

Ask the Expert '21

～機械学習～



上野 陽子 (Yoko Ueno)

Google Cloud

Technical Account Manager,
Professional Services



- Gaming Industry
- Data Engineer
- IoT, Raspberry Pi



松田実法 (Minori Matsuda)



Vending Machine Division
Data Scientist Manager
MBA



- Statistics
- Analytics
- Machine Learning
- Natural Language Processing
- Google Developer Expert(ML)

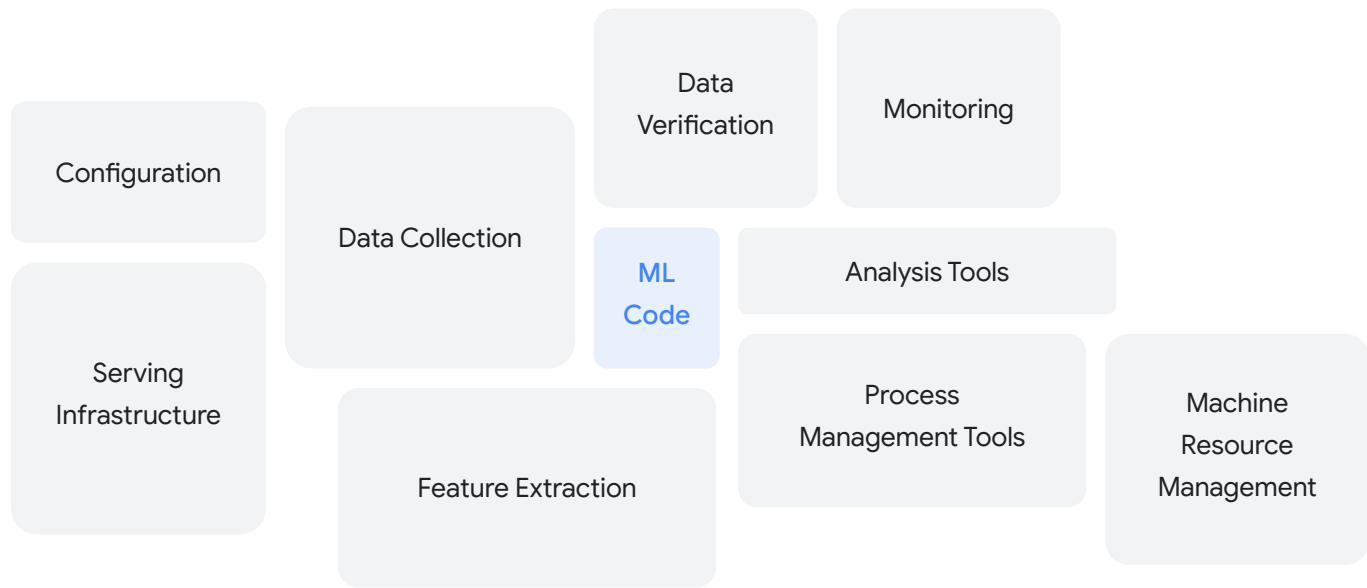


アジェンダ

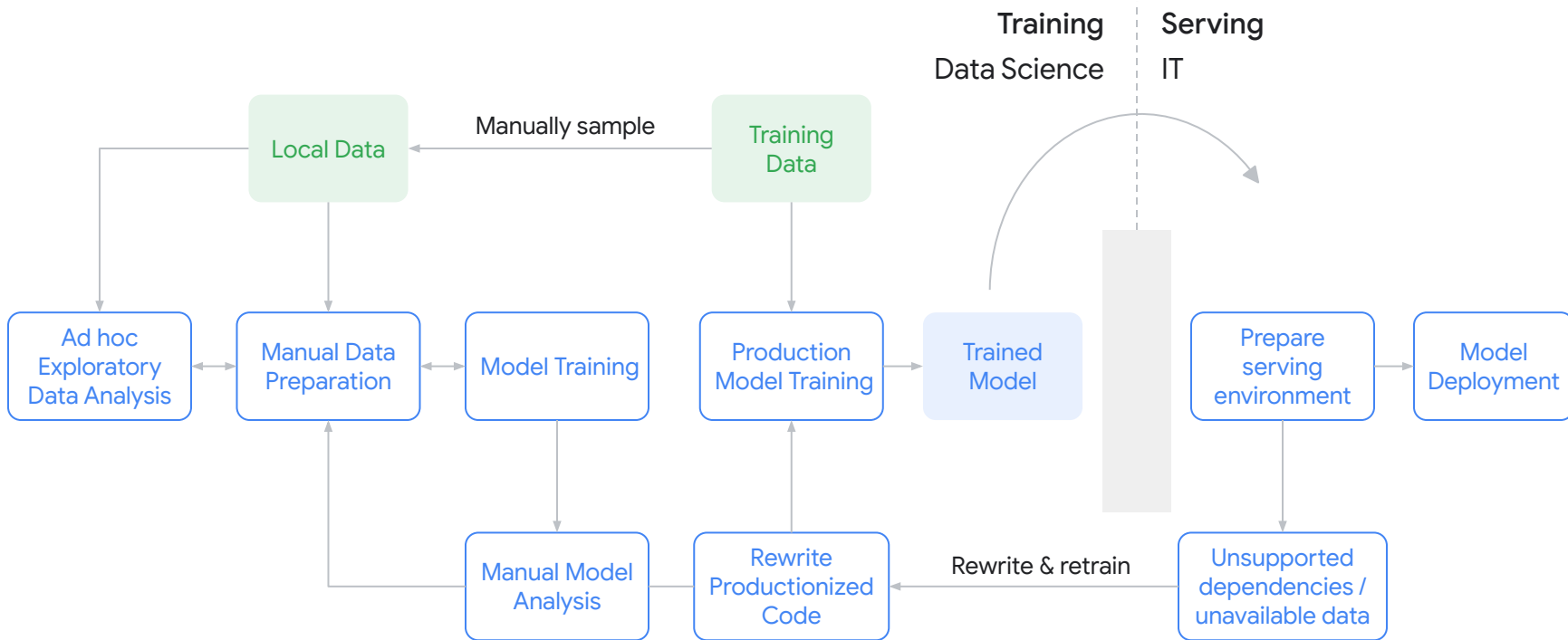
1. Vertex AI
2. AutoML Forecast
3. まとめ

Vertex AI

