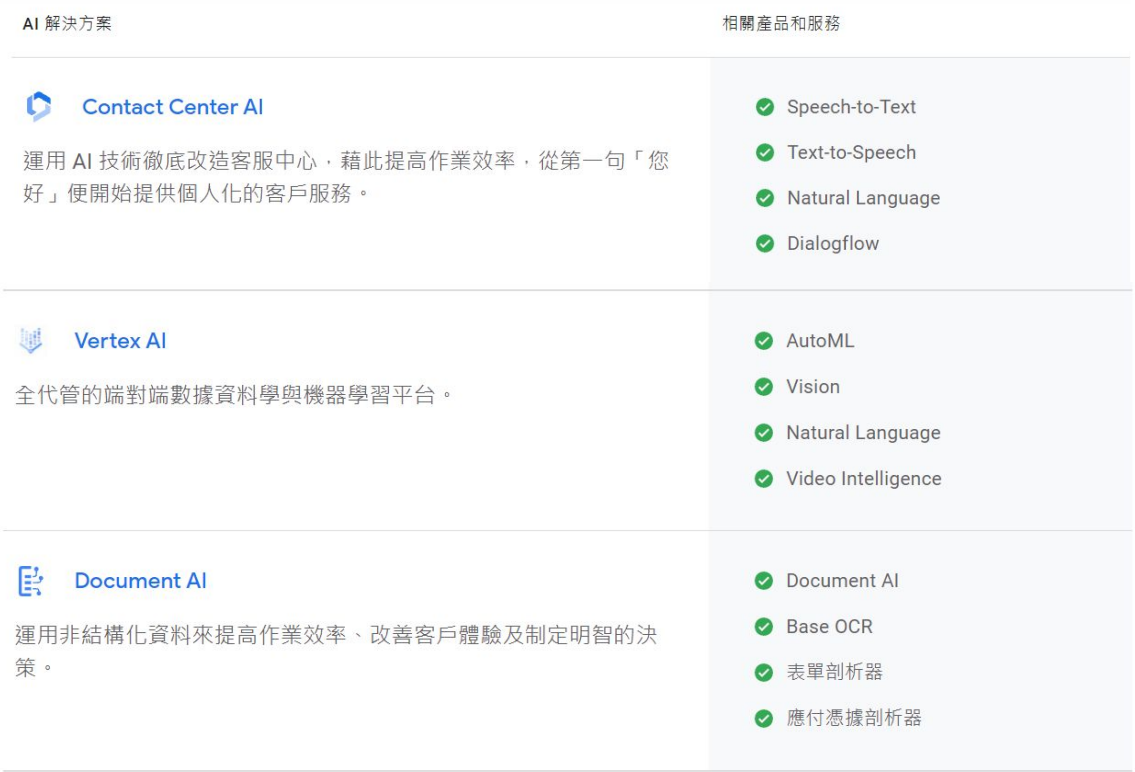
[1] Amazon SageMaker Studio – 第一個用於機器學習的DE
[2] Announcing Amazon SageMaker Canvas – a Visual, No Code Machine Learning Capability for Business Analysts | AWS News Blog

地盤怎麼分 (cont.)

以ML為切入，看看官網

Google Cloud

- 以技術出發
- 完整 Solution 鎖定的行業為**客服中心**，以及以**文件處理為主的需求**；最近有針對零售與製造業有新的 solution [1][2]
- Vertex AI 是一個完整開發平台，適用於各種角色開發、使用。



AI 解決方案	相關產品和服務
Contact Center AI 運用 AI 技術徹底改造客服中心，藉此提高作業效率，從第一句「您好」便開始提供個人化的客戶服務。	Speech-to-Text Text-to-Speech Natural Language Dialogflow
Vertex AI 全代管的端對端數據資料學與機器學習平台。	AutoML Vision Natural Language Video Intelligence
Document AI 運用非結構化資料來提高作業效率、改善客戶體驗及制定明智的決策。	Document AI Base OCR 表單剖析器 應付憑據剖析器

[1] AI and Machine Learning Solutions | Google Cloud
[2] AI & Machine Learning Products | Google Cloud

How to become a solution architect

- 雲端架構師在幹嘛

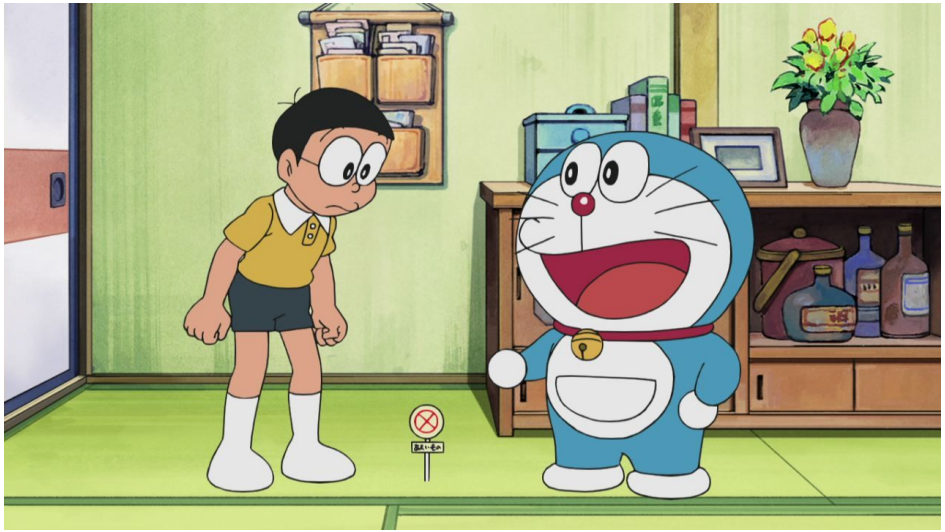
03

Google Cloud

雲端架構師在幹嘛

協助客戶做數位轉型、上雲需求、架構規劃、~~把屎把尿~~...等等

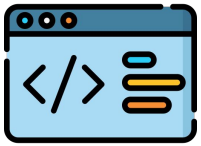
- Infra 的搬遷 (VM、Database)
- 架構設計，以符合情境需求 (High availability / Disaster Recovery / Scalability / Hybrid)
- Database Optimization
- Cost effective
- Application Refactoring
- AI/ML Enablement
- Resources Management
- Monitoring
- Security
- …



雲端架構師在幹嘛 (cont.)

