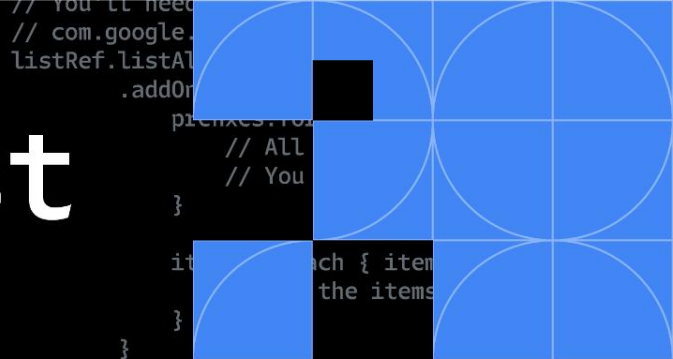




# devfest




## The Progress of Using Google Vertex AI

 Google Developer Groups  
Yucheng Wang ML/IoT GDE

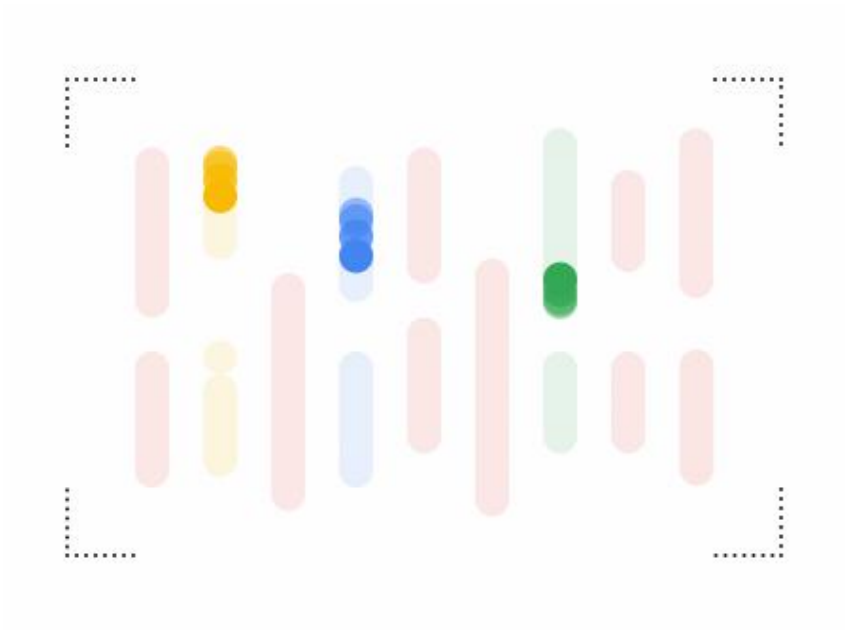


# Agenda

- 
- 1 Vertex AI Overview
  - 2 Unified Data & AI Platform
  - 3 End-to-End MLOps
  - 4 Open and Scalable AI Infrastructure