01

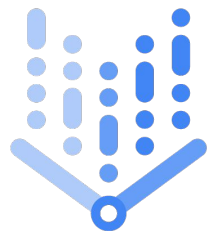


データサイエンスとITを隔てる壁



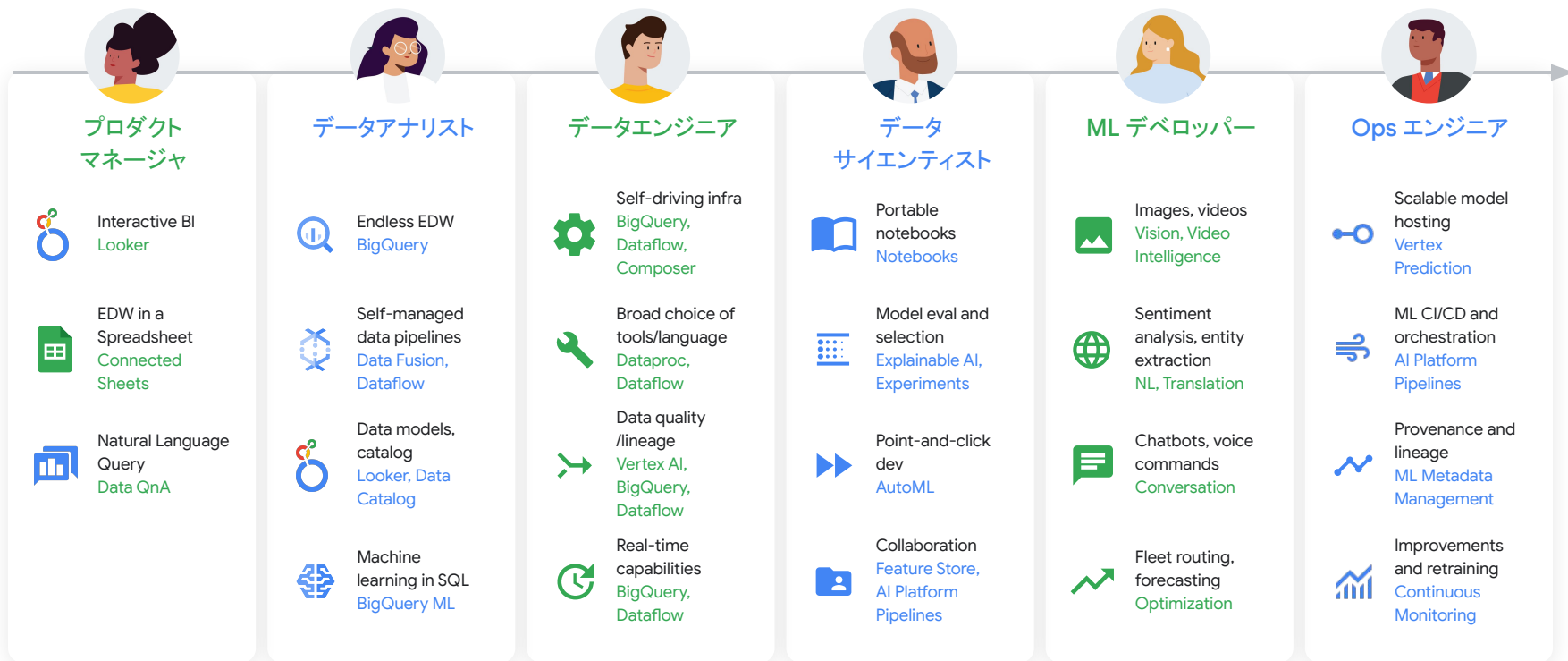
2018 年の MLOps



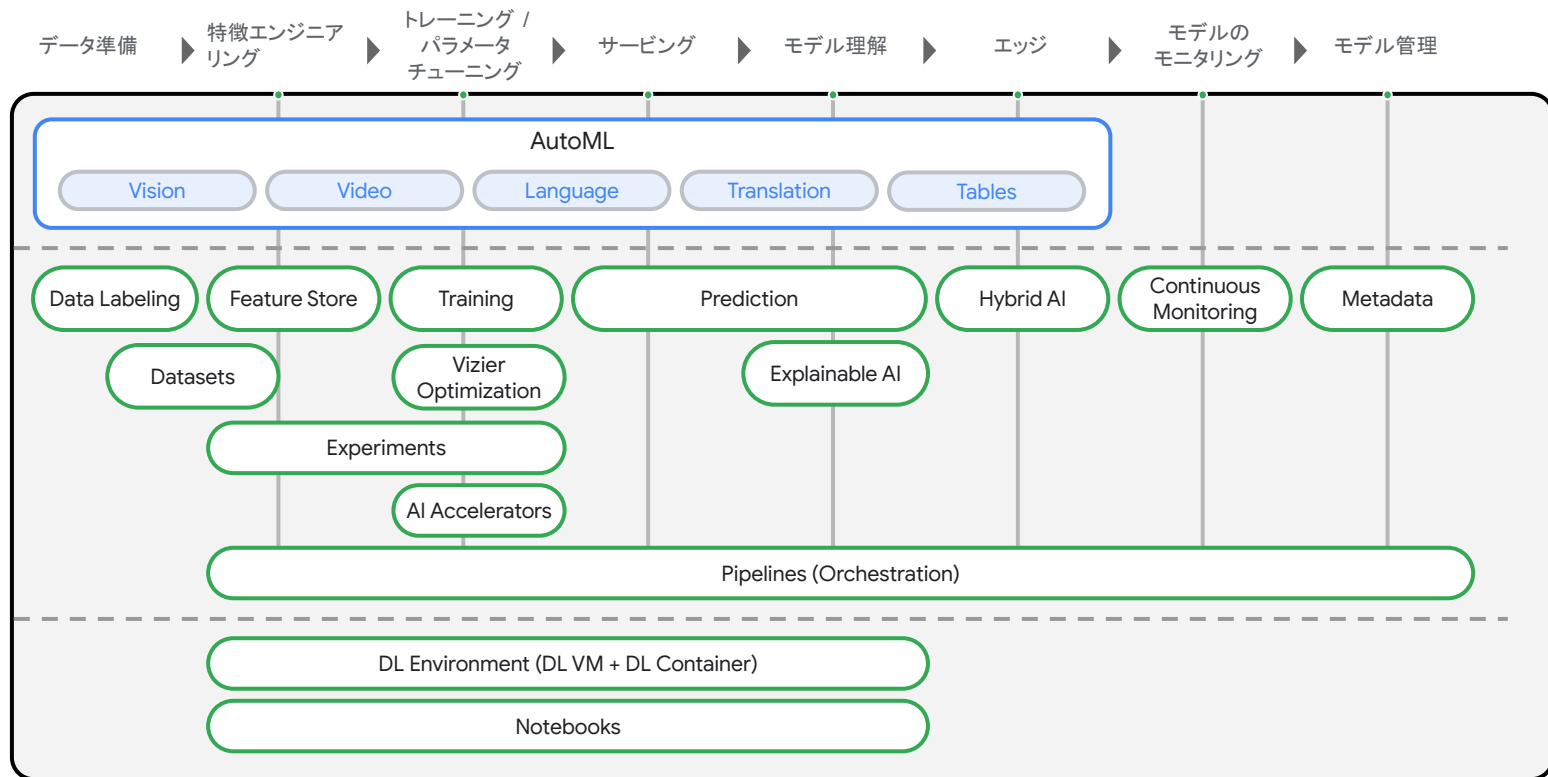


Vertex AI は
機械学習システムの
開発から実運用までを支える
統合プラットフォーム

ML ライフサイクルに関わるすべての担当者が共有



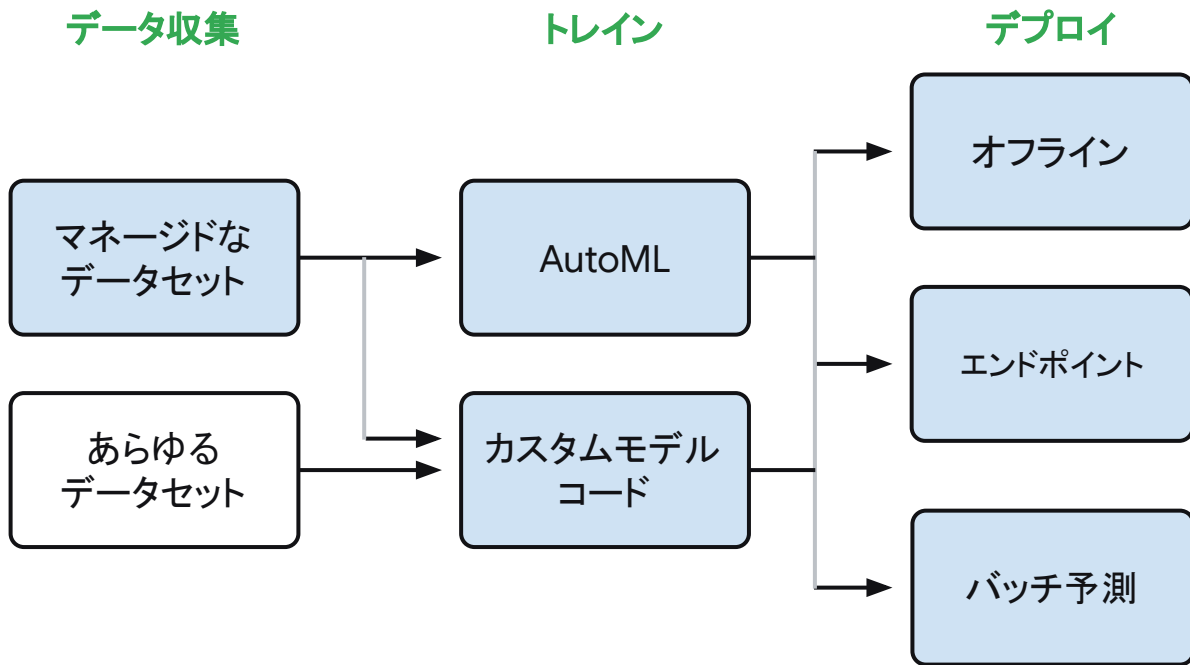
Vertex AI を構成する機能



3つの学習方法を統合、モデルの共有とデプロイが可能

Train with	Data Analyst	ML Developer	Data Scientist	Serve with
Vertex Training			✓	
AutoML		✓	✓	Vertex Prediction
BigQuery ML	✓	✓	✓	

Vertex AI でのモデルのデプロイ



Datasets

Select an objective

An objective is an outcome you want to achieve with a trained model. Don't worry, you can use this dataset for other image-based objectives later.

IMAGE

TABULAR

TEXT

VIDEO



☒ Image classification (Single-label)

Predict the one correct label that you want assigned to an image.



☐ Image classification (Multi-label)

Predict all the correct labels that you want assigned to an image.



☐ Image object detection

Predict all the locations of objects that you're interested in.