系統設計
~~(積木組合)~~

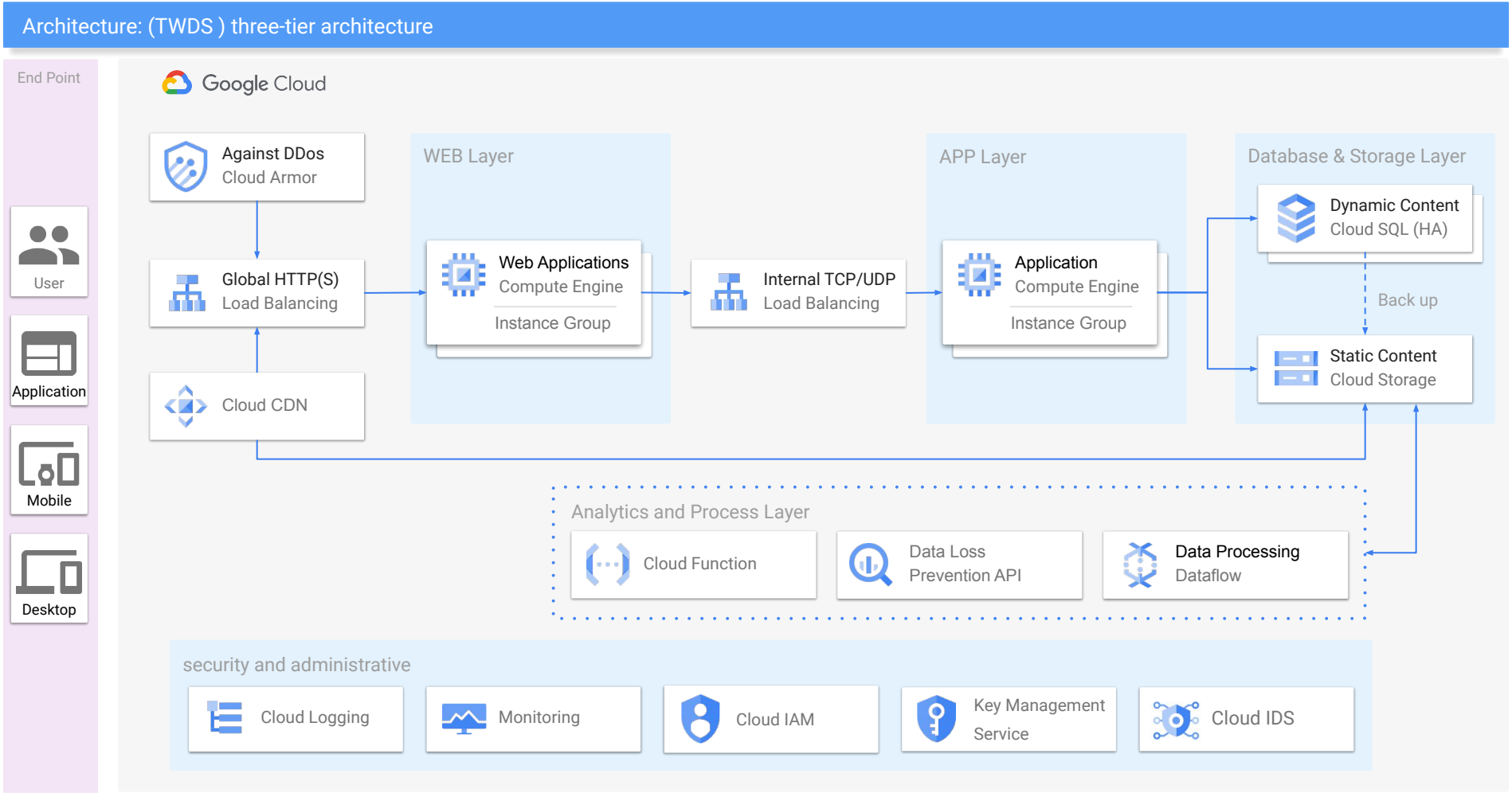
溝通
~~(通靈)~~



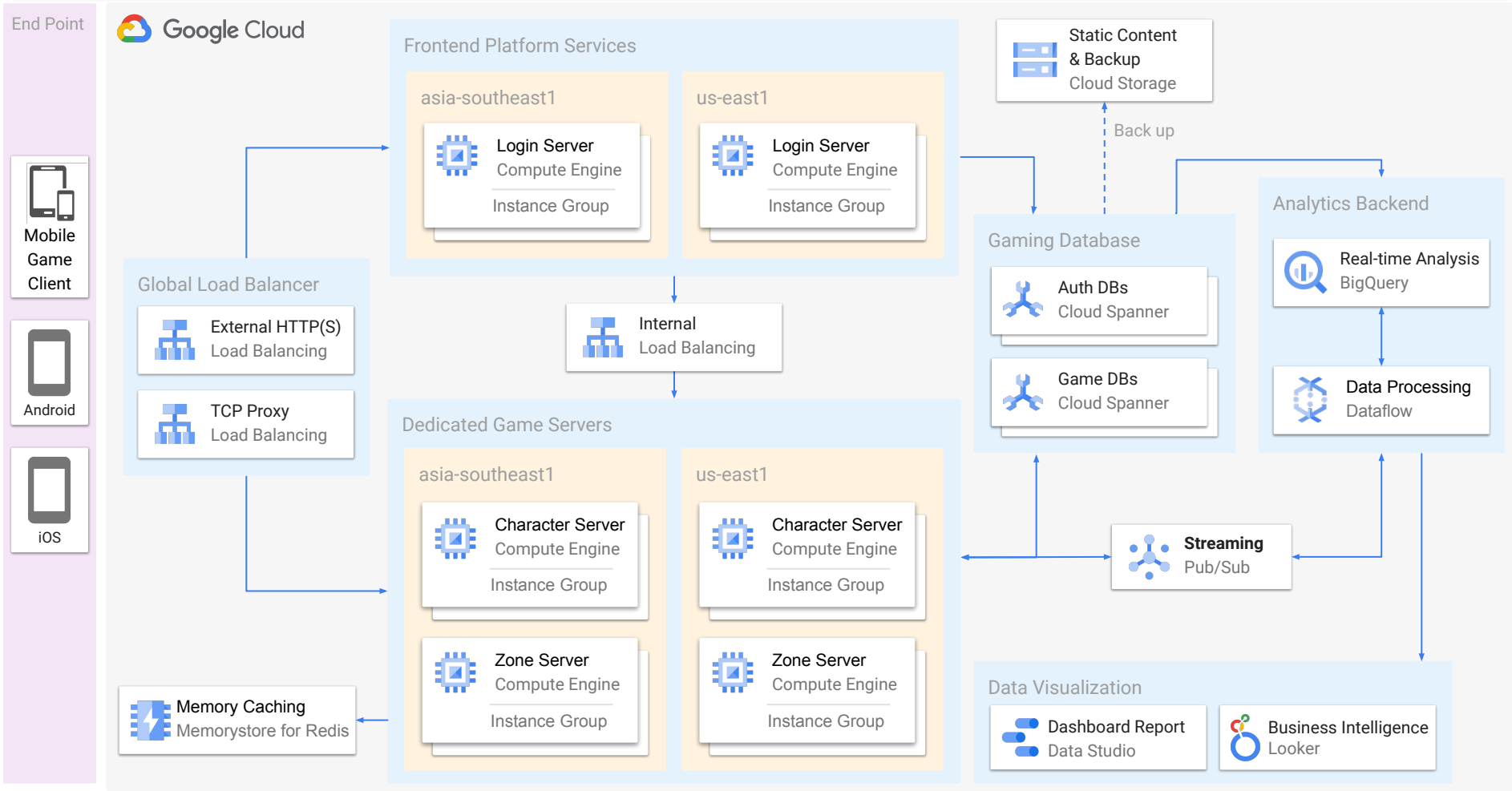
程式設計
~~(邏輯要好)~~

自學能力
~~(熱愛工作)~~

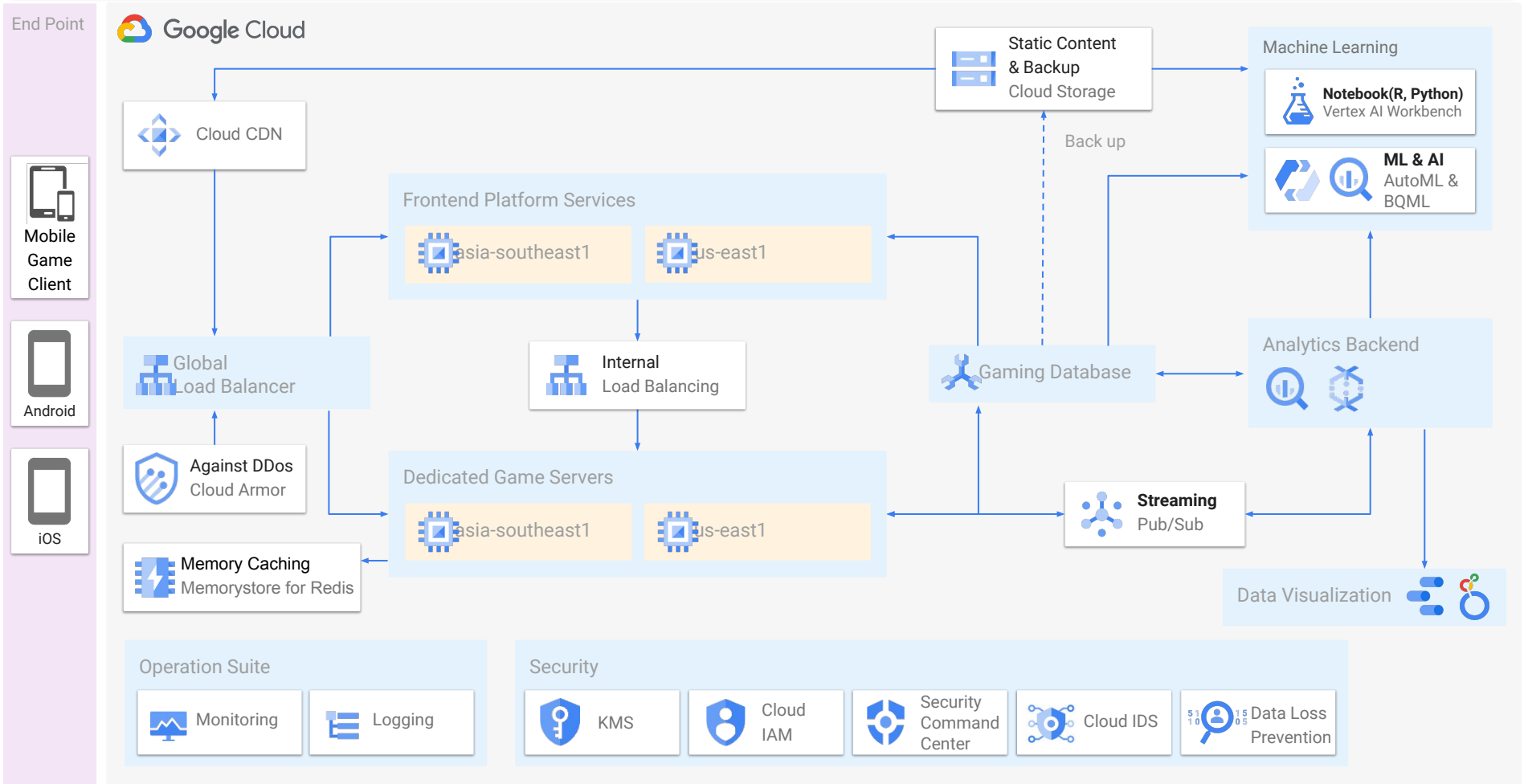
Architecture: (TWDS) three-tier architecture



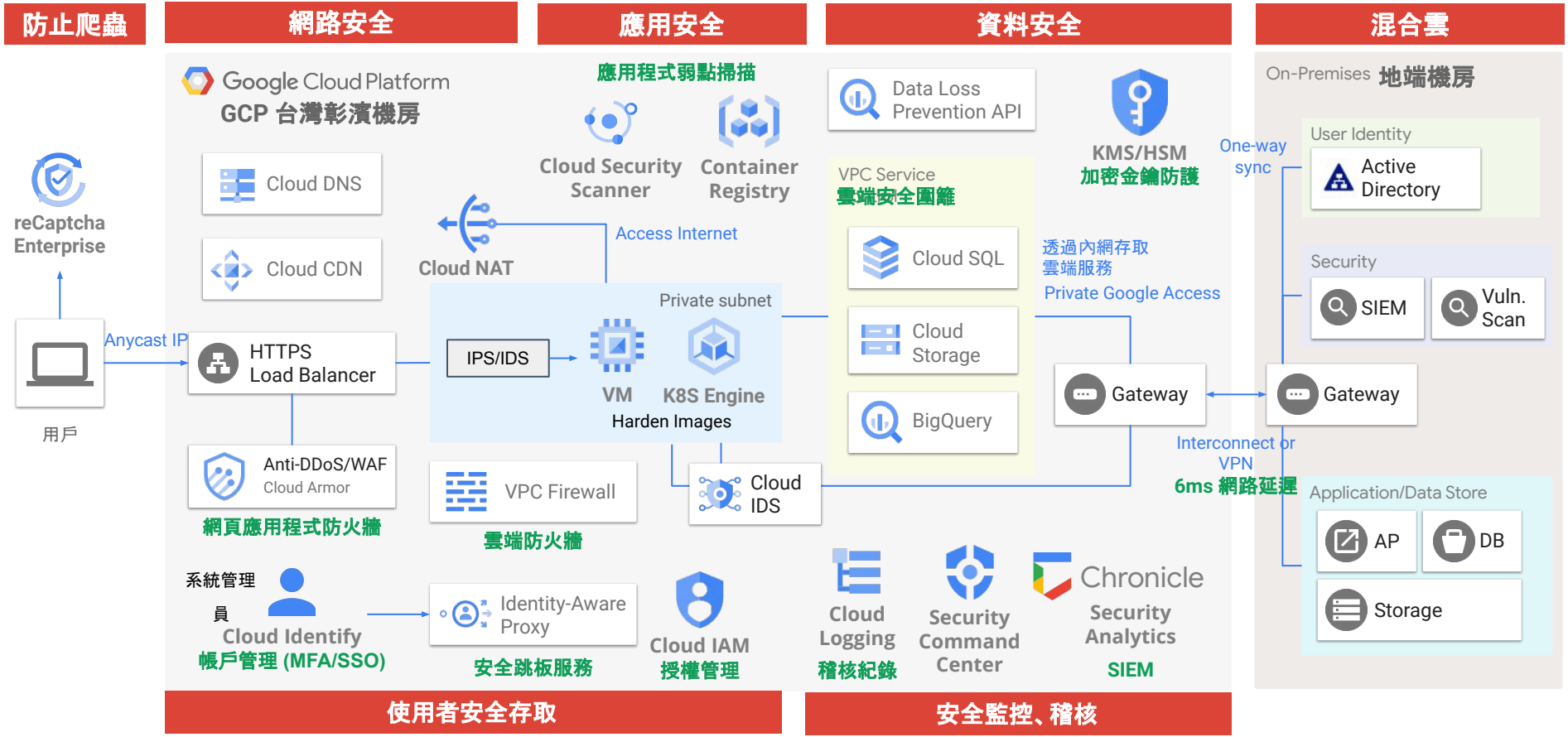
Architecture: (TWDS) MMORPG Gaming



Architecture: (TWDS) MMORPG Gaming (cont.)



Google Cloud 安全架構



AI/ML in Cloud

- Google Cloud (BQML, Vertex AI)
- Amazon Web Service (SageMaker)

Google Cloud AI/ML Strategy

Tools & Solutions

Solution	senario or tools
ML API	for tech (e.g. develop)
AutoML Services	build GUI for ML API. (e.g. document AI, NLP, image, Vision, Table, etc.)
AI Platform	integrate all AutoML Service. (ready to deprecate)
Vertex AI	integrate AutoML, Kubeflow, and TensorFlow Extended (TFX)

PaaS for tech and non-tech

Tools & Solutions

Solution	senario or tools
Contact Center AI (CCAI)	integrate Dialogflow, STT, TTS, NLP
Retail AI	integrate Vision Product Search, Recommendations AI, Retail Search

AI for every level of expertise

Services and Solutions
Ease of Implementation

Horizontal solutions
- Document Understanding AI
- Contact Center AI
- Dialogflow

Industry solutions
- Talent Solution
- Recommendation AI

Collaboration
- AI Hub

Building Blocks

APIs
Pre-trained Models

Sight
- Vision
- Video Intelligence

Language
- Natural Language
- Translate

Conversation
- Speech-to-Text
- Text-to-Speech

AutoML
Custom Models

Sight
- Vision
- Video

Language
- Translate
- Natural Language

Structured Data
- Tables

Platform

Vertex AI (AI Platform)
Development Environment

Built-in Tools for Data Science and Machine Learning
- Datasets
- Data Labeling
- Pre-built Algorithms
- Workbench (Notebook)
- VM Images
- Training
- Predictions
- Continuous evaluation
- Explainability
- Pipelines

On-prem
- Kubeflow

Data Analytics & Management

Ingestion and Processing
- Cloud Pub/Sub
- Cloud Dataflow
- Cloud Dataproc
- Data Fusion
- Dataprep

Storage and Analytics
- Cloud Storage
- BigQuery & BQML
- Cloud Bigtable
- Cloud SQL
- Data Catalog
- Data Studio