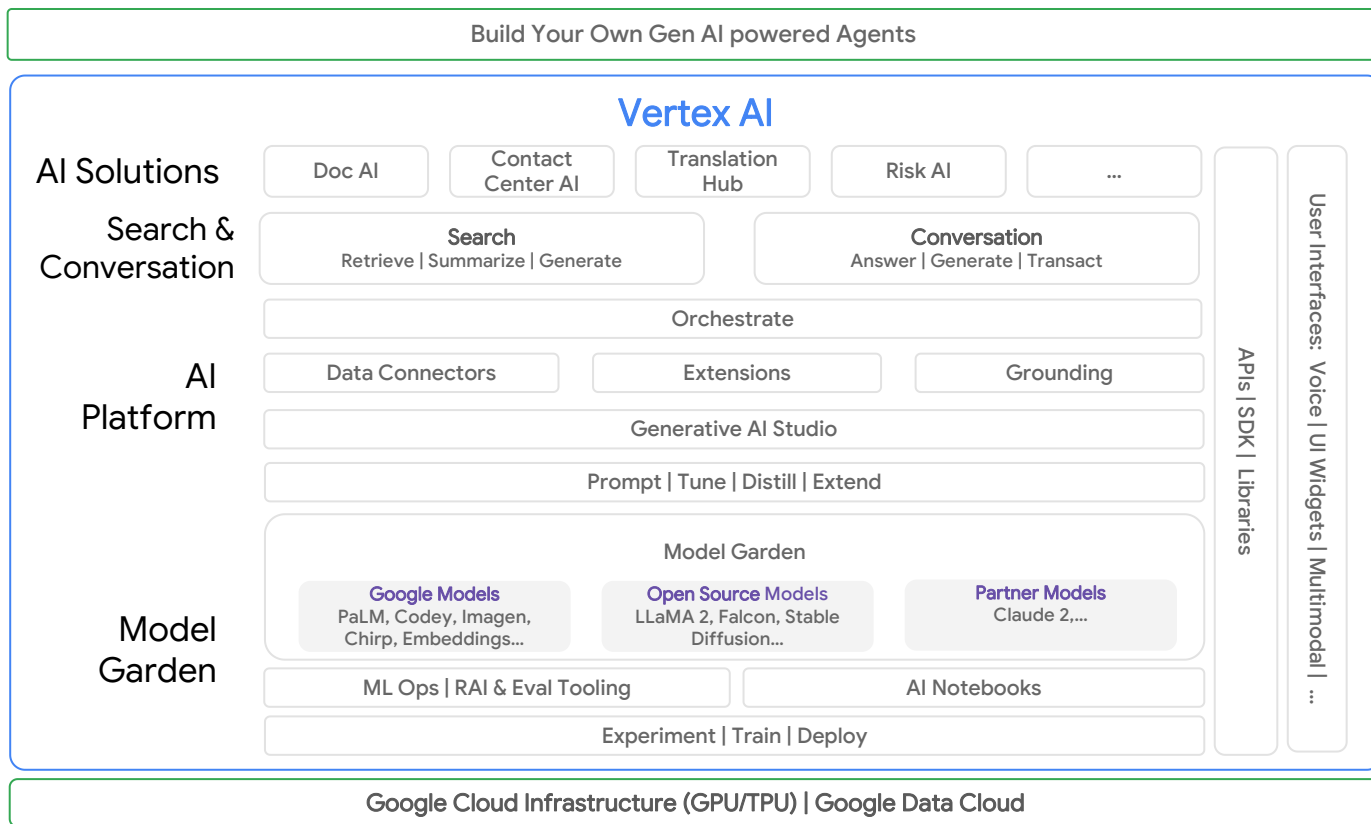
# Vertex AI Overview

Unified development and deployment platform for  
data science, machine learning, and generative AI