Notebooks

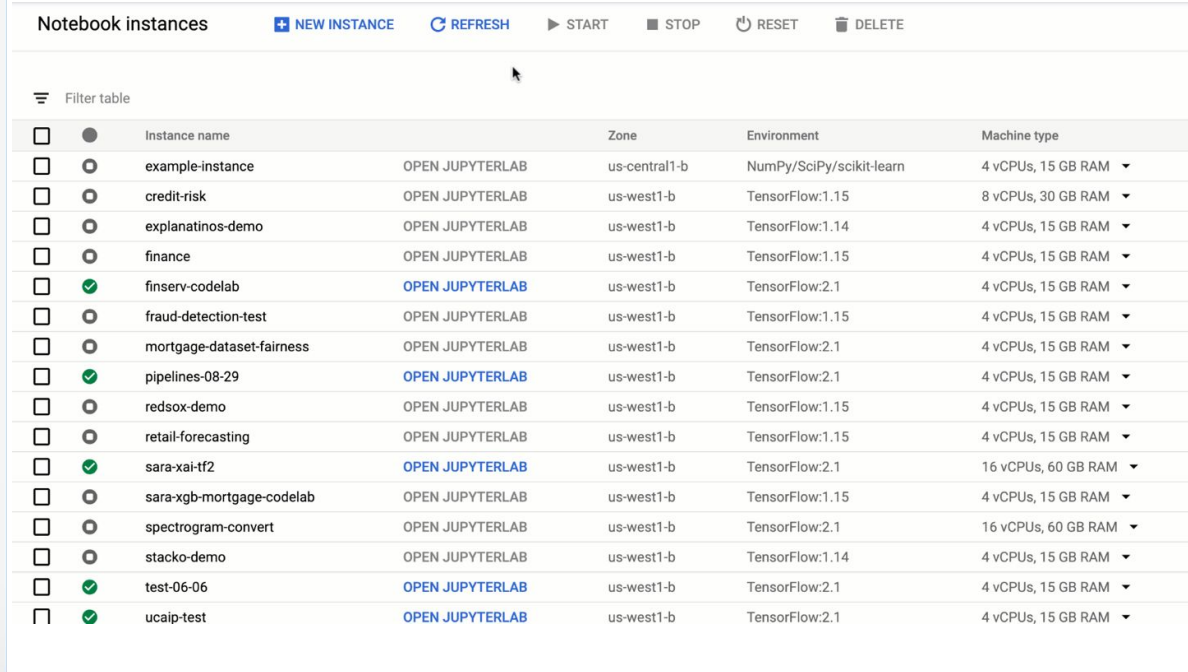
クラウドで動く Jupyter Notebook

データ分析系ツールを導入済み

BigQuery や Dataflow 等と統合

1 CPU から 16 A100 まで

Git 統合



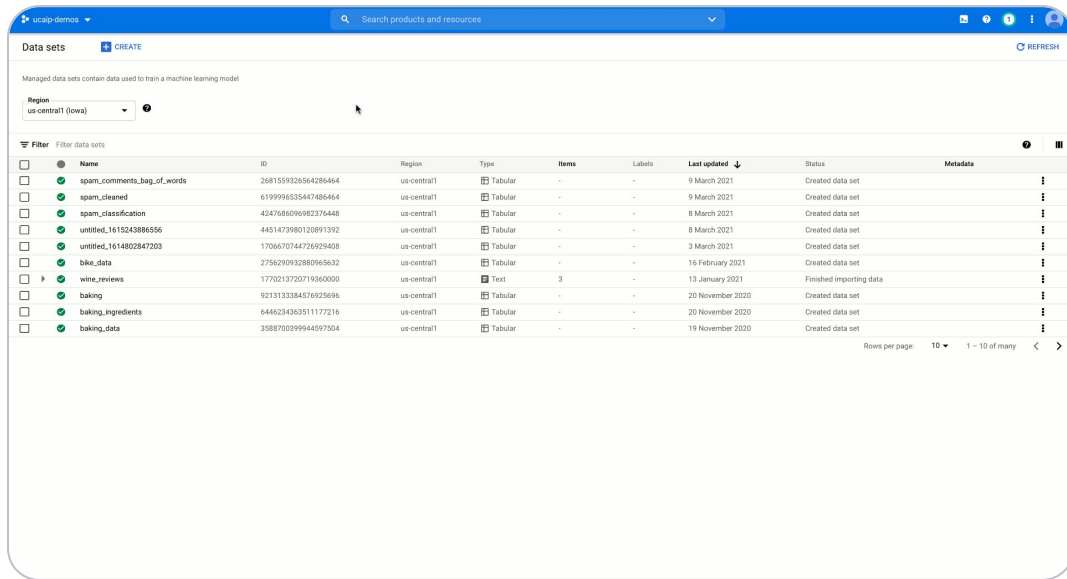
The screenshot shows the 'Notebook instances' management page in Google Cloud. At the top, there are buttons for 'NEW INSTANCE', 'REFRESH', 'START', 'STOP', 'RESET', and 'DELETE'. Below the buttons is a 'Filter table' section. The main table lists various notebook instances with columns for selection, status, instance name, environment, zone, and machine type. Some instances are running (green dot) and others are stopped (grey dot). The 'OPEN JUPYTERLAB' link is visible for several instances.

☐	☒	Instance name	Environment	Zone	Machine type
☐	☐	example-instance	OPEN JUPYTERLAB	us-central1-b	NumPy/SciPy/scikit-learn 4 vCPUs, 15 GB RAM
☐	☐	credit-risk	OPEN JUPYTERLAB	us-west1-b	TensorFlow:1.15 8 vCPUs, 30 GB RAM
☐	☐	explanatinos-demo	OPEN JUPYTERLAB	us-west1-b	TensorFlow:1.14 4 vCPUs, 15 GB RAM
☐	☐	finance	OPEN JUPYTERLAB	us-west1-b	TensorFlow:1.15 4 vCPUs, 15 GB RAM
☐	☒	finserve-codelab	OPEN JUPYTERLAB	us-west1-b	TensorFlow:2.1 4 vCPUs, 15 GB RAM
☐	☐	fraud-detection-test	OPEN JUPYTERLAB	us-west1-b	TensorFlow:1.15 4 vCPUs, 15 GB RAM
☐	☐	mortgage-dataset-fairness	OPEN JUPYTERLAB	us-west1-b	TensorFlow:2.1 4 vCPUs, 15 GB RAM
☐	☒	pipelines-08-29	OPEN JUPYTERLAB	us-west1-b	TensorFlow:2.1 4 vCPUs, 15 GB RAM
☐	☐	redsox-demo	OPEN JUPYTERLAB	us-west1-b	TensorFlow:1.15 4 vCPUs, 15 GB RAM
☐	☐	retail-forecasting	OPEN JUPYTERLAB	us-west1-b	TensorFlow:1.15 4 vCPUs, 15 GB RAM
☐	☒	sara-xai-tf2	OPEN JUPYTERLAB	us-west1-b	TensorFlow:2.1 16 vCPUs, 60 GB RAM
☐	☐	sara-xgb-mortgage-codelab	OPEN JUPYTERLAB	us-west1-b	TensorFlow:1.15 4 vCPUs, 15 GB RAM
☐	☐	spectrogram-convert	OPEN JUPYTERLAB	us-west1-b	TensorFlow:2.1 16 vCPUs, 60 GB RAM
☐	☐	stacko-demo	OPEN JUPYTERLAB	us-west1-b	TensorFlow:1.14 4 vCPUs, 15 GB RAM
☐	☒	test-06-06	OPEN JUPYTERLAB	us-west1-b	TensorFlow:2.1 4 vCPUs, 15 GB RAM
☐	☒	ucaip-test	OPEN JUPYTERLAB	us-west1-b	TensorFlow:2.1 4 vCPUs, 15 GB RAM

AutoML トレーニング

コードを記述せずに、
最新のモデルを短時間で
トレーニングできます

Explainable AI を利用して、
詳細なモデル評価指標を
取得します



☐	Name	ID	Region	Type	Items	Labels	Last updated	Status	Metadata
☐	spam_comments_bag_of_words	2681559326564286464	us-central1	Tabular	-	-	9 March 2021	Created data set	
☐	spam_cleaned	6199996535647486464	us-central1	Tabular	-	-	9 March 2021	Created data set	
☐	spam_classification	424768696982376448	us-central1	Tabular	-	-	8 March 2021	Created data set	
☐	untitled_161524388656	4461479980120891392	us-central1	Tabular	-	-	8 March 2021	Created data set	
☐	untitled_1614802847203	1706670744726929408	us-central1	Tabular	-	-	3 March 2021	Created data set	
☐	bike_data	2756290932880965632	us-central1	Tabular	-	-	16 February 2021	Created data set	
☒	wine_reviews	1770213720719360000	us-central1	Text	3	-	13 January 2021	Finished importing data	
☐	baking	9213133384576925696	us-central1	Tabular	-	-	20 November 2020	Created data set	
☐	baking_ingredients	644623436351177216	us-central1	Tabular	-	-	20 November 2020	Created data set	
☐	baking_data	3588700399944597304	us-central1	Tabular	-	-	19 November 2020	Created data set	