Infrastructure
AI Foundation

Accelerators, Compute
- Compute Engine
- Cloud GPU
- Cloud TPU

Orchestration
- Cloud scheduler
- Cloud Composer

Instrumentation
- Cloud Build
- Container Registry

Frameworks
- TensorFlow
- Kubernetes
- PyTorch
- scikit-learn
- Spark

Google Cloud

AI for every level of expertise

ML developer
Intelligent apps

Data analyst
Query and analyze

Data scientist
Models that work

Data engineer
Get clean, useful data

ML engineer
Models in production

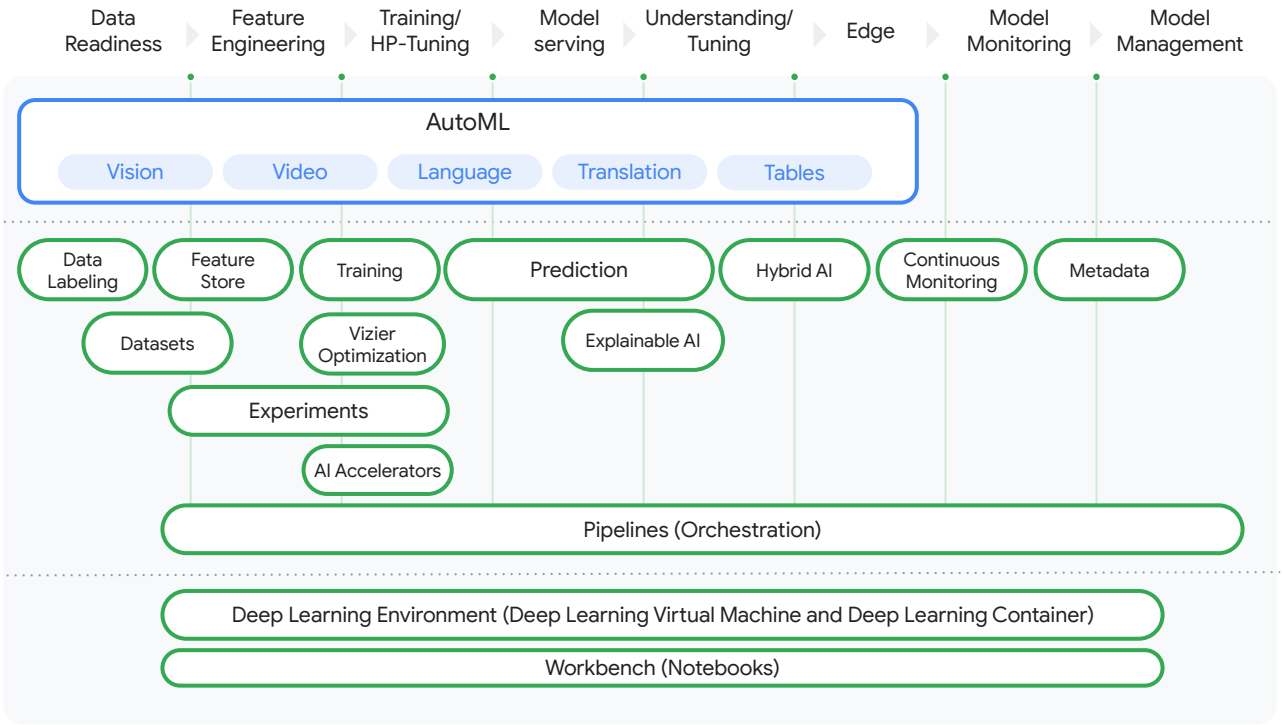
Services and Solutions Ease of Implementation	Fastest way to start using AI today
APIs Pre-trained Models	No training data needed, get started right away
AutoML Custom Models	Easily create custom models (A no-code approach)
AI Platform Development Environment	Lots of control possible, but need Data Science / ML expertise
Data Analytics & Management	
Infrastructure AI Foundation	

Google Cloud

GCP – Vertex AI

Build, deploy, and scale ML models faster, with pre-trained and custom tooling within One unified end-to-end AI platform for everything.

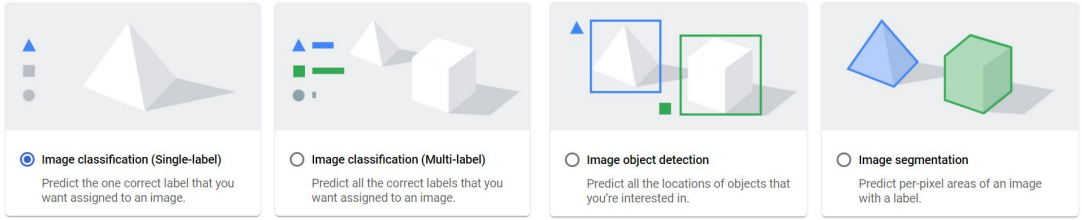
Vertex AI



GCP – Vertex AI (AutoML)

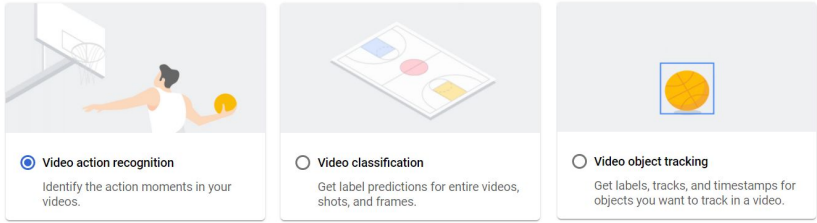
Image

- 圖像分類(單標籤、多標籤)
- 圖像物件偵測
- 圖像分割



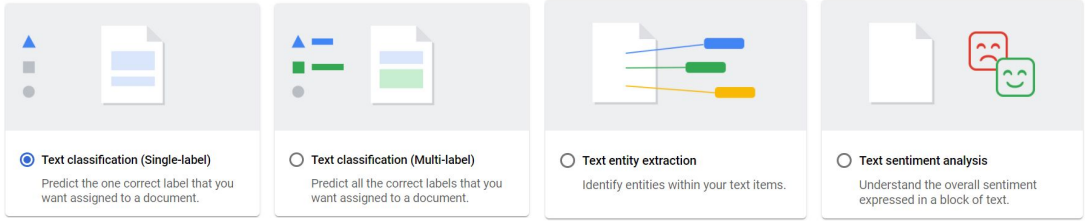
Video

- 影像動作分類
- 影像分類
- 影像物件追蹤



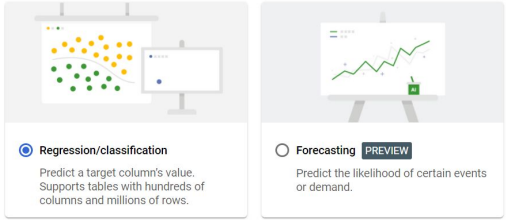
Language

- 文章分類(單標籤、多標籤)
- 文字實體識別
- 文字情緒分析



Tabular

- 回歸 / 分類
- 預測

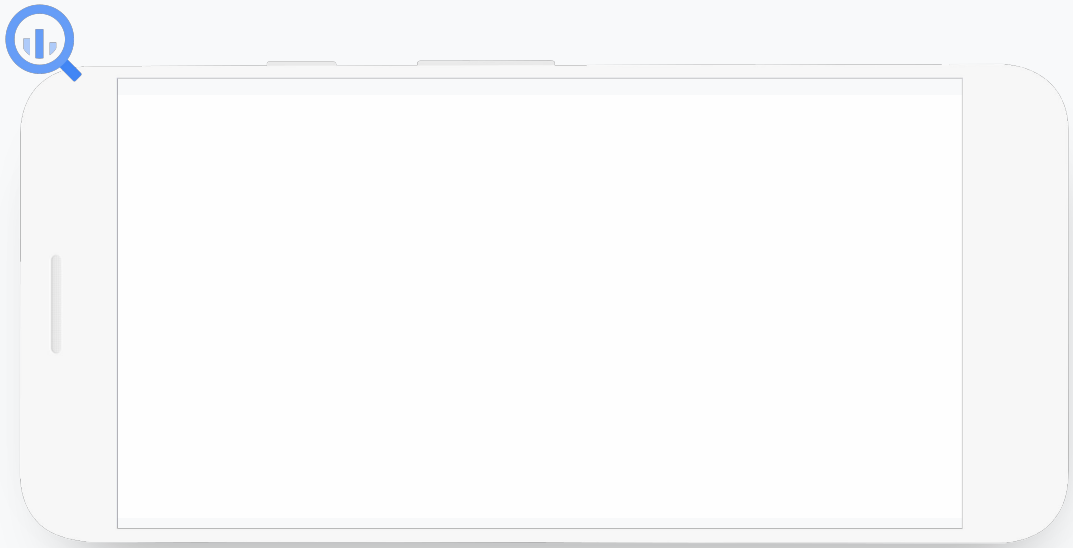


GCP – BigQuery ML

> BigQuery is NOT just analytics

Built-in intelligence
BigQuery ML - build custom models with standard SQL

Google Cloud provides extensive integrated AI and ML services for data analytics including; BigQuery ML, Auto ML, Cloud ML Engine, Tensorflow, and more.



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
4. **Built-in** infrastructure management, security & compliance

The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities

AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
Amazon Rekognition	Amazon Polly	Amazon Transcribe *+Medical*	Amazon Comprehend *+Medical*	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens *For Amazon Connect*

ML SERVICES

Amazon SageMaker

Ground Truth

AWS Marketplace for ML

SageMaker Studio IDE

Built-in algorithms | Notebooks | Experiments | Processing | Model training & tuning | Debugger | Autopilot | Model hosting | Model Monitor

Neo

Augmented AI

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow | mxnet | PYTORCH

GLUON | Keras | DeepGraphLibrary

Deep Learning AMIs & Containers | GPUs & CPUs | Elastic Inference | Inferentia | FPGA

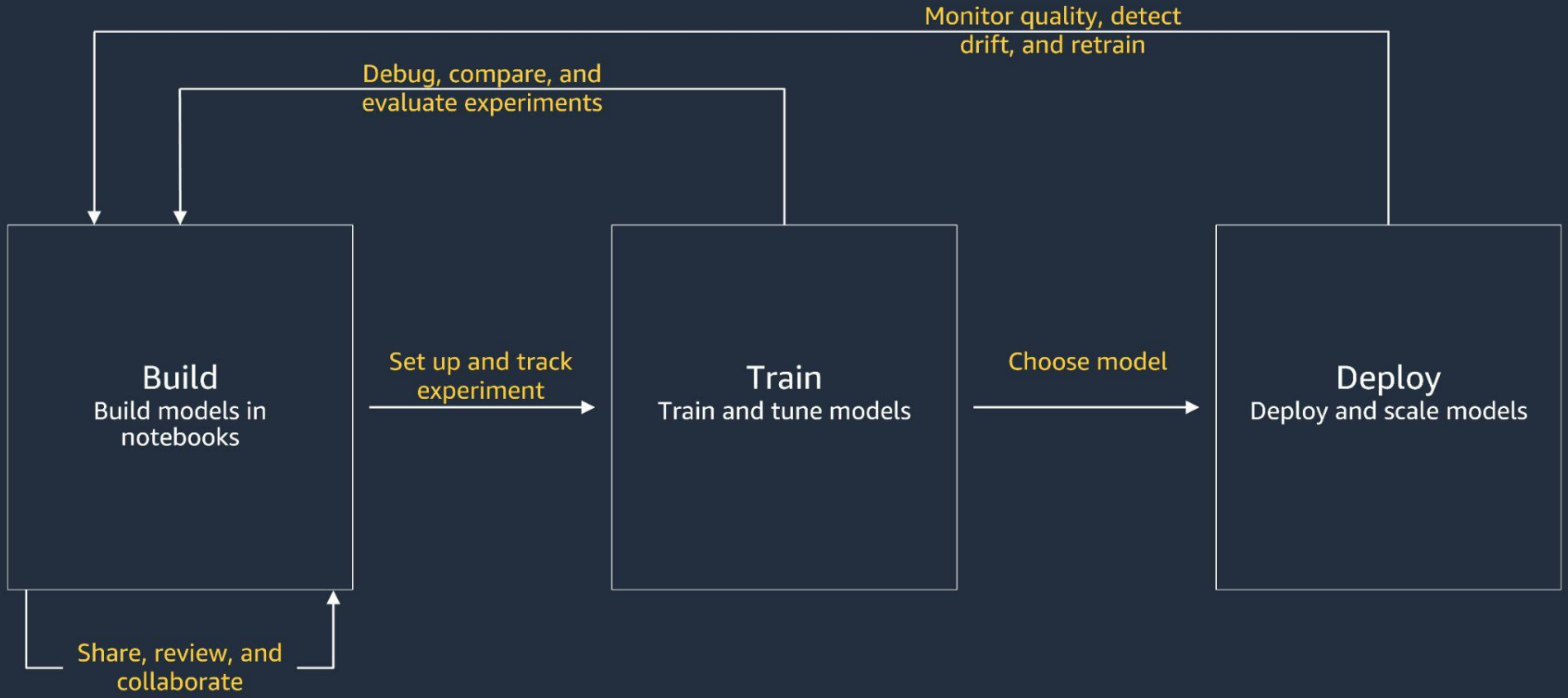
aws

Amazon SageMaker

a fully-managed platform that enables developers and data scientists to quickly and easily **build**, **train**, and **deploy** machine learning

aws machine learning

AWS – SageMaker: Build, train, and deploy models



aws

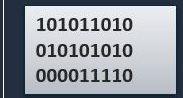
AWS – SageMaker: Build, train, and deploy models

Prepare | **Build** | **Train & tune** | **Deploy & manage**

Web-based IDE for ML

Automatically build and train models

Fully managed data processing jobs and data labeling workflows



Collect and prepare training data

One-click collaborative notebooks and built-in, high performance algorithms and models

Choose or build an ML algorithm

One-click training

Set up and manage environments for training

Debugging and optimization

Train, debug, and tune models

Visually track and compare experiments

Manage training runs

One-click deployment and auto-scaling

Deploy model in production

Automatically spot concept drift

Monitor models

Add human review of predictions

Validate predictions

Fully managed with auto-scaling for 75% less

Scale & manage the production environment

aws

AWS – SageMaker Overview

aws

AutoML with Amazon SageMaker Autopilot

SageMaker Autopilot covers all steps

- *Problem identification*: looking at the data set, what class of problem are we trying to solve?
- *Algorithm selection*: which algorithm is best suited to solve the problem?
- *Data preprocessing*: how should data be prepared for best results?
- *Hyperparameter tuning*: what is the optimal set of training parameters?