# Google Cloud Generative AI

Empower enterprises to innovate faster with **enterprise-ready** generative AI



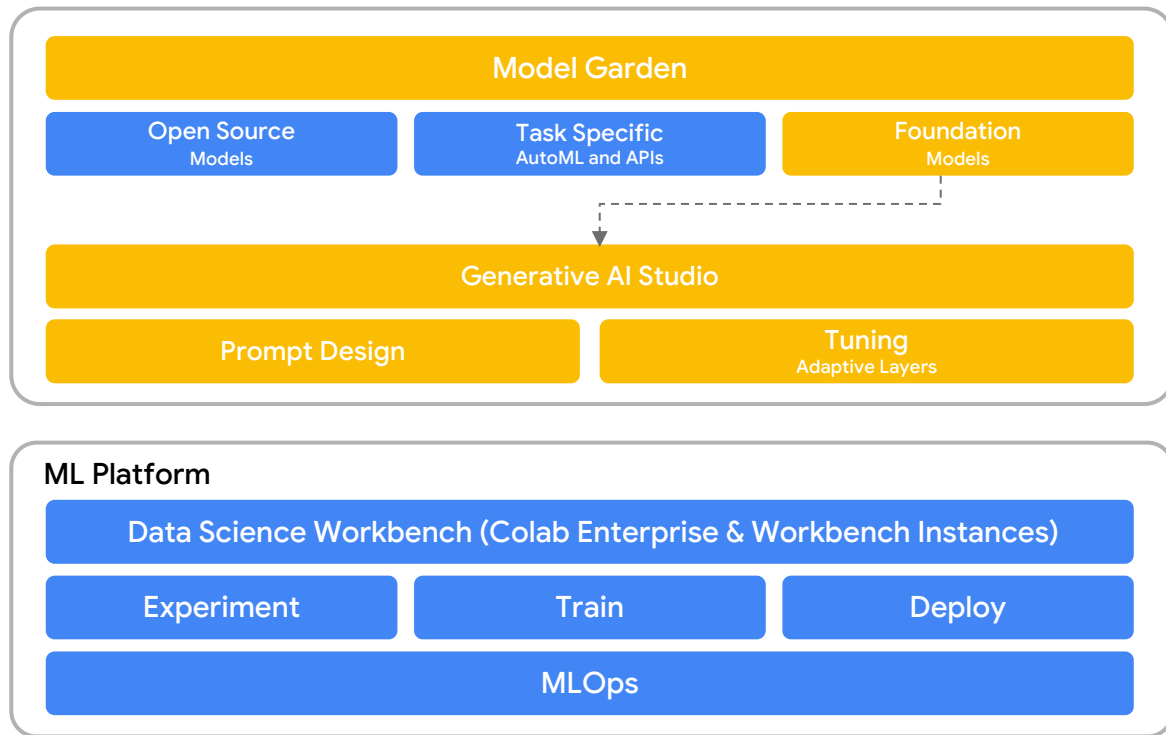
Broad Ecosystem of Partners

Enterprise Readiness

- Data Governance
- Data Privacy
- Responsible AI
- Security & Compliance

Google Cloud

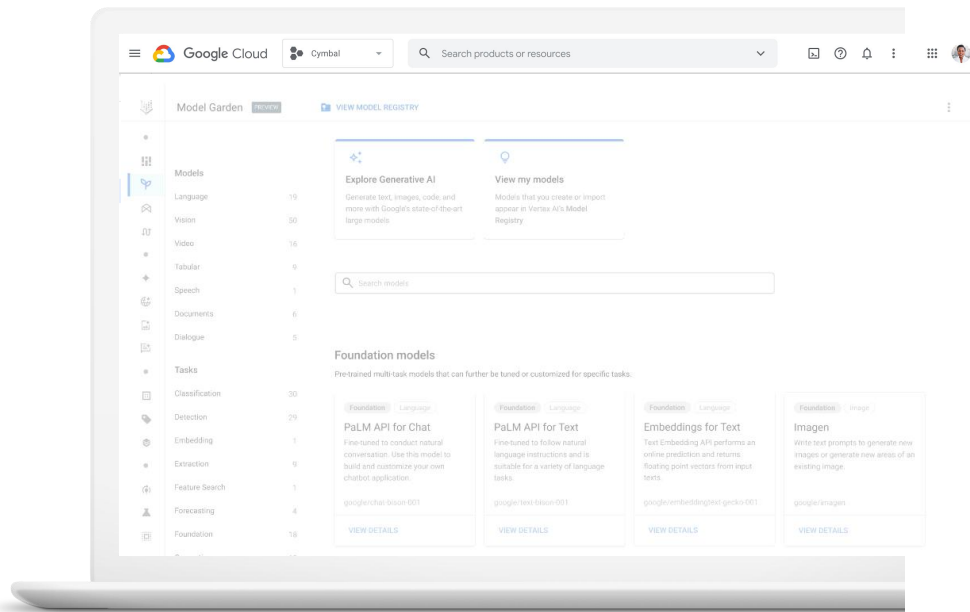
# Vertex AI Platform provides data science, ML, and Gen AI practitioners with tools to customize and build



# Vertex AI Model Garden

One place to launch various enterprise user journeys

- 01 | One stop shop
- 02 | Use Foundation Models directly
- 03 | Tune Foundation Models
- 04 | Customize popular open source models
- 05 | API access to task-specific solutions



# Generative AI Studio

Simple UI for interfacing, experimenting, and tuning models

The screenshot displays the 'Generative AI Studio' interface with three tabs: 'PROMPT GALLERY', 'PROMPT DESIGN' (active), and 'TUNING'. The 'PROMPT DESIGN' tab contains a 'Prompt' section with a text area and three buttons: 'SAVE TO GALLERY', 'VIEW CODE', and 'PROMPT EXAMPLES'. The prompt text is: 'We want to create a multi-media campaign that highlights the simple ingredients and sustainable sourcing practices from our most popular granola bar. We want to highlight: 

- We use only 6 simple ingredients: organic oats, organic almonds, biodynamic, local honey, organic whole grain flour, dried organic cranberries, free-range eggs
- Our packaging is made out of 100% compostable materials.
- We use rainwater harvesting in our oat and almond farming to reduce the need for irrigation.
- We use solar energy to power irrigation systems and other farm equipment
- Wherever we can, we support local farmers

 Blog Headline: Blog Post: Instagram Caption: Instagram Hashtags:'. Below the prompt is an 'Output' section with 'REFRESH OUTPUT' and 'CLEAR OUTPUT' buttons. The output text is: 'Blog Headline: 6 Simple Ingredients, 100% Sustainable: Our New Granola Bar is a Wholesome Treat for the Whole Family Blog Post: At Cymbal, we believe that the best food is simple and sustainable. That's why we sustainably source our eggs from local farmers who use sustainable farming practices. Our most popular granola bar is made with just six simple, organic ingredients: oats, almonds, honey, whole grain flour, cranberries, and eggs.' On the right side, there is a 'Model' dropdown set to 'PaLM Bison Text', and four sliders: 'Temperature' (set to 0), 'Max output tokens' (set to 0), 'Top K' (set to 4), and 'Top P' (set to 0). At the bottom right are 'RESET' and 'SUBMIT' buttons.