Rows per page: 10 1 - 10 of many

カスタムトレーニング

カスタム コンテナを介して、選択した
フレームワークで構築されたモデルを
トレーニングします

分散トレーニング ジョブを実行し、
GPU を使用してトレーニングを加速します

Vizier の最適化アルゴリズムによる
ハイパーパラメータ調整を行います

Train new model

- ✓ Define your model
- ✓ Choose training method
- 3 Container settings**
- 4 Hyperparameter tuning
- 5 Compute and pricing
- 6 Inference settings

START TRAINING CANCEL

Select a pre-built container or build a custom container using ML frameworks (as well as non-ML dependencies, libraries and binaries) that are not otherwise supported. [Learn more](#)

☒ Pre-built container
View the list of [supported runtimes](#) including TensorFlow, scikit-learn and PyTorch versions

☐ Custom container
Build a custom Docker container. Must be stored in [Container Registry](#)

Pre-built container settings

Before you begin, you need to package and upload your application code and dependencies to a Cloud Storage bucket. [Learn more](#)

In order to run in a pre-built container, your code needs to be in Python 3.7

Model framework *
TensorFlow

Model framework version *
2.1

gs:// Package location (Cloud Storage path) * BROWSE
Learn how to [package and upload](#) your application code and dependencies

[+ ADD PACKAGE](#)

Python module *

gs:// Model output directory BROWSE
Your model artifacts and other data needed for training will be stored on Cloud Storage. You should specify a path here if you do not set an output directory in your application code or arguments.

スケーラブルなエンドポイントへのデプロイ

統合エンドポイント EX

AutoMLで
作成されたモデル

カスタムコードで作
成されたモデル

Endpoints [+ CREATE ENDPOINT](#)

Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.

To create an endpoint, you need at least one machine learning model

Region
us-central1 (Iowa) ?

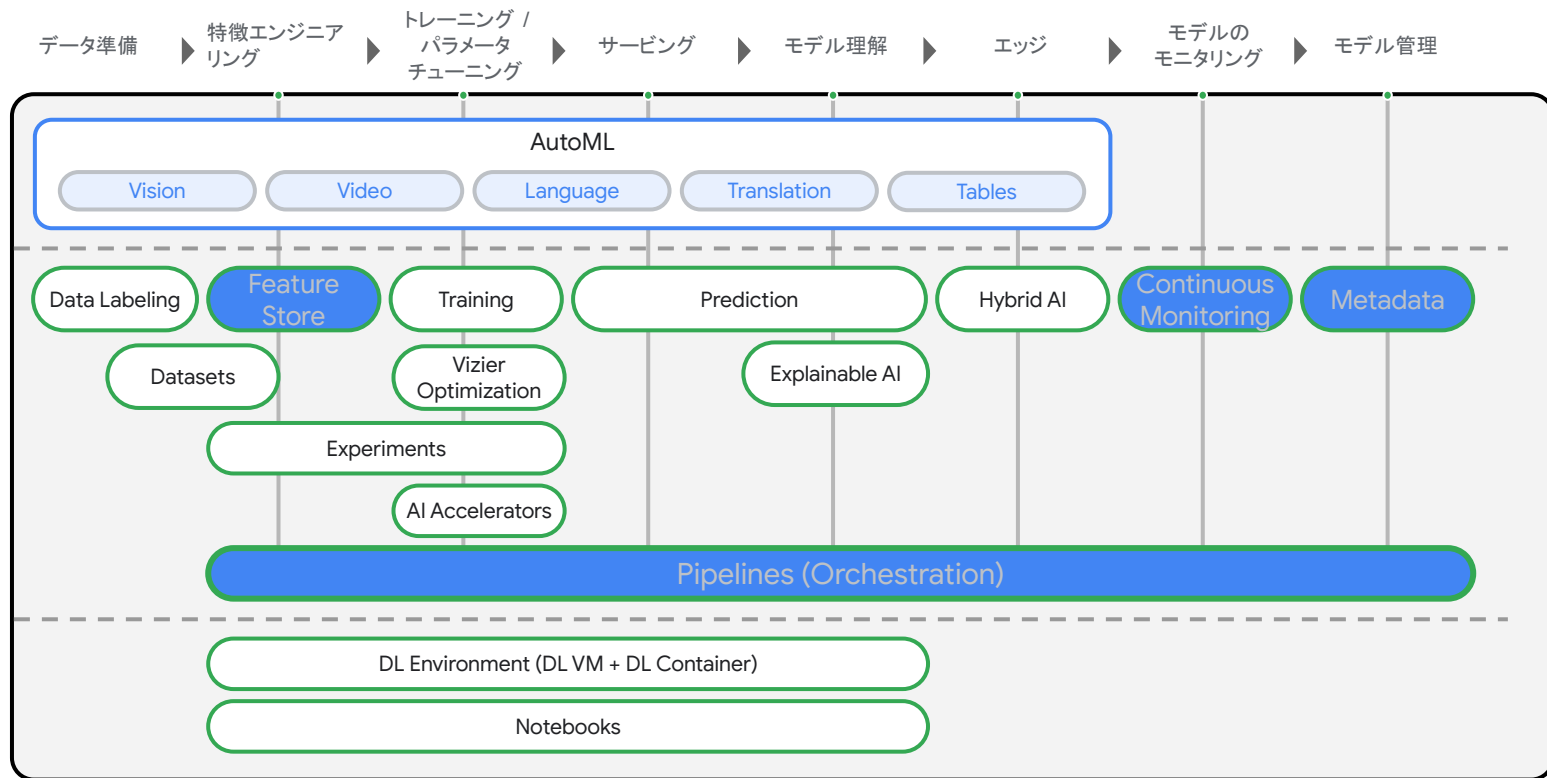
Filter endpoints

☐	Endpoint	ID	Models	Region
☐	v1	8619010077484908544	1	us-central1
☐	v2	2518180694769795072	1	us-central1