Supported algorithms at launch:

Linear Learner, XGBoost

Amazon SageMaker Autopilot

Core Features

Automatic model creation for tabular data with full visibility and control

Quick to start

Provide your data in a tabular form and specify target prediction

Automatic model creation

Get ML models with feature engineering & model tuning automatically done

Visibility and control

Get notebooks for your models with source code

Recommendations and optimization

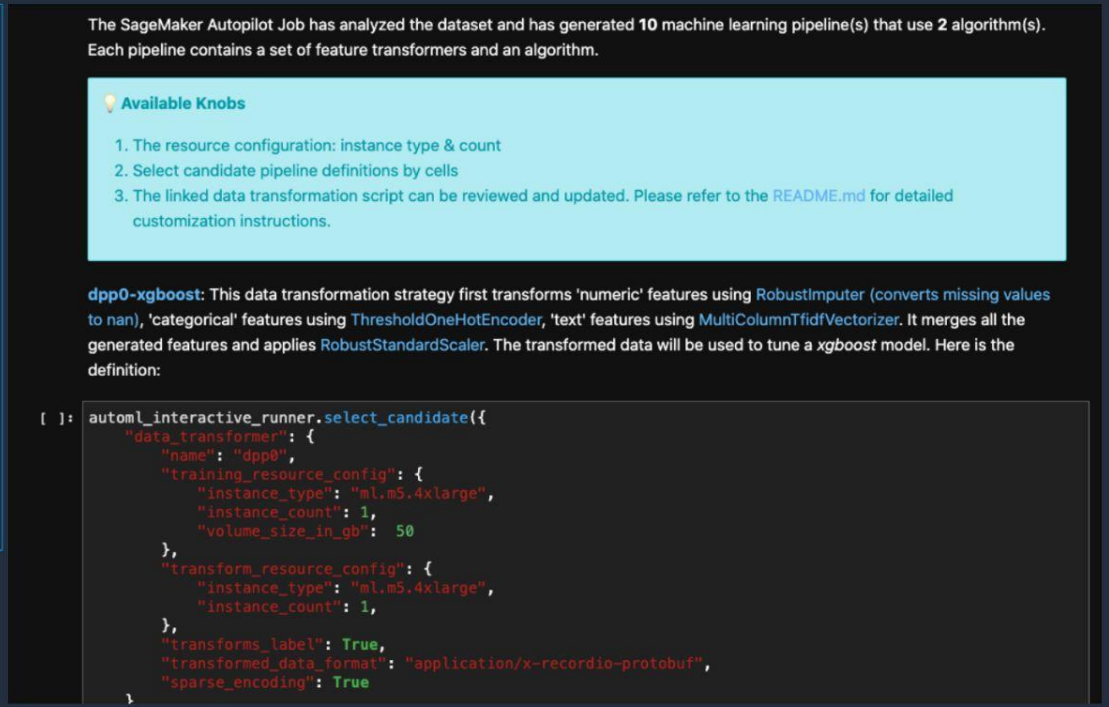
Get a leaderboard & continue to improve your model

aws

Autopilot for model candidates

Fully runnable Model Candidate Notebook:

- Data transformers
- Featurization techniques applied
- Override points:
  - Algorithms considered
  - Evaluation metric
  - Hyper-parameter ranges
  - Model search strategy
  - Instances used



The SageMaker Autopilot Job has analyzed the dataset and has generated **10** machine learning pipeline(s) that use **2** algorithm(s). Each pipeline contains a set of feature transformers and an algorithm.

Available Knobs

1. The resource configuration: instance type & count
2. Select candidate pipeline definitions by cells
3. The linked data transformation script can be reviewed and updated. Please refer to the README.md for detailed customization instructions.

dpp0-xgboost: This data transformation strategy first transforms 'numeric' features using RobustImputer (converts missing values to nan), 'categorical' features using ThresholdOneHotEncoder, 'text' features using MultiColumnTfidfVectorizer. It merges all the generated features and applies RobustStandardScaler. The transformed data will be used to tune a *xgboost* model. Here is the definition:

```
[ ]: automl_interactive_runner.select_candidate({
    "data_transformer": {
        "name": "dpp0",
        "training_resource_config": {
            "instance_type": "ml.m5.4xlarge",
            "instance_count": 1,
            "volume_size_in_gb":  50
        },
        "transform_resource_config": {
            "instance_type": "ml.m5.4xlarge",
            "instance_count": 1,
        },
        "transforms_label": True,
        "transformed_data_format": "application/x-recordio-protobuf",
        "sparse_encoding": True
    }
```


aws

AWS – SageMaker: Build-in algorithm

Amazon SageMaker
has built-in algorithms
or bring your own

Computer vision
Image classification | Object detection | Semantic segmentation

Topic modeling
LDA | NTM

Classification
Linear Learner | XGBoost | KNN

Recommendation
Factorization machines

Forecasting
DeepAR

Working with text
BlazingText | Supervised | Unsupervised

Regression
Linear Learner | XGBoost | KNN

Clustering
KMeans

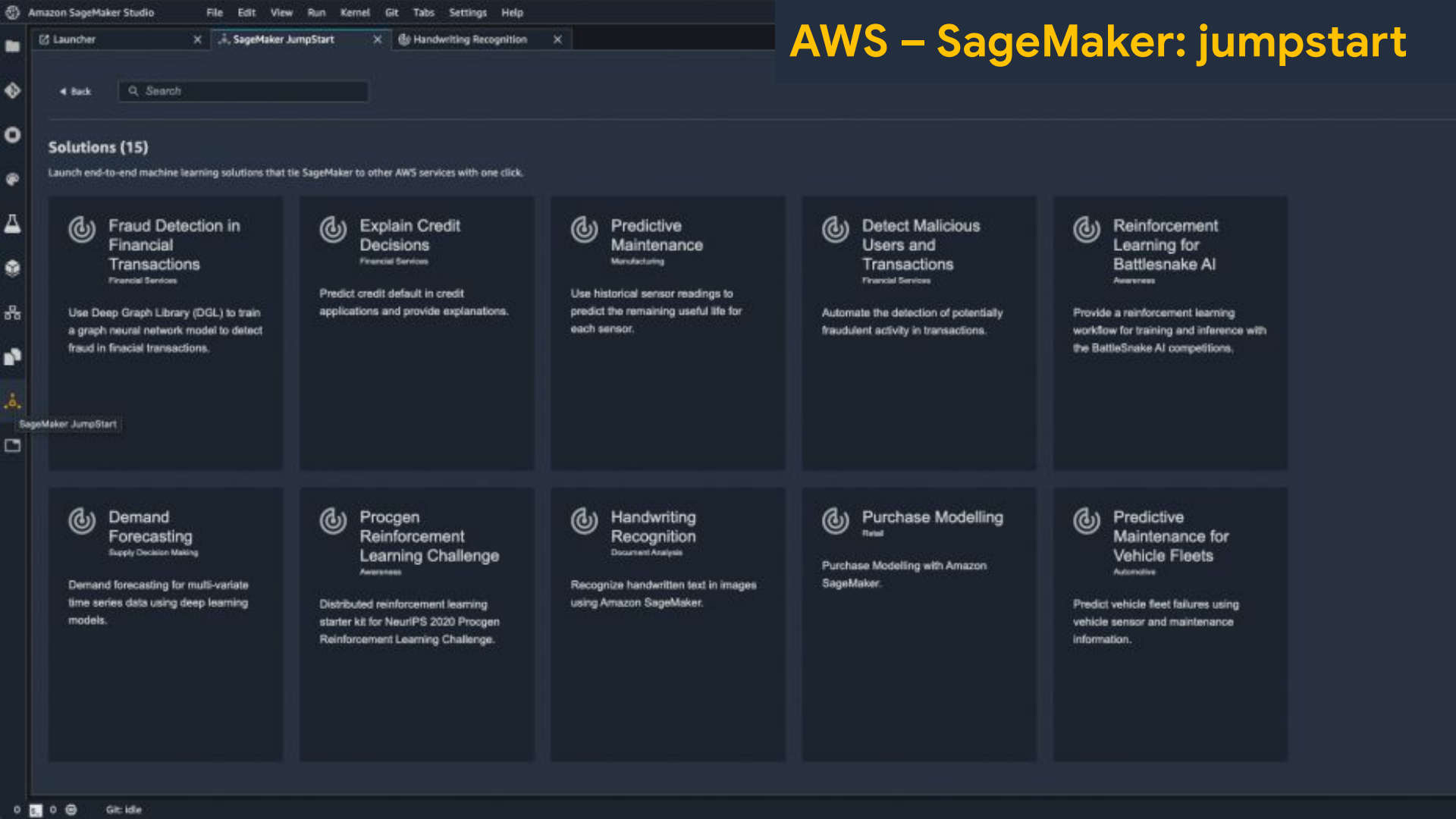
Sequence translation
Seq2Seq

Anomaly detection
Random cut forests | IP Insights

Feature reduction
PCA

aws

AWS – SageMaker: jumpstart



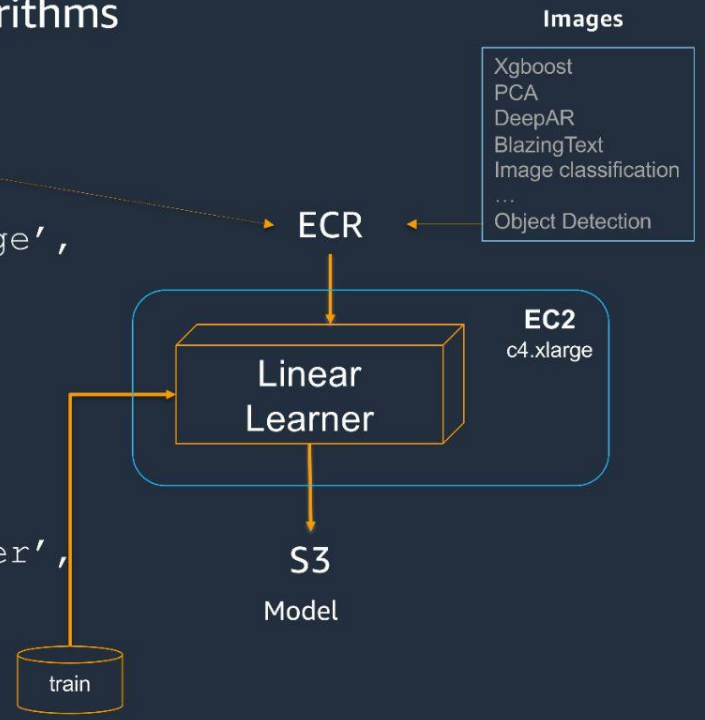
Amazon SageMaker | Training

Use built-in algorithms

```
linear = Estimator('linear-learner',
                   train_instance_count=1,
                   train_instance_type='ml.c4.xlarge',
                   output_path=output_location,
                   sagemaker_session=sess)
```

```
linear.set_hyperparameters(
                   feature_dim=784,
                   predictor_type='binary_classifier',
                   mini_batch_size=200)
```

```
linear.fit({'train': s3_train_data})
```



aws

Every model run on a SageMaker training job has its own ephemeral cluster.