- 01 | A simple UI for interacting with models
- 02 | Tune models with your own data
- 03 | Use models in production
- 04 | Multiple modalities



## Vertex AI Pre-Trained Models

Best in class tools allowing customers to leverage Google's leadership in AI to solve common problems



### Vision



Vision



AutoML Vision



Video Intelligence



AutoML Video Intelligence



### Language



Translation



AutoML Translation



Natural Language



AutoML Natural Language



### Conversation



Dialogflow



Speech-to-Text



Text-to-Speech



Speaker ID



### Structured data

● AutoML Tables

● Time Series Insights API

● Vertex AI Forecast

● TabNet

● Fleet Routing API



## Vertex AI areas of investment



1

Unified Data & AI  
platform for all users  
to accelerate time  
to value



2

End-to-End MLOps  
to efficiently and  
responsibly manage  
and govern AI



3

Open and Scalable AI  
Infrastructure  
to flexibly and  
successfully deploy AI



4

State-of-the-Art AI to drive innovation and improve customer outcomes

# Unified Data & AI Platform

# An Integrated Data $\Rightarrow$ Value Journey

ANY USER



**Data engineers**  
Clean, useful data



**ML engineers**  
Integrated intelligence



**Data scientists**  
Models that work



**Developers**  
Intelligent apps



**Data analysts**  
Query and analyze



**Business users**  
Insights Everywhere



**Consumers**  
Value

ANY DATA



Corporate and  
3rd Systems



E-Commerce



SaaS  
Applications



IoT Devices



Social Media



Geospatial

ANY  
WORKLOAD

DATA ENGINEERING



Dataproc



Dataflow

DATA ANALYSIS



BigQuery

MODEL DEVELOPMENT



Vertex  
AI

ML ENGINEERING



Vertex  
AI

INSIGHTS ACTIVATION



Vertex  
AI



Looker



Cloud  
Run



Cloud Composer



Vertex AI Pipelines

Google's Data Cloud

ANY  
INNOVATION



Decision  
Making



App  
Development



SaaS  
Applications



Exception  
Management



Operational  
Intelligence



Data  
Monetization

# BigQuery



Query acceleration



Multi-cloud



GIS



Integrated ML



# Vertex AI



Call our  
perception APIs



Build your  
own models



Train state of  
the art models



Deploy, monitor, &  
govern confidently

## 27%

Lower TCO than cloud data warehouse  
alternatives

## 99.99%

Uptime SLA ensures reliable access to  
data and insights

## 5X

Build and train models 5X faster,  
compared to traditional notebooks

## 80%

Fewer lines of code needed to build  
custom models

# BigQuery ML

Train models in SQL; manage, orchestrate, and deploy directly to Vertex AI

## Rich Built-in Models

- **Supervised Learning:** Linear / Logistic regression, Boosted trees, Random Forest, DNN, Wide & Deep models, AutoML Tables
- **Unsupervised Learning:** K-Means, PCA, Matrix Factorization, Autoencoder
- **Time Series:** ARIMA\_PLUS univariate and multivariate

## Business Applications

- Forecasting
- Anomaly Detection
- Recommendation
- Prediction
- Natural Language Processing
- Computer Vision

## Inference Engine

- **Remote Models:** with Vertex Endpoint
- **Imported Models:** with TF, TFLite, XGBoost & ONNX (enabling sklearn, Pytorch, SparkML)
- **Pretrained Models for Unstructured Data:** LLMs, Vision, NL, Translate

## ML Ops & Cycles

- Hyperparameter Tuning using Vertex
- Explainable AI
- Integration with Vertex Model Registry
- Components for Vertex Pipelines
- Monitoring with Vertex Model Monitoring

Pre-Announcing - Generally Available on Vertex AI

# Colab Enterprise on Vertex AI

Combines the ease of use of Google Colab notebooks with the enterprise-level security and compliance capabilities of Google Cloud

## Collaboration & Productivity

IAM-based notebook sharing

Automatic Versioning

Commenting (coming soon!)

Co-editing (coming soon!)

Generative AI powered code completion and generation

## Zero-Config & Flexible Compute

Provides both zero-config compute options, as well as access to a wide range of machine-shapes and compute

## Enterprise Ready

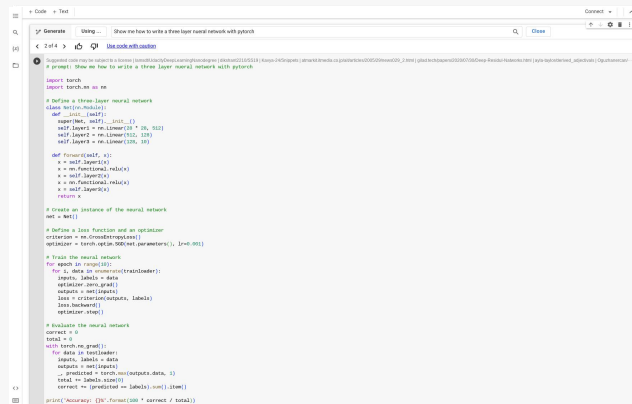
Will support a wide range of security and management capabilities including:

- VPC-SC
- CMEK
- Regionalization
- Cloud Monitoring
- Cloud Logging

## Available across Google Cloud

Available in BigQuery and Vertex AI (Dataproc and Dataflow coming soon), making it easy to work across data and AI workloads

**Use Cases:** Data science, data analysis, data engineering, ML engineering



# Vertex AI Workbench

A one-stop surface for Data Science



## Fully managed compute with admin control

A Jupyter-based fully managed, scalable, enterprise-ready compute infrastructure with easily enforceable policies and user management



## Fast workflow for data tasks

Seamless visual and code-based integrations with data & analytics services



## At-your-fingertips integration

Load and share notebooks alongside your AI and data tasks. Run tasks without extra code

The screenshot displays the Google Cloud Platform Vertex AI Workbench interface. The top panel shows the 'Workbench' page with a list of managed notebooks. The bottom panel shows the 'mchrestkha-sandbox' notebook environment with a file explorer and a launcher for various data science tools.

Notebook name	Location	Access mode
managed-notebook-1643391246	us-central1-f	Single user only
managed-notebook-1643393492	us-central1-c	Single user only
managed-notebook-1647318683	us-central1-c	Single user only
mchrestkha-sandbox	us-central1-a	Single user only
nvidia-ngc	us-central1-f	Single user only
tf-mnist-ngc	us-central1-f	Single user only

The bottom panel shows the 'mchrestkha-sandbox' notebook environment. The file explorer on the left shows a directory structure with 'src' and 'tutorials' folders. The launcher on the right provides a grid of icons for various data science tools, including Python (Local), PySpark (Local), PySpark on cluster-5977-m, Python 3 on cluster-5977-m, PyTorch (Local), R (Local), R on cluster-5977-m in us, scylla-kernel on cluster-, TensorFlow 2 (Local), XGBoost (Local), and Console. A 'Modify hardware' dialog is open, showing the current configuration: 4 vCPUs (34.7%), 15 GB RAM (6.1%), and 1 Tesla T4 GPU (0%).



“From several months to same day set up of development environment with Vertex AI Workbench, Training, and NVIDIA A100s”

# Simplify and Accelerate Data to AI Journey

1 Unified interface for data analysis and AI development

2 Consistent workflow with minimal data movement

3 MLOps tooling for BigQuery ML

Data science surface

Colab Enterprise & Vertex AI Workbench

Data

Analytics

ML

BigQuery (including BigQuery ML)