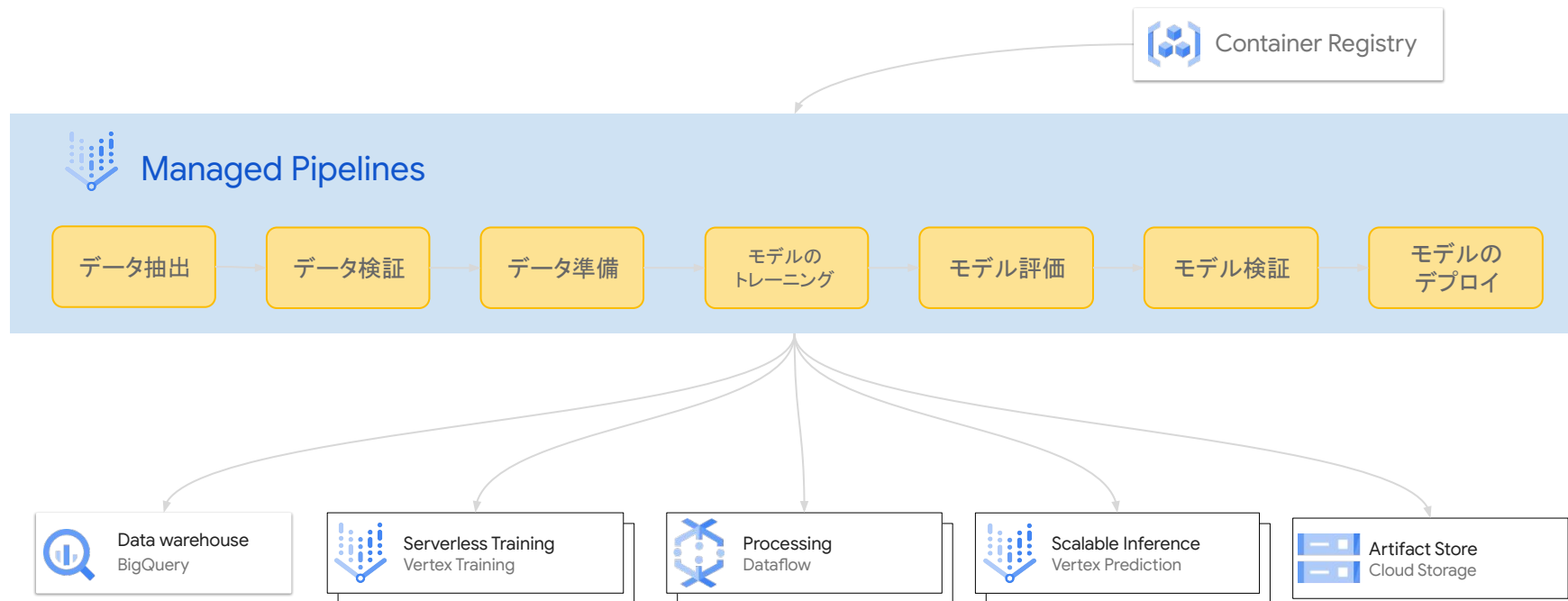
汎用フレームワークで構築されたモデルの提供
モデル間でトラフィックを分散
エンドポイントのマシントップをカスタマイズ



ML Ops で活用される Vertex AI の機能



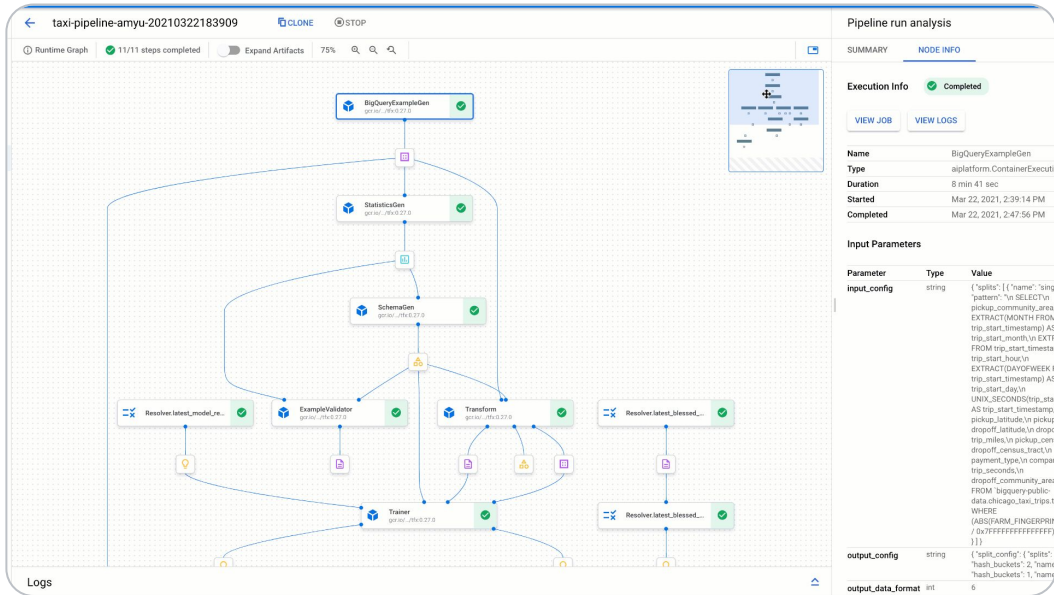
Vertex Pipelines で ML Ops をシンプルに



Vertex Pipelines

TFX や Kubeflow pipeline などの
使い慣れたオープンソース
Python SDK を使用して
パイプラインの構築

自動化され、スケーラブルで、
サーバーレス、費用対効果が高い

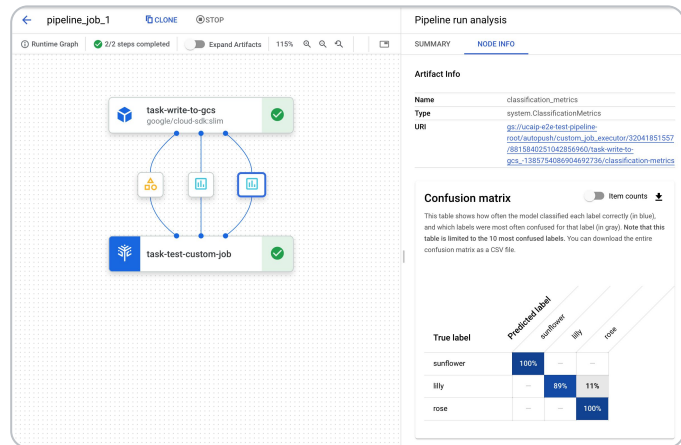
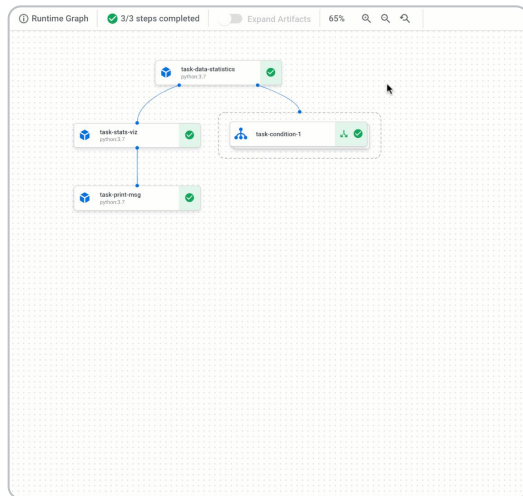