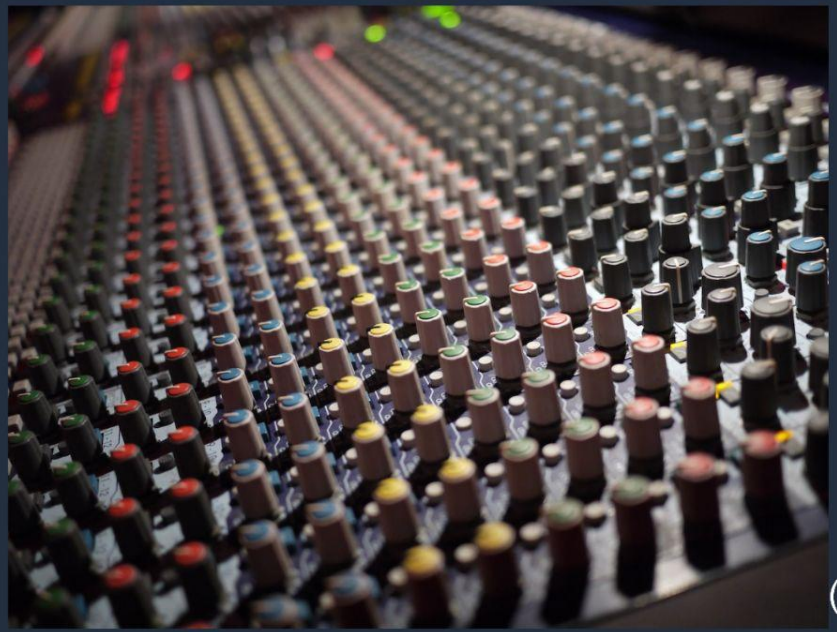
That means you have a dedicated EC2 instance alive for the number of seconds your model needs to train.

This cluster comes down immediately after the model finished training.

aws

Amazon SageMaker Automatic Model Tuning

Hyperparameter Optimizer



Decision Trees	Neural Networks
Tree depth	Number of layers
Max leaf nodes	Hidden layer width
Gamma	Learning rate
Eta	Embedding
Lambda	dimensions
Alpha	Dropout
…	…

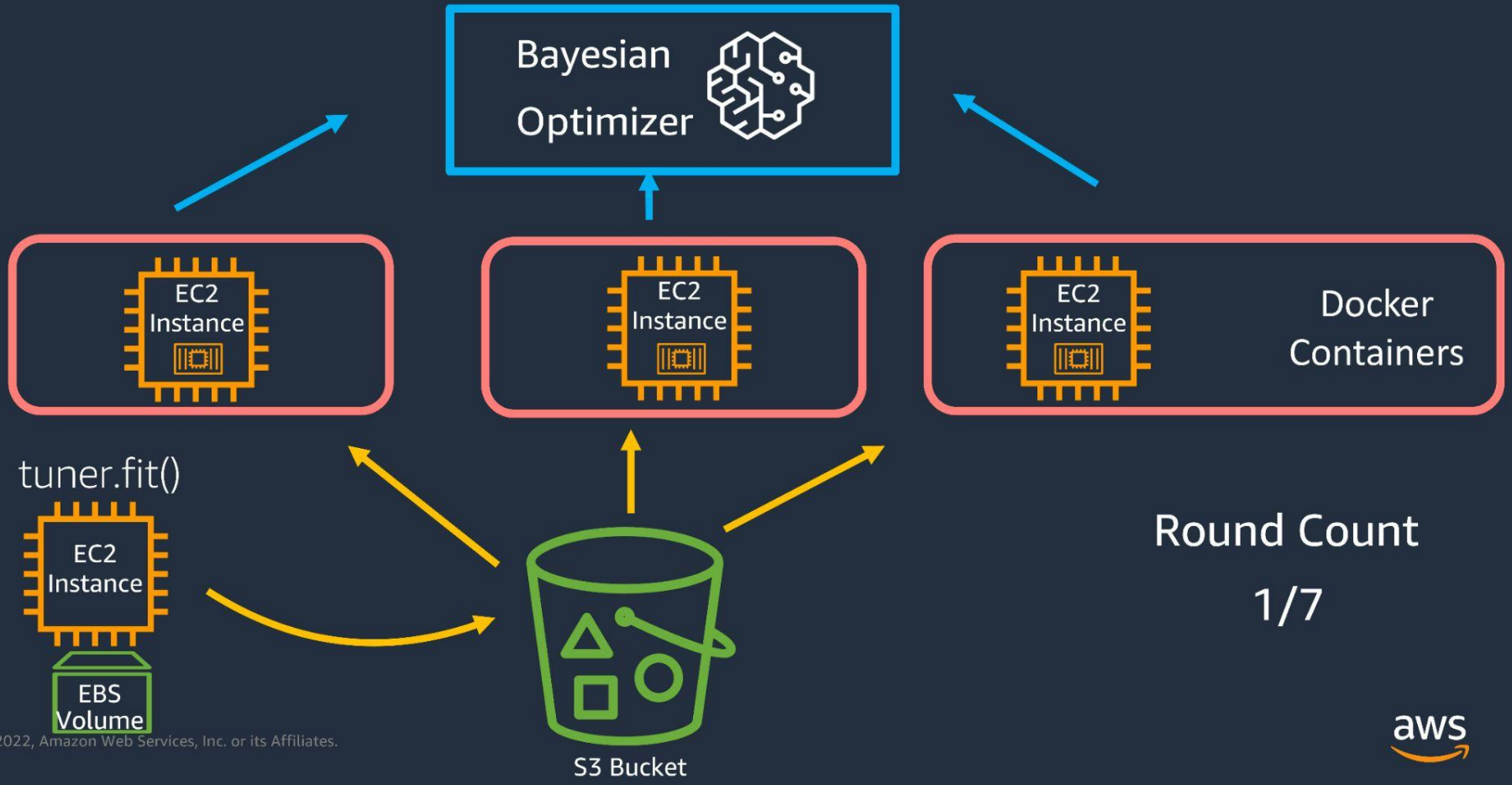
…

"Hyperparameters"

(algorithm parameters that significantly affect model quality)

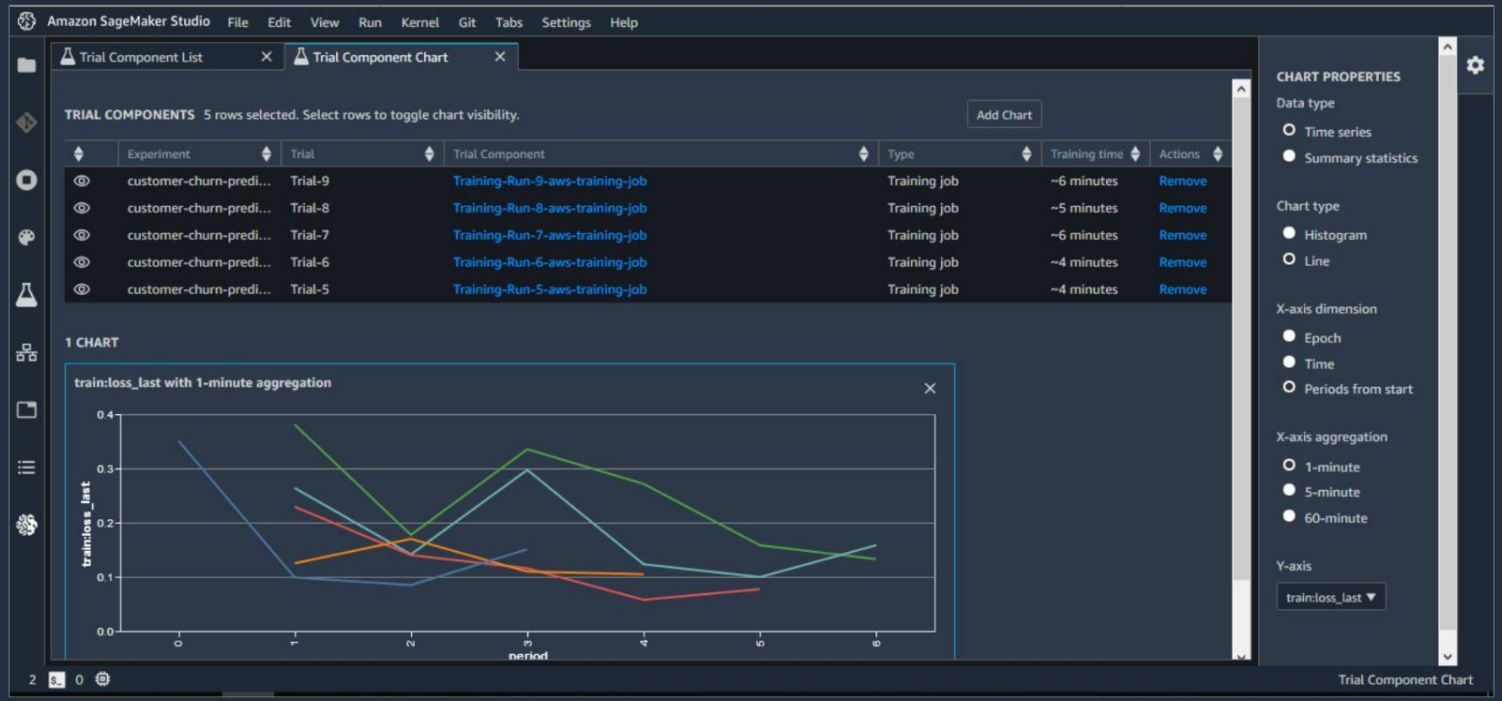
aws

Hyperparameter Tuning Jobs



aws

Use Amazon SageMaker Experiments to track and manage thousands of experiments



Deployment options

Model in Amazon S3

Amazon SageMaker Real-time endpoint

1 line of code
Vanilla HTTPS
Post data, get a prediction
Any tool, any language
Auto Scaling available

Amazon SageMaker batch transform

1 line of code
Predict data stored in S3
Read results from S3

Amazon container services (ECS, EKS, Fargate)

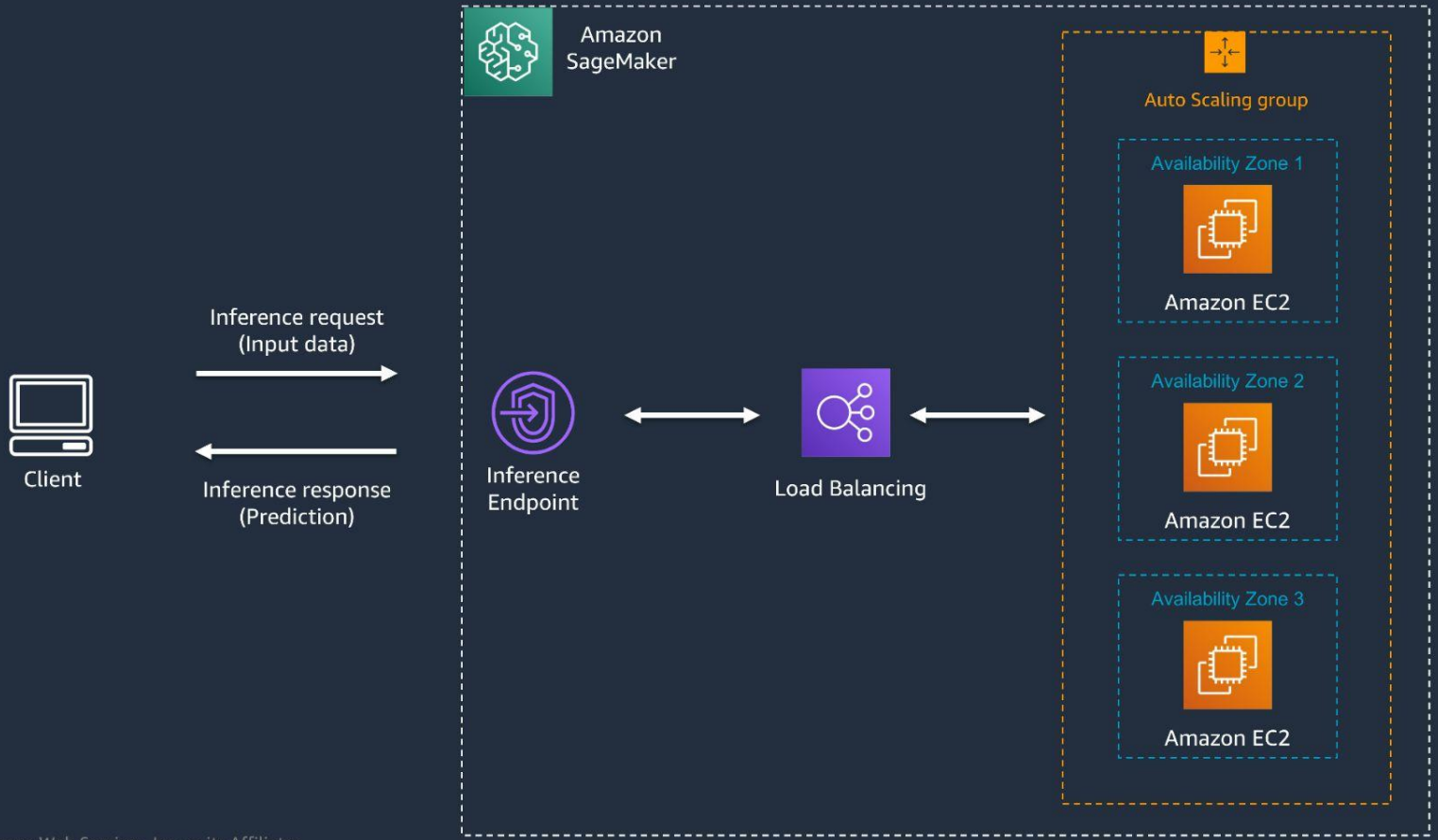
Use AWS Deep Learning containers
Use your own container