Vertex AI

Spark

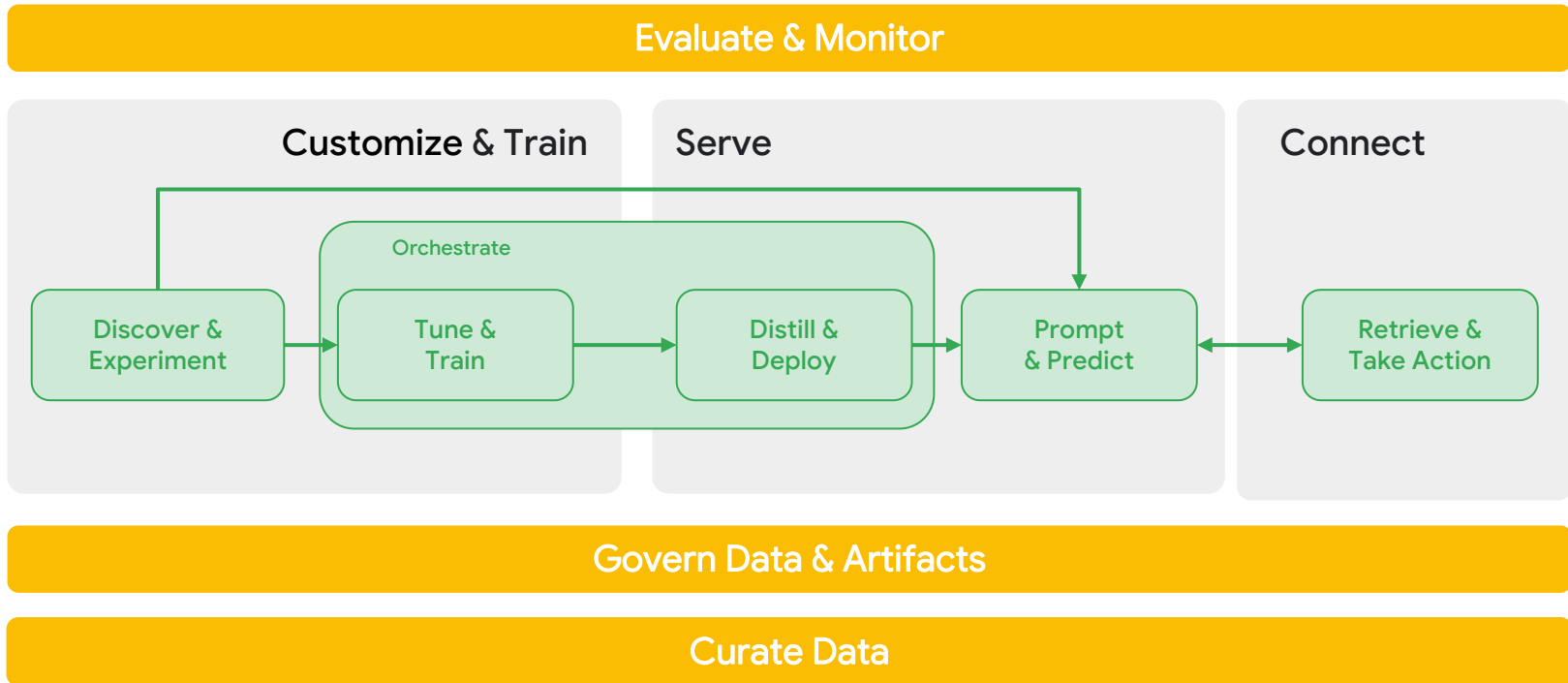
ML Workflow management

Vertex AI: Model Registry, Metadata, Pipelines






# End-to-End MLOps

# MLOps for predictive & generative AI






# Manage and govern your ML models




## Feature Store

-  **Share** and **reuse** ML features across use cases
-  Serve ML Features **at scale** with **low latency**
-  **Alleviate** training serving skew




## ML Metadata

-  **Automatically track** inputs / outputs to all components
-  Track custom metadata **directly from your code**
-  **Visualize, analyze,** and **compare** detailed ML lineage

## Model Registry

-  **Register, organize, track,** and **version** your trained and deployed ML models.
-  **Govern** the model launch process
-  **Maintain** model documentation and reporting

## Model Evaluation

-  Iteratively **run model evaluations** on new datasets at scale
-  Visualize and compare model evaluations to identify the **best model for prod deployment**
-  **Assess the performance of models** on different slices and evaluated annotations

# Vertex AI Experiments

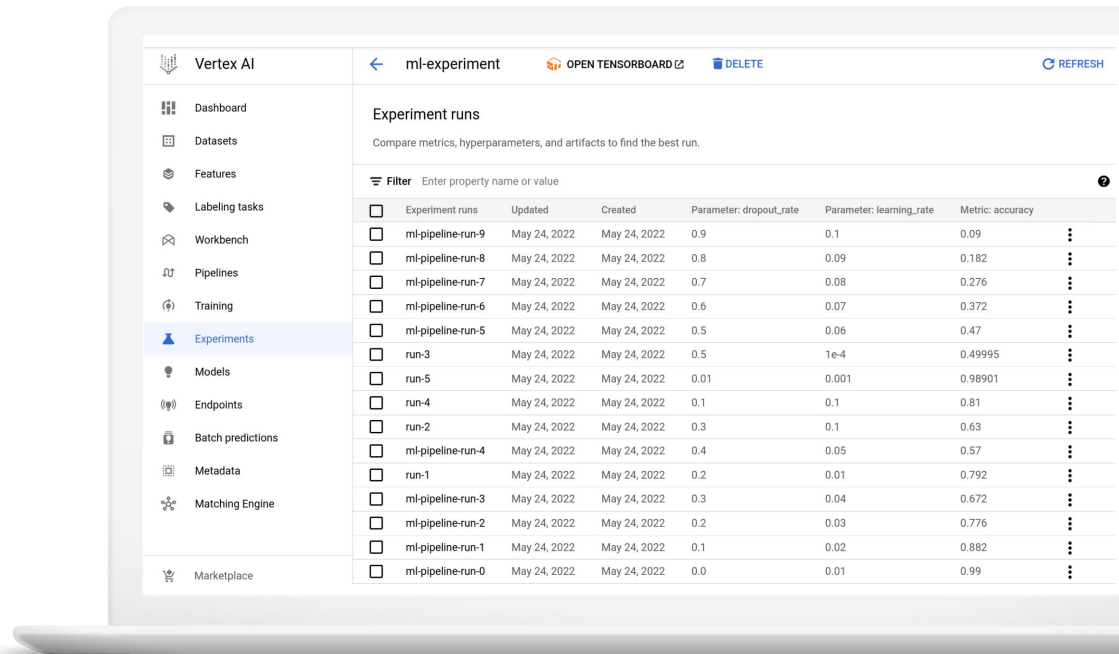
Track and compare multiple experiment runs and analyze key model metrics

Vary and track parameters and metrics as you experiment.

Organize Vertex Pipeline runs and compare their parameters, metrics, and artifacts.

Track steps and artifacts to capture the lineage of experiments.