Vertex Pipelines

パイプラインに条件付きロジックと
ブランチの追加

パイプラインによって生成されたすべての
アーティファクトのメタデータを保存します

ML ワークフロー全体でアーティファクト、
系統、メトリック、および実行の追跡



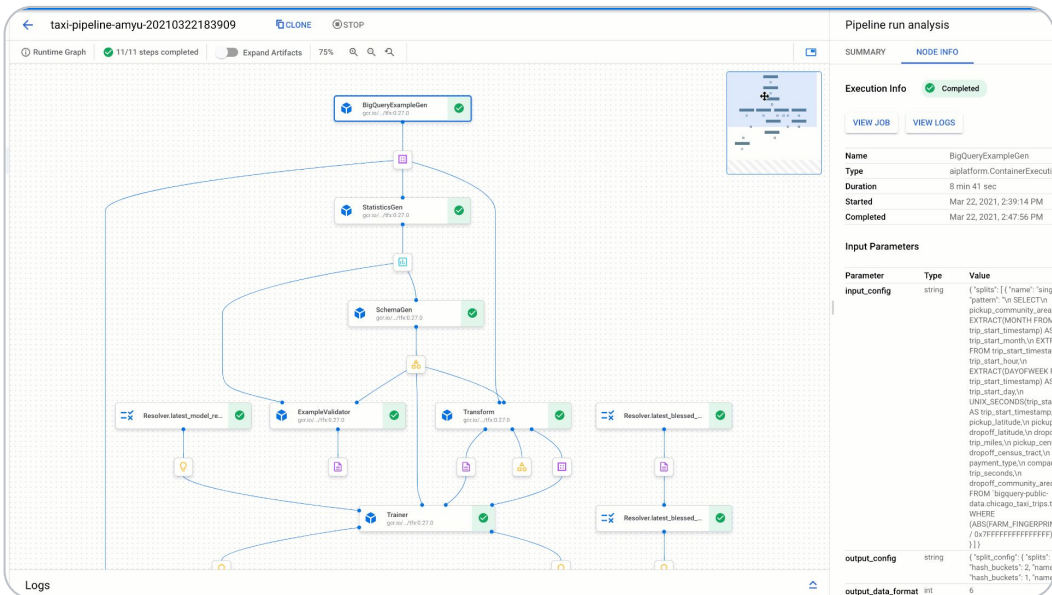
Vertex Pipelines

ML ライフサイクルにおける CI/CD と実験管理

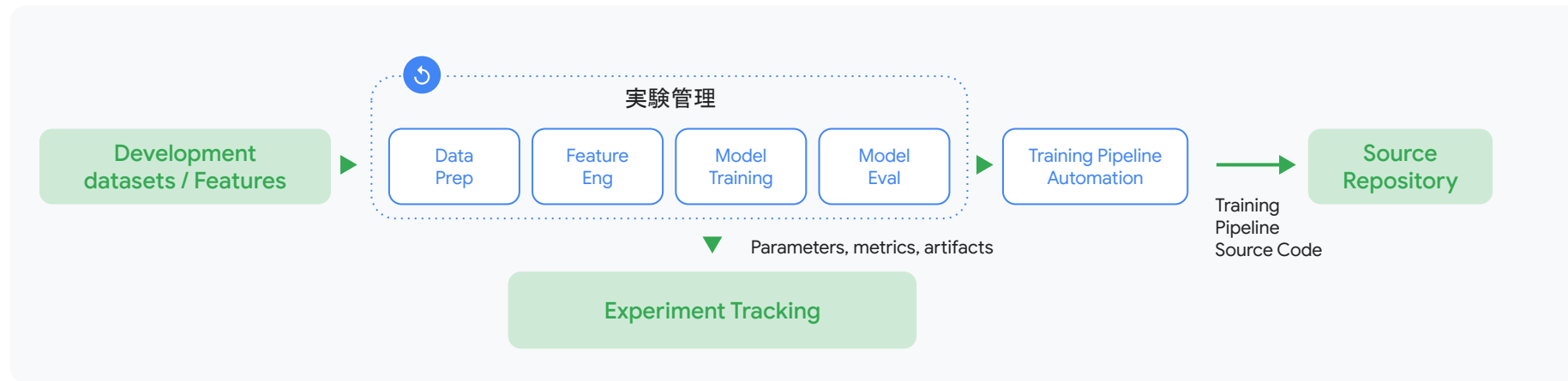
アーティファクト(データ、特徴量、
モデル、パイプライン、実験結果)の
再現可能性、監査、ガバナンス

Kubeflow Pipelines と TensorFlow Extension (TFX) の サーバレス実行環境

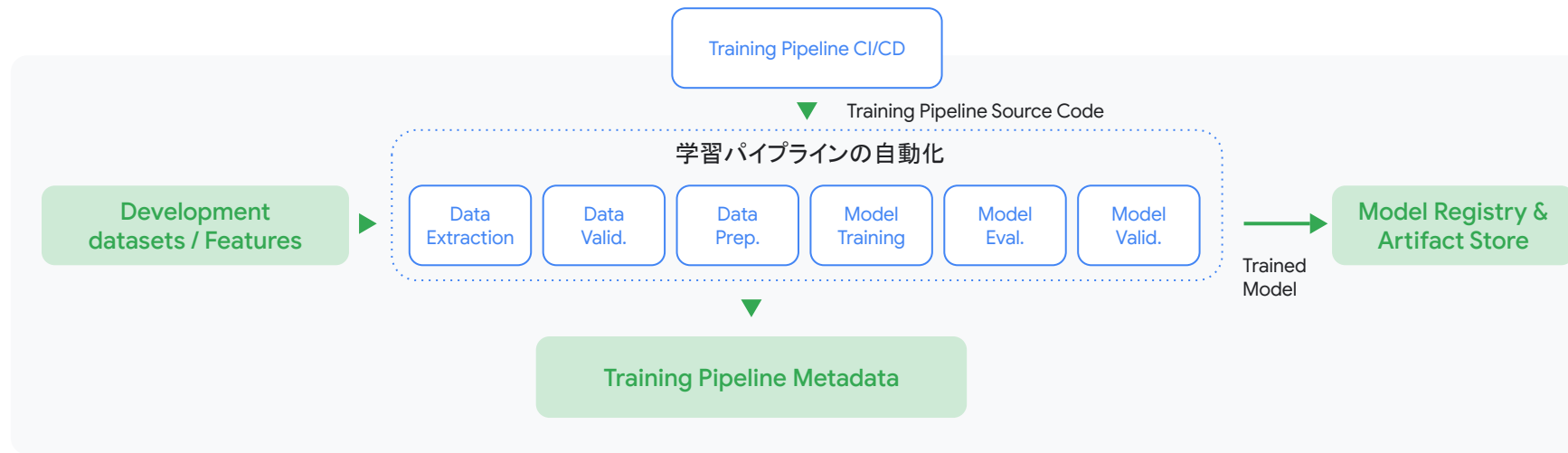
様々な ML フレームワークに対応:
TensorFlow, Pytorch,
scikit-learn, XGBoost など