Anywhere you like

Grab the model in S3 and run

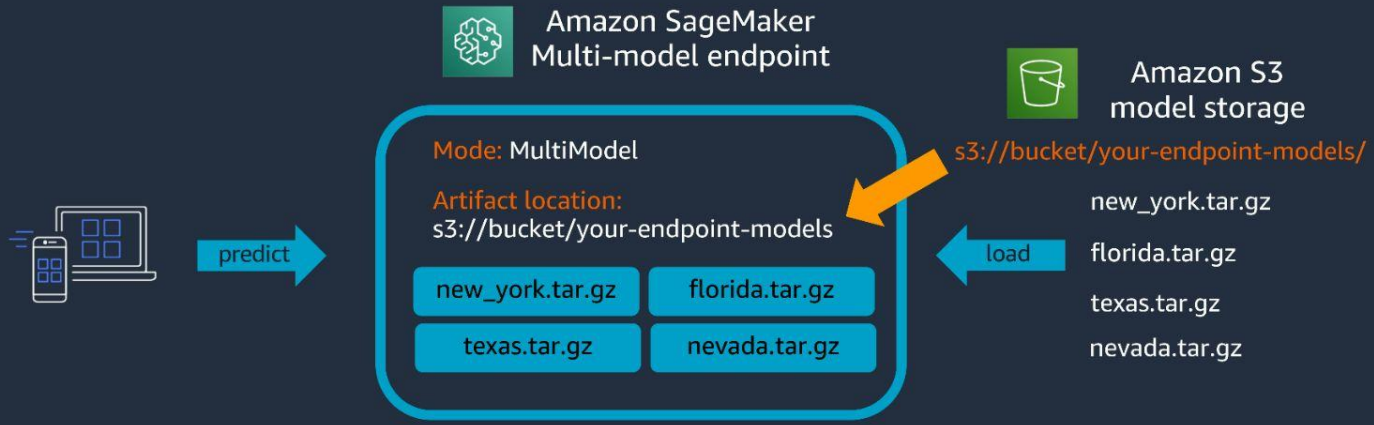
aws

Amazon SageMaker Real-time endpoint



aws

Multi-Model Endpoints



Key Capabilities

- Deploy tens to tens of thousands of models; target model for each inference request
- Two modes for model caching: Caching Enabled, Caching Disabled
- Support for built in algorithms, SageMaker frameworks, and custom models
- Support for inference pipelines and condition keys to restrict access to models

aws

Multi-model endpoints

Significant savings for large-scale deployments

10 separate endpoints
$3,430/month

1 multi-model endpoint
$343/month

Sample scenario: ml.c5.xlarge, $0.238/hour, 2 instances running 24/7

aws

SageMaker Serverless Inference (preview)

Deploying ML models using SageMaker Serverless Inference (Preview)

by Ram Vegiraju, Michael Pham, Rishabh Ray Chaudhury, and Shelbee Eigenbrode | on 05 JAN 2022 | in Amazon Machine Learning, Amazon SageMaker, Artificial Intelligence | Permalink | Comments | Share

Amazon SageMaker Serverless Inference (Preview) was recently announced at re:Invent 2021 as a new model hosting feature that lets customers serve model predictions without having to explicitly provision compute instances or configure scaling policies to handle traffic variations. Serverless Inference is a new deployment capability that complements SageMaker's existing options for deployment that include: SageMaker Real-Time Inference for workloads with low latency requirements in the order of milliseconds, SageMaker Batch Transform to run predictions on batches of data, and SageMaker Asynchronous Inference for inferences with large payload sizes or requiring long processing times.

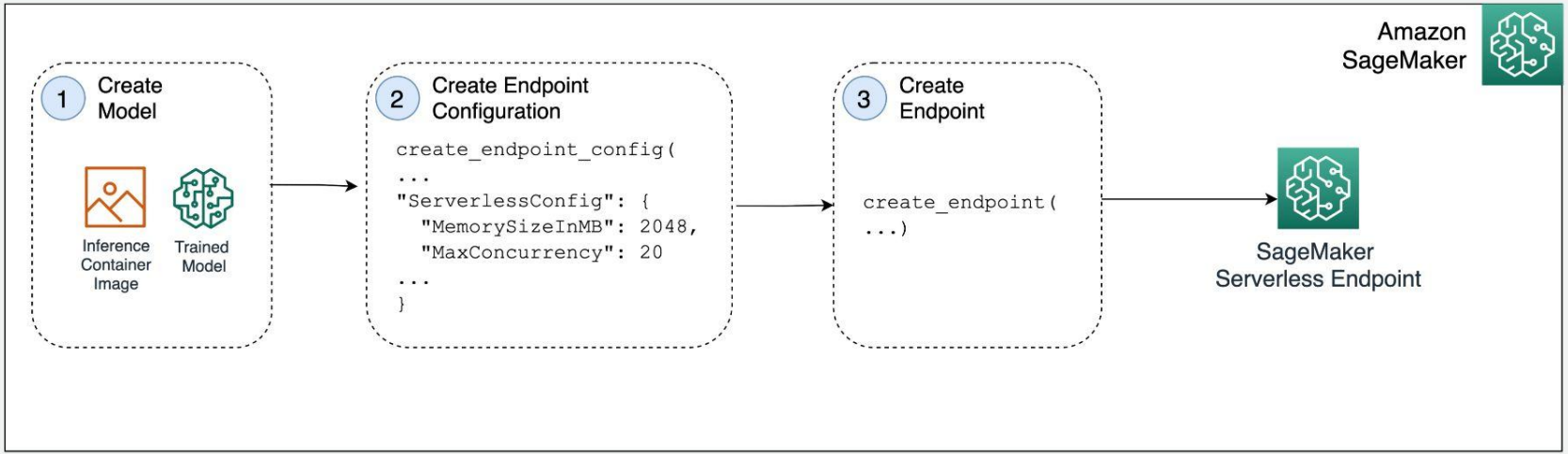
Serverless Inference means that you don't need to configure and manage the underlying infrastructure hosting your models. When you host your model on a Serverless Inference endpoint, simply select the memory and max concurrent invocations. Then, SageMaker will automatically provision, scale, and terminate compute capacity based on the inference request volume. SageMaker Serverless Inference also means that you only pay for the duration of running the inference code and the amount of data processed, not for idle time. Moreover, you can scale to zero to optimize your inference costs.

Source: https://aws.amazon.com/blogs/machine-learning/deploying-ml-models-using-sagemaker-serverless-inference-preview/

aws

How it works

Creating a Serverless Endpoint



Our mission at AWS

Put machine learning in the hands of every developer

aws

How to Build a AI/ML Kingdom

- Use what you need

05

Google Cloud

How to Build a AI/ML Kingdom

每個雲端都有擁護者。

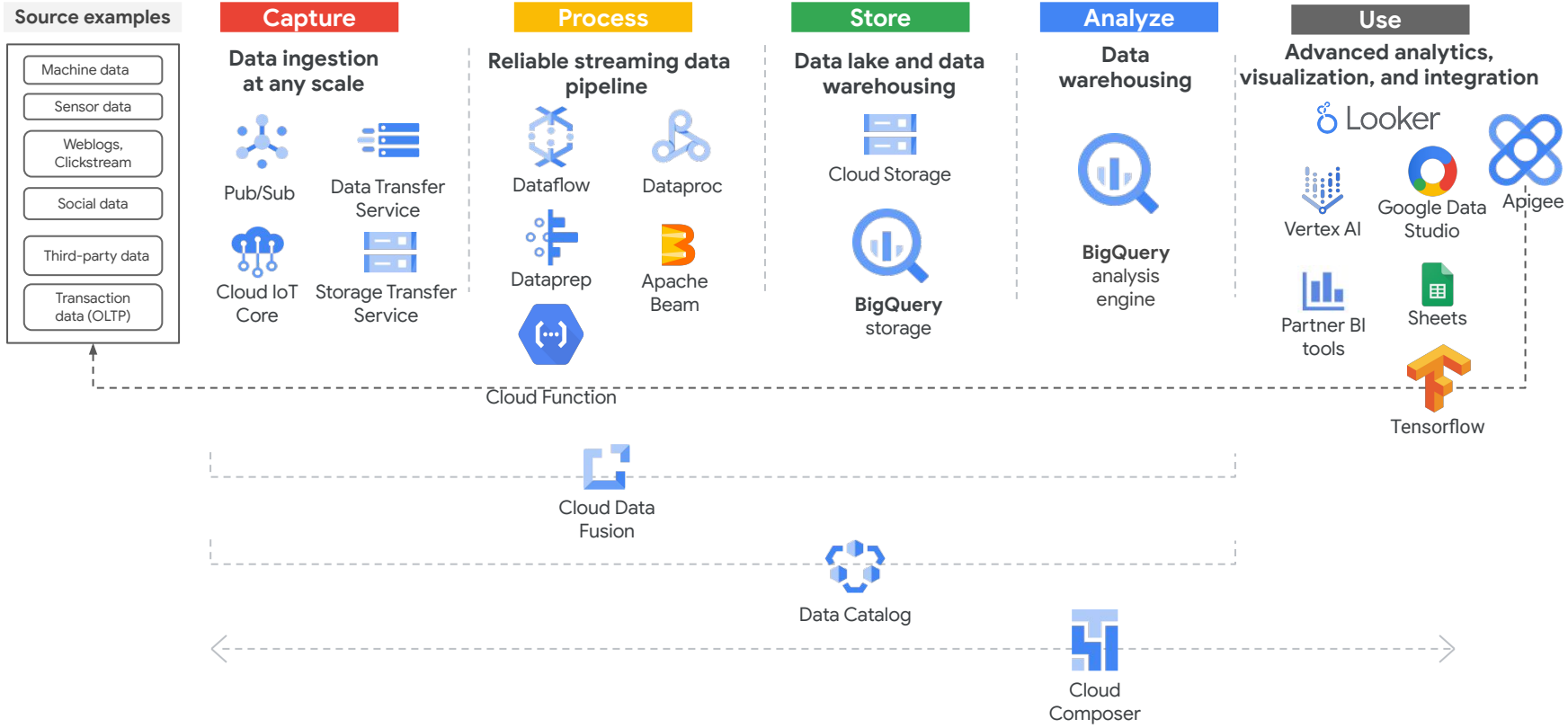
個人的建議

- Coding 開發狂熱者 → 用用看 AWS - SageMaker (真的節省很多時間)
- 落實 no/low code → 試試 GCP - Vertex AI (確實做到 no/low code 的模型訓練到部署)