Compare Vertex Pipelines against Notebook experiments.



The screenshot displays the Vertex AI Experiments interface. On the left is a navigation sidebar with options: Dashboard, Datasets, Features, Labeling tasks, Workbench, Pipelines, Training, Experiments (selected), Models, Endpoints, Batch predictions, Metadata, Matching Engine, and Marketplace. The main panel shows the 'ml-experiment' view with a title 'Experiment runs' and a subtitle 'Compare metrics, hyperparameters, and artifacts to find the best run.' Below this is a filter bar and a table of experiment runs.

<input type="checkbox"/>	Experiment runs	Updated	Created	Parameter: dropout_rate	Parameter: learning_rate	Metric: accuracy	
<input type="checkbox"/>	ml-pipeline-run-9	May 24, 2022	May 24, 2022	0.9	0.1	0.09	⋮
<input type="checkbox"/>	ml-pipeline-run-8	May 24, 2022	May 24, 2022	0.8	0.09	0.182	⋮
<input type="checkbox"/>	ml-pipeline-run-7	May 24, 2022	May 24, 2022	0.7	0.08	0.276	⋮
<input type="checkbox"/>	ml-pipeline-run-6	May 24, 2022	May 24, 2022	0.6	0.07	0.372	⋮
<input type="checkbox"/>	ml-pipeline-run-5	May 24, 2022	May 24, 2022	0.5	0.06	0.47	⋮
<input type="checkbox"/>	run-3	May 24, 2022	May 24, 2022	0.5	1e-4	0.49995	⋮
<input type="checkbox"/>	run-5	May 24, 2022	May 24, 2022	0.01	0.001	0.98901	⋮
<input type="checkbox"/>	run-4	May 24, 2022	May 24, 2022	0.1	0.1	0.81	⋮
<input type="checkbox"/>	run-2	May 24, 2022	May 24, 2022	0.3	0.1	0.63	⋮
<input type="checkbox"/>	ml-pipeline-run-4	May 24, 2022	May 24, 2022	0.4	0.05	0.57	⋮
<input type="checkbox"/>	run-1	May 24, 2022	May 24, 2022	0.2	0.01	0.792	⋮
<input type="checkbox"/>	ml-pipeline-run-3	May 24, 2022	May 24, 2022	0.3	0.04	0.672	⋮
<input type="checkbox"/>	ml-pipeline-run-2	May 24, 2022	May 24, 2022	0.2	0.03	0.776	⋮
<input type="checkbox"/>	ml-pipeline-run-1	May 24, 2022	May 24, 2022	0.1	0.02	0.882	⋮
<input type="checkbox"/>	ml-pipeline-run-0	May 24, 2022	May 24, 2022	0.0	0.01	0.99	⋮

# Model Monitoring



## Monitor and alert

Monitor signals for model's predictive performance (batch and online), and alert when those signals deviate.



## Diagnose

Help identify the cause for the deviation i.e. what changed, how and how much?



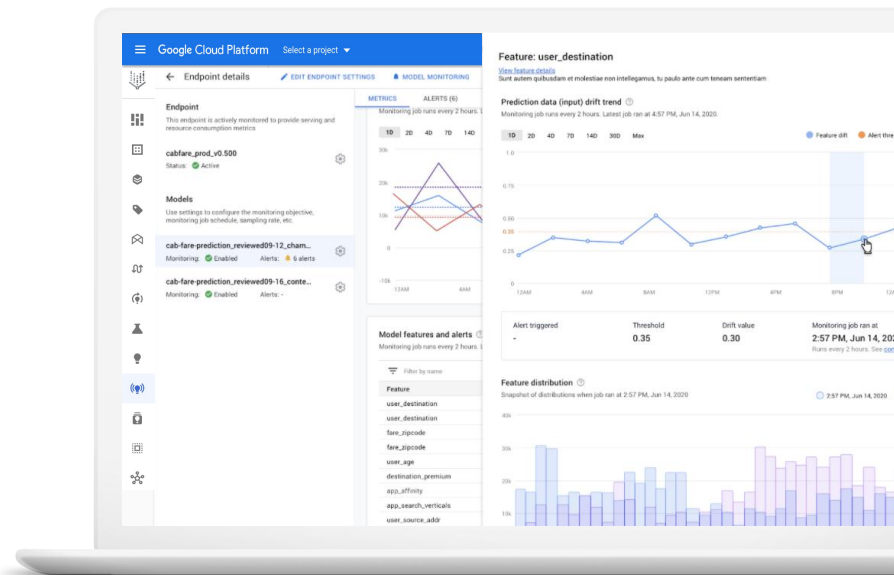
## Update Model

Trigger model re-training pipeline or collect relevant training data to address performance degradation.