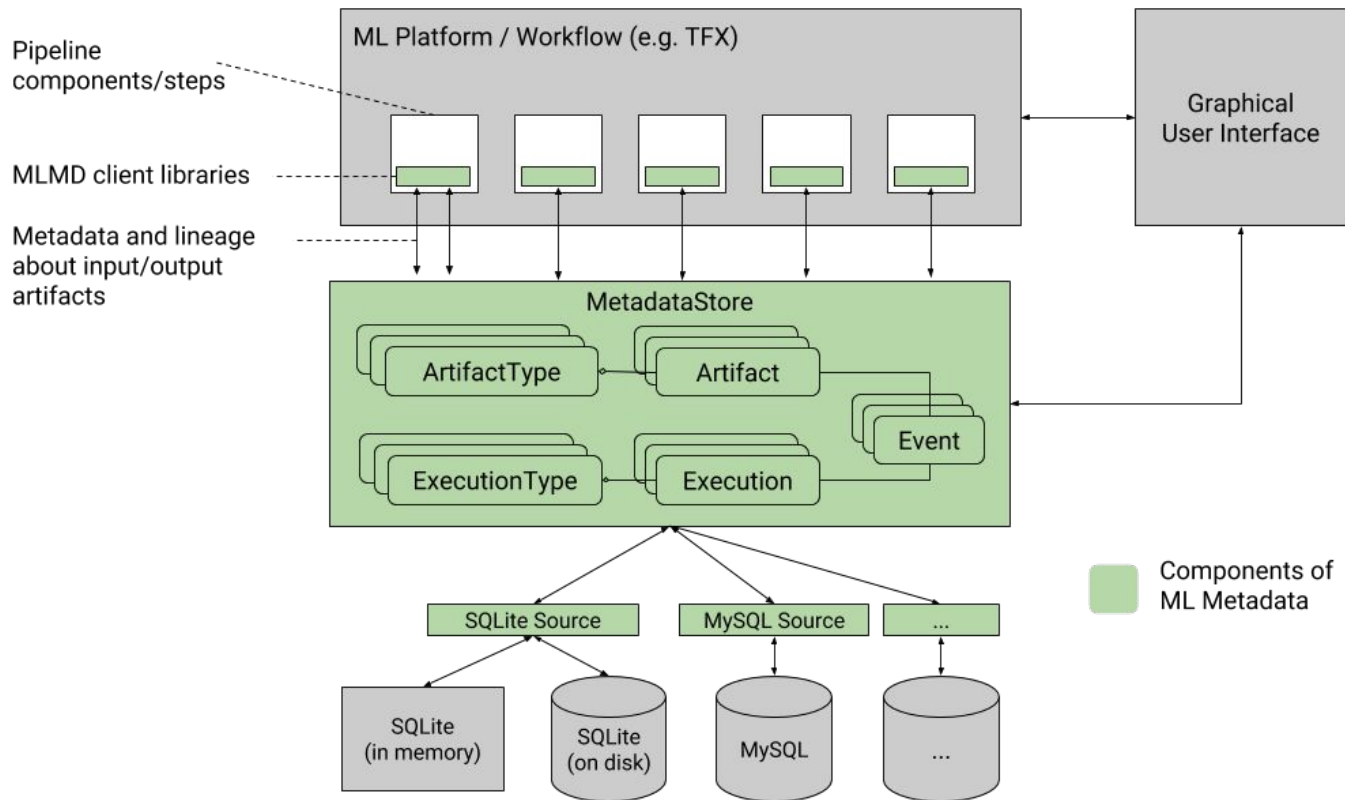
Vertex Pipelines による実験管理



Vertex Pipelines による継続学習



Vertex ML Metadata で ML ワークフローの 全てのメタデータを管理

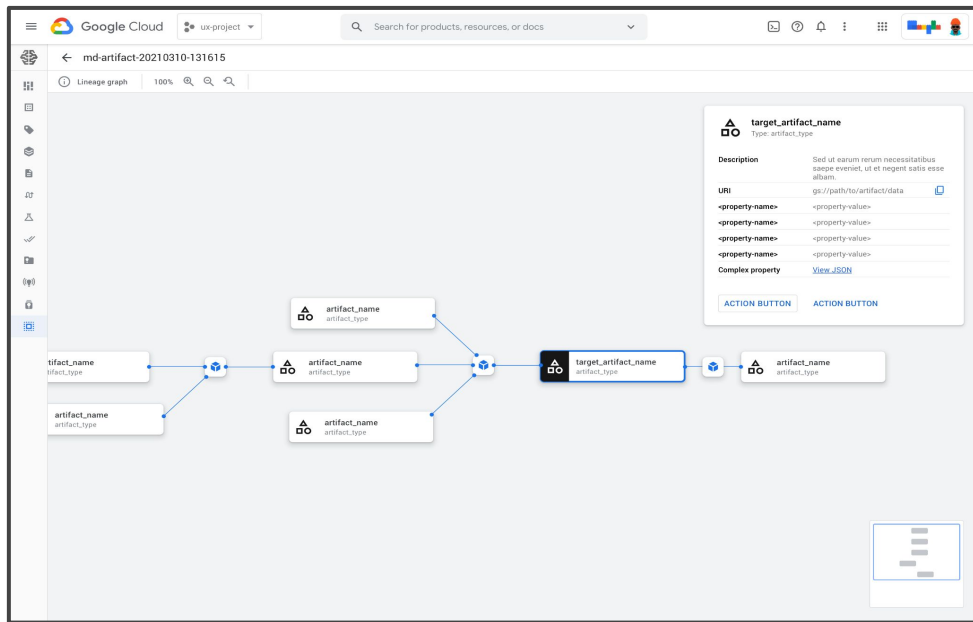


Vertex ML Metadata