> **沒有最好的路，只有最適合自己的路**

About Data Pipeline

Part of Google Cloud comprehensive data analytics portfolio



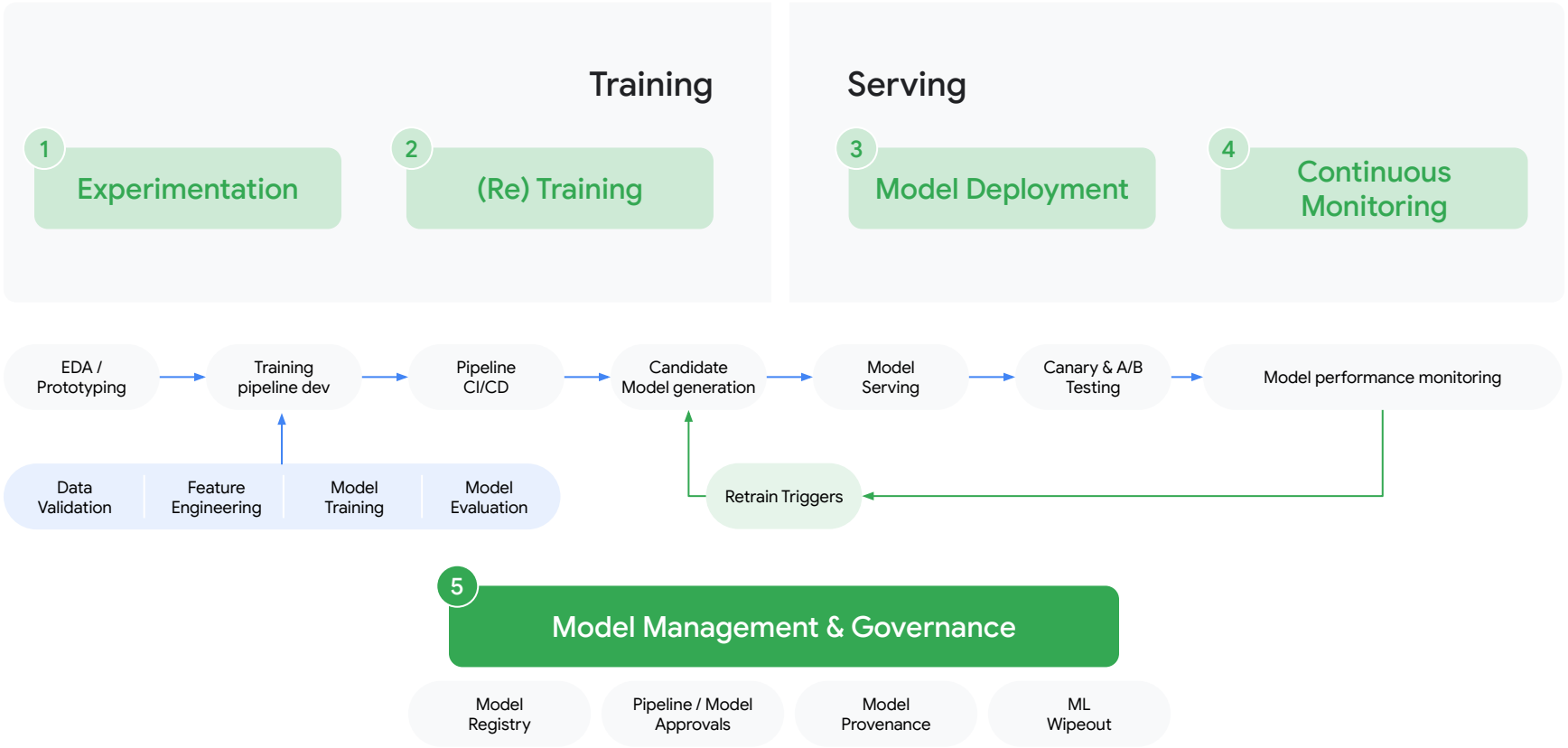
What's Next

- MLOps
- Data Governance

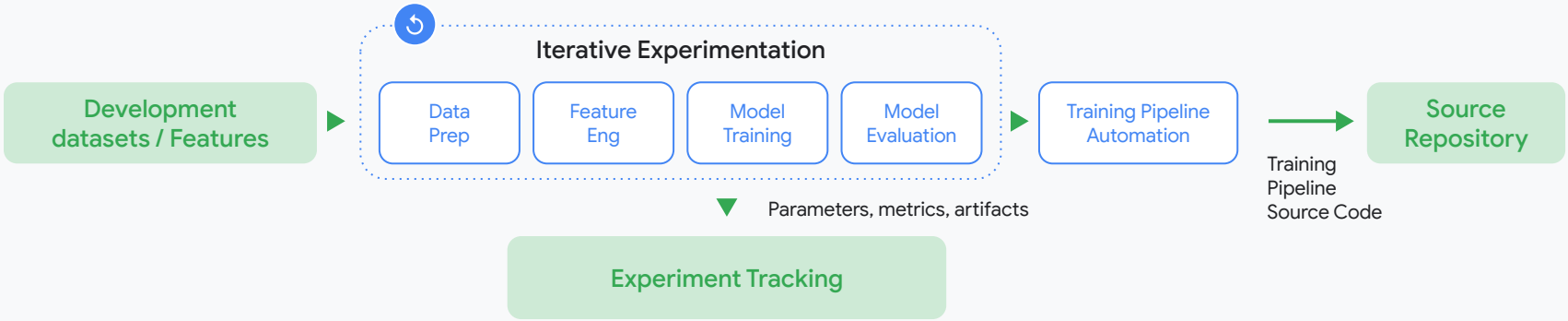
06

Google Cloud

High Level MLOps Framework that aligns with Vertex AI



Experimentation



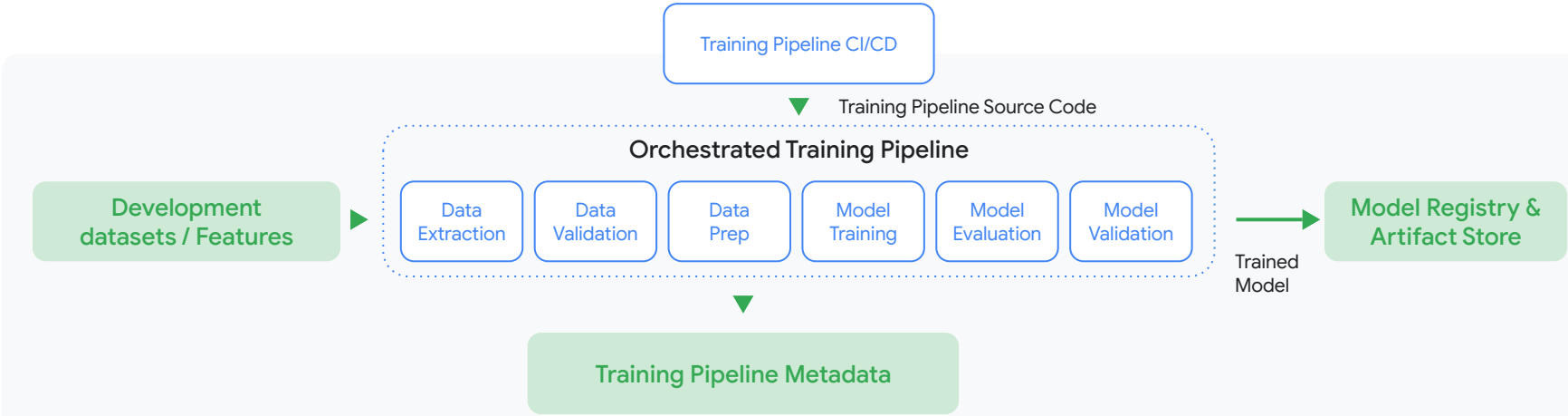
Tools and services

- Notebooks
- Vertex Training
- AutoML in Vertex AI
- BigQuery ML
- Vertex Tensorboard
- Vertex Pipelines
- Vertex Feature Store
- What-if Tool

Key artifacts

- Development datasets
- Features
- Experiments
- Parameters, metrics

Training



Tools and services

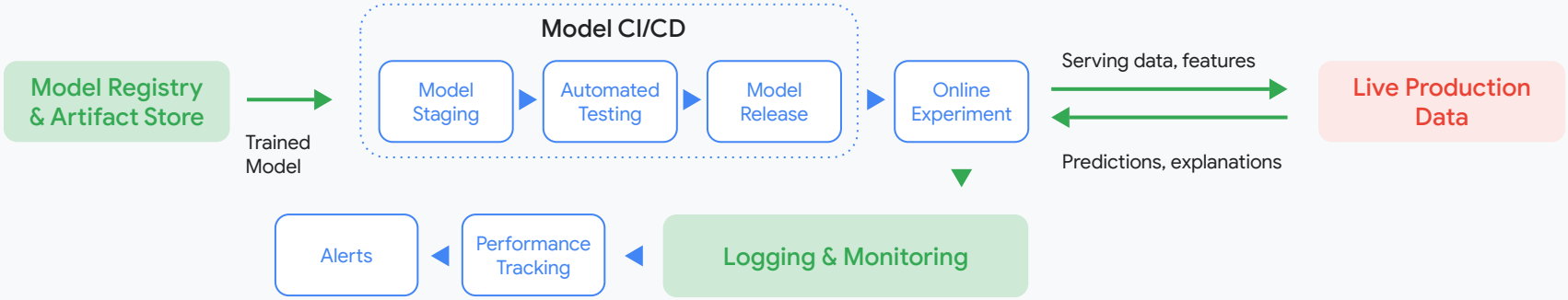
- Vertex Pipelines
- Vertex Training
- Cloud Build
- Artifact Store
- Vertex Explainable AI
- Vertex ML Metadata
- Vertex Feature Store

Key artifacts

- Training datasets
- Features
- Pipeline source code & containers
- Pipeline Metadata
- Models

Model deployment with monitoring

3 & 4



Tools and services

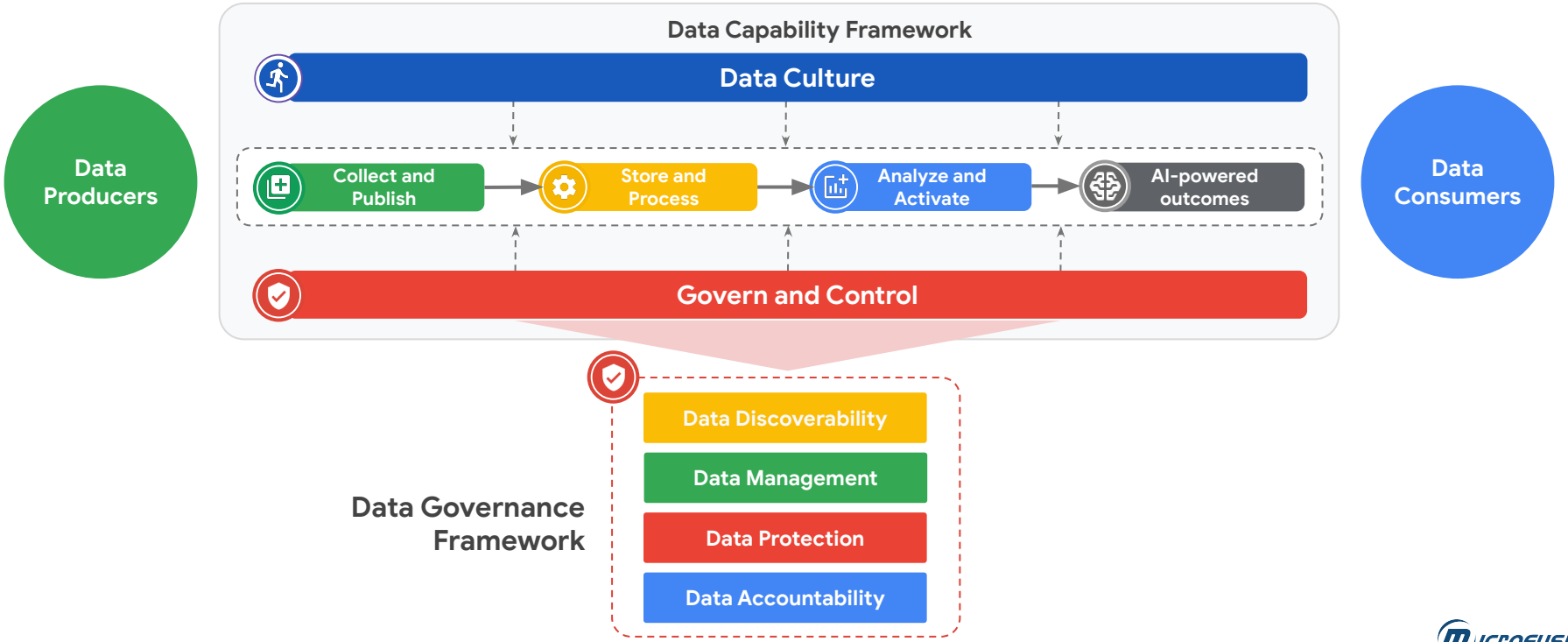
- Vertex Prediction
- Vertex Pipelines
- Vertex Explainable AI
- Artifact Store
- Vertex Model Monitoring
- Vertex ML Metadata
- Vertex Feature Store