## Integrated with Feature Store

Monitor and set up alerts for Feature Store performance and resource utilization, and track how much a feature's value distribution changes over time



# Open and Scalable AI Infrastructure

# Accelerate getting models to production

While maintaining the flexibility of ML frameworks and compute options

ANY DATA SOURCE



BigQuery



Cloud Storage



Filestore



Persistent Disk



Multi-Cloud Data

ANY FRAMEWORK

Spark

TensorFlow

beam

PyTorch

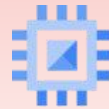
pandas

XGBoost

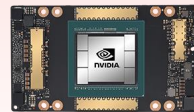
DASK



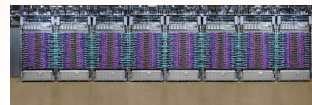
ANY HARDWARE



CPUs



GPUs



# Run and scale your code with high availability using Vertex AI Training

Serverless experience,  
no provisioning

Rapid cluster orchestration

Built-in logging and  
monitoring



Ephemeral clusters on-  
demand

Automatic job queuing

Pay for only what you use



*"Vertex AI Workbench and Vertex Training have both accelerated our adoption of highly scalable model development and training capabilities, with a focus on being purpose built at scale for Wayfair's needs. It has received really enthusiastic adoption from our own data science community."*

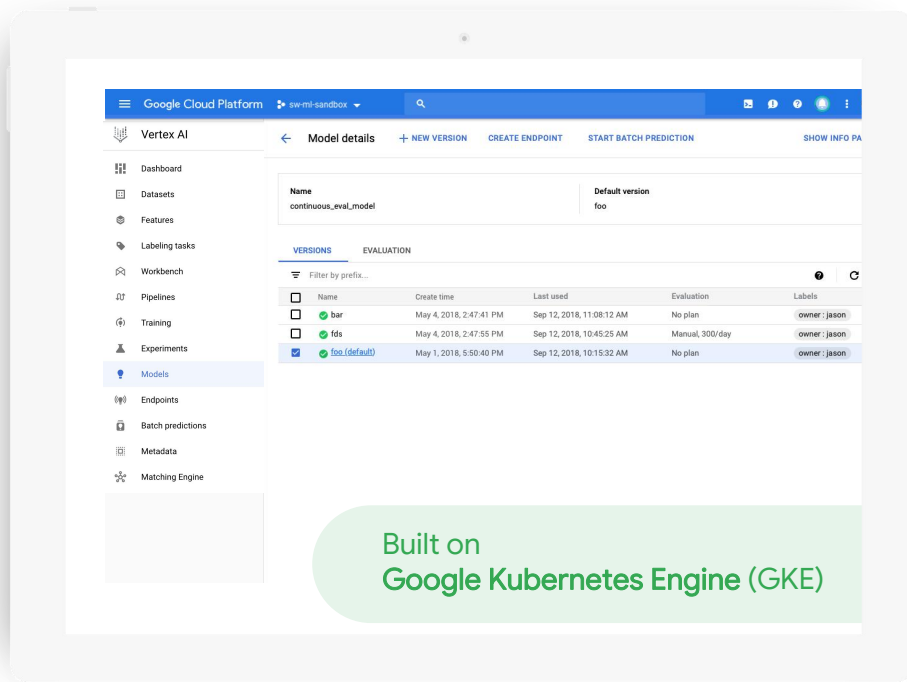


# Robust and reliable model hosting with Vertex AI Prediction

- Serve **online** endpoints for low-latency predictions, or predictions on massive **batches** of data.
- **Built-in security and compliance:** VPC peering and security perimeter. Custom managed encryption keys. Fine-tuned access control.
- **Low TCO:** Scale **automatically** based on your traffic, and alleviate operational overhead.
- **Fast inference on GPUs:** Support for a broad range of machine types specialized for ML, such as GPUs.



“Large Transformer models like ours are more cost-effective when running on GPU even for inference (this was confirmed by our load test experiments with a 6x cost reduction factor between T4 GPU and CPU).”



# Vertex AI Platform is for Builders

Accelerate time to value by building on top of Google's world class infrastructure and services that help ensure security and reliability



## State of the art

- Built on Google research and continuous innovation
- Best in class selection of 1P, OSS and 3P models



## End-to-end governance

- **Prompting, Tuning & Distillation:** Customize LLMs for your domain and use case. Transfer and distill large scale learning to your models. Leverage Vertex AI for LLMOps managements
- **MLOps:** Leverage Vertex AI's capabilities for model evaluation, model management, prompt engineering, prompt management, and deployment



## Enterprise readiness

- **Your Data, Your Terms:** Control and protect your data at every step of training and deployment
- **Responsible AI:** Tooling, enablement, and support to empower customers to build responsible Generative AI
- **Enterprise-ready, out of the box:** Accelerate time to value and developer efficiency with developer-friendly tooling built on enterprise security, reliability, and scalability

Q&A

Thank you