Key artifacts

- Deployed Models
- Production Serving Data
- Features
- Online Predictions

Data Governance

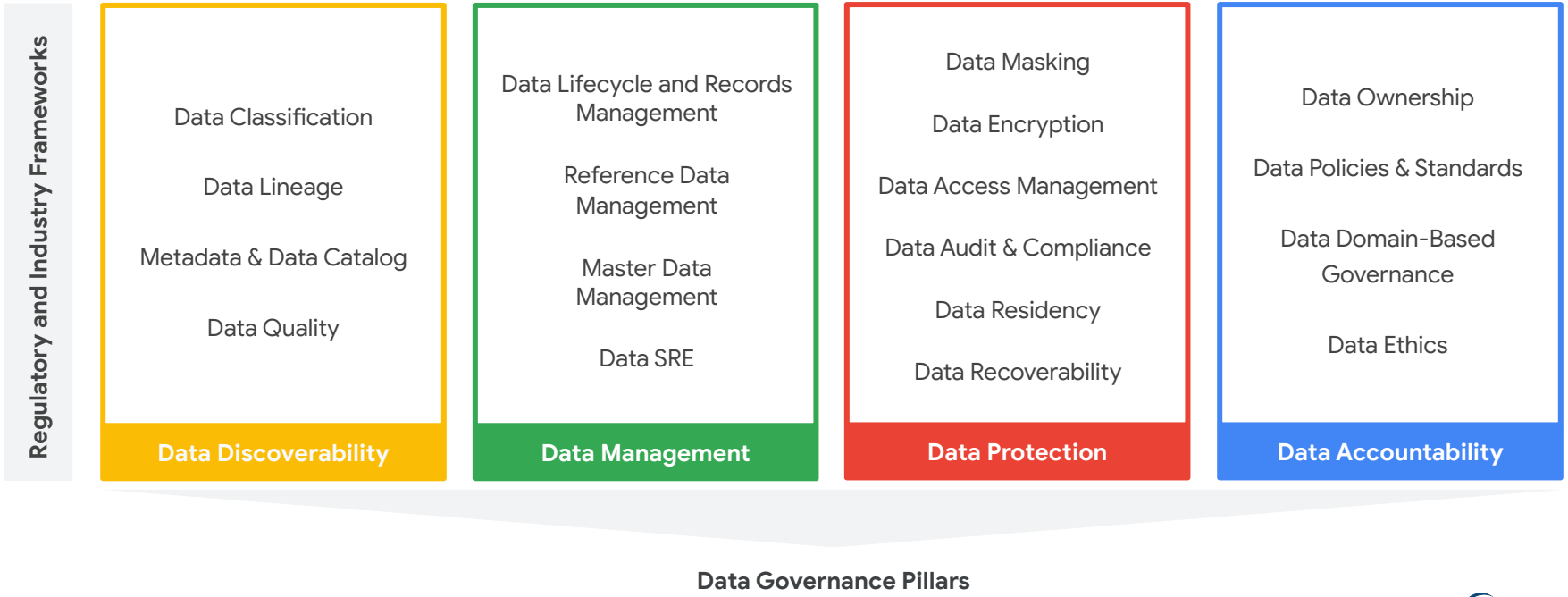
數據驅動是從數據的生產方到數據的消費者。
如果有一套數據治理的框架將整個旅程，結構化、系統化，那就能讓 資料應用加倍放大



Data Governance

At its most basic level, data governance is the practice of enhancing an organization's data such that it is **discoverable, understood, protected, and trusted.**

Data Governance Framework



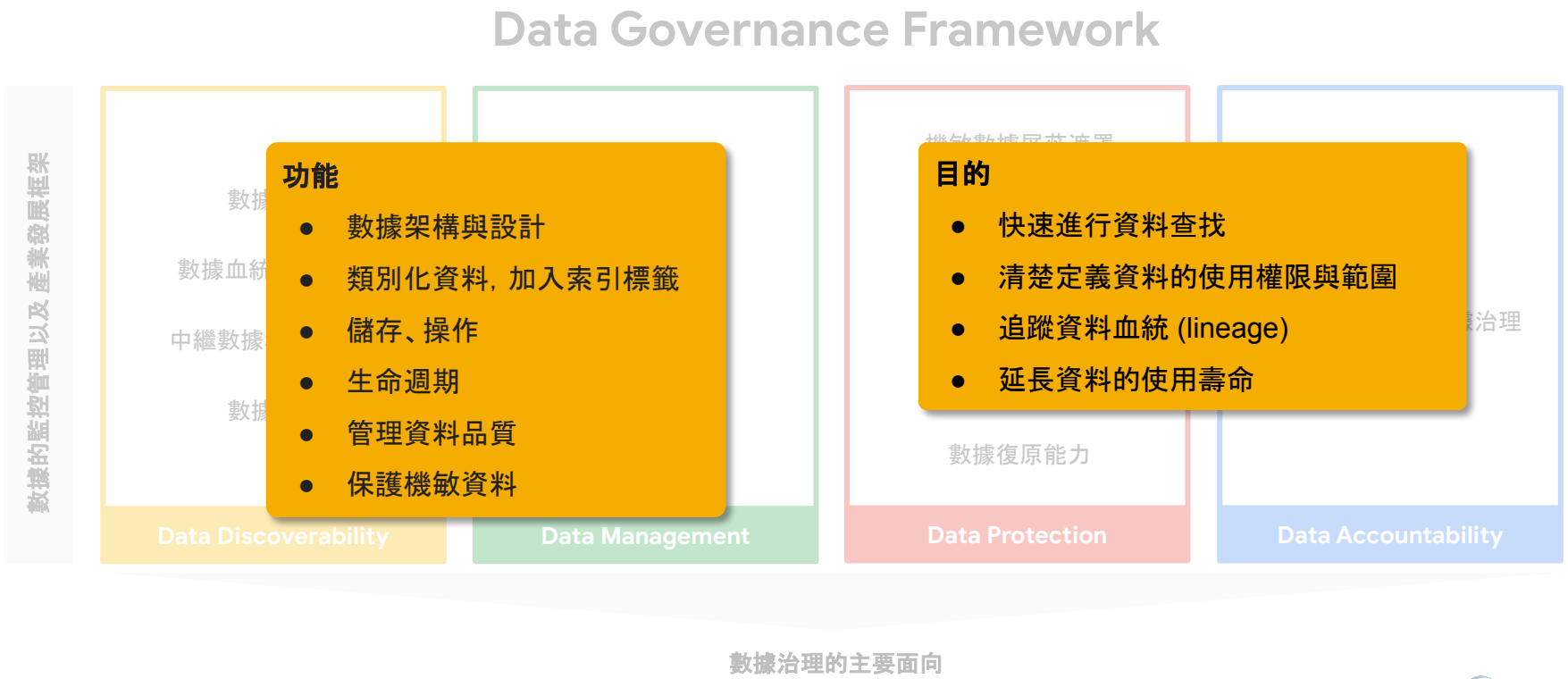
Data Governance

「數據治理」是增強整個組織在數據驅動上的實踐，使其可被 **發現、理解、管理、保護和信任** 。

Data Governance Framework

資料膨脹與治理

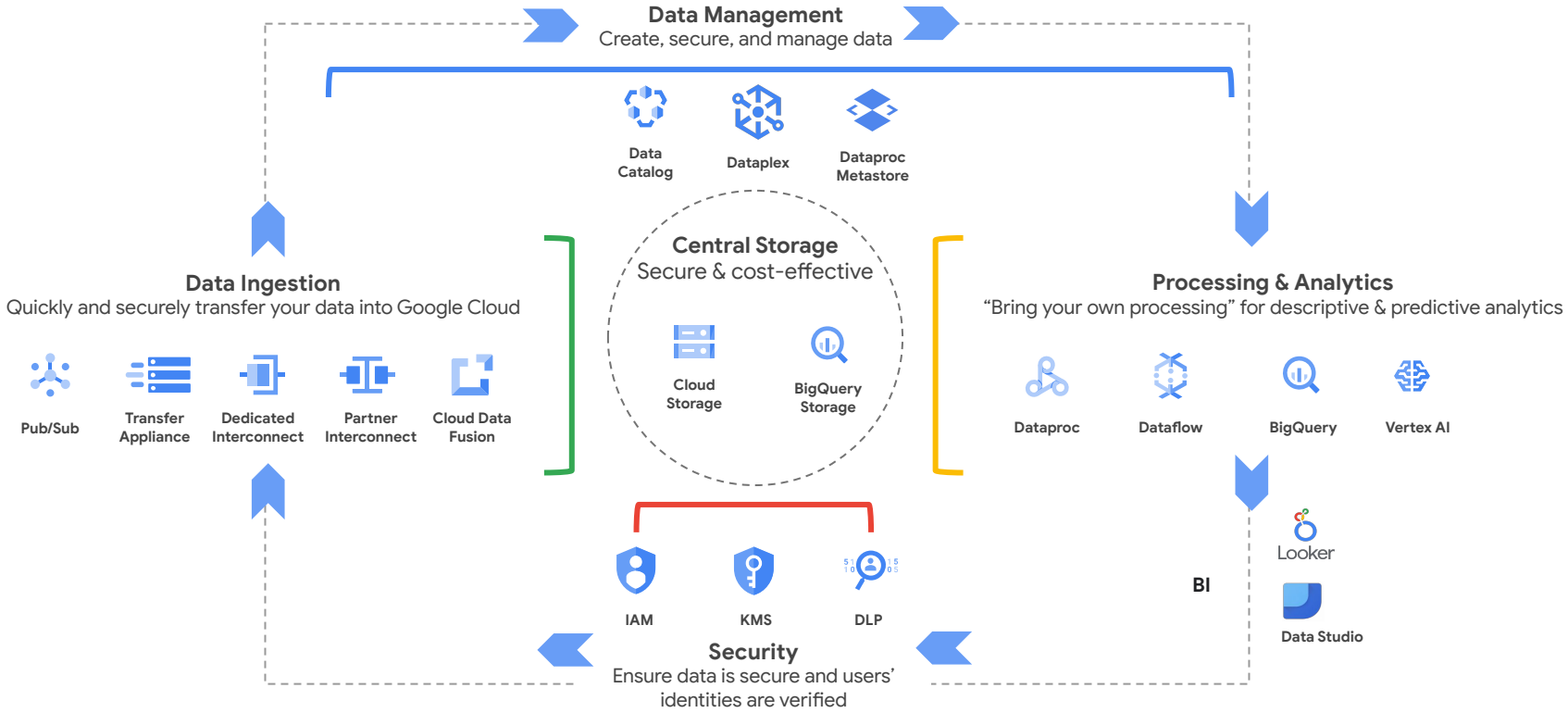
「數據治理」是增強整個組織在數據驅動上的實踐，使其可被 **發現、理解、管理、保護和信任** 。



Data governance in Google Cloud

你終究要做數據治理，何不一開始就用?

How Google Cloud services can help you govern your data



工商一下啦

07

Google Cloud

本公司有在招人XDD

- 雲端資安工程師｜博弘雲端科技股份有限公司｜台北市中山區－ 104人力銀行
- GCP助理雲端架構師｜博弘雲端科技股份有限公司｜台北市中山區
- GCP雲端架構師｜博弘雲端科技股份有限公司｜台北市中山區
- 資料分析師｜博弘雲端科技股份有限公司｜台北市中山區

Any Questions

linktr.ee/youjun

履歷、專案
社群分享

linktr.ee/youjun_talk

今日簡報
歡迎下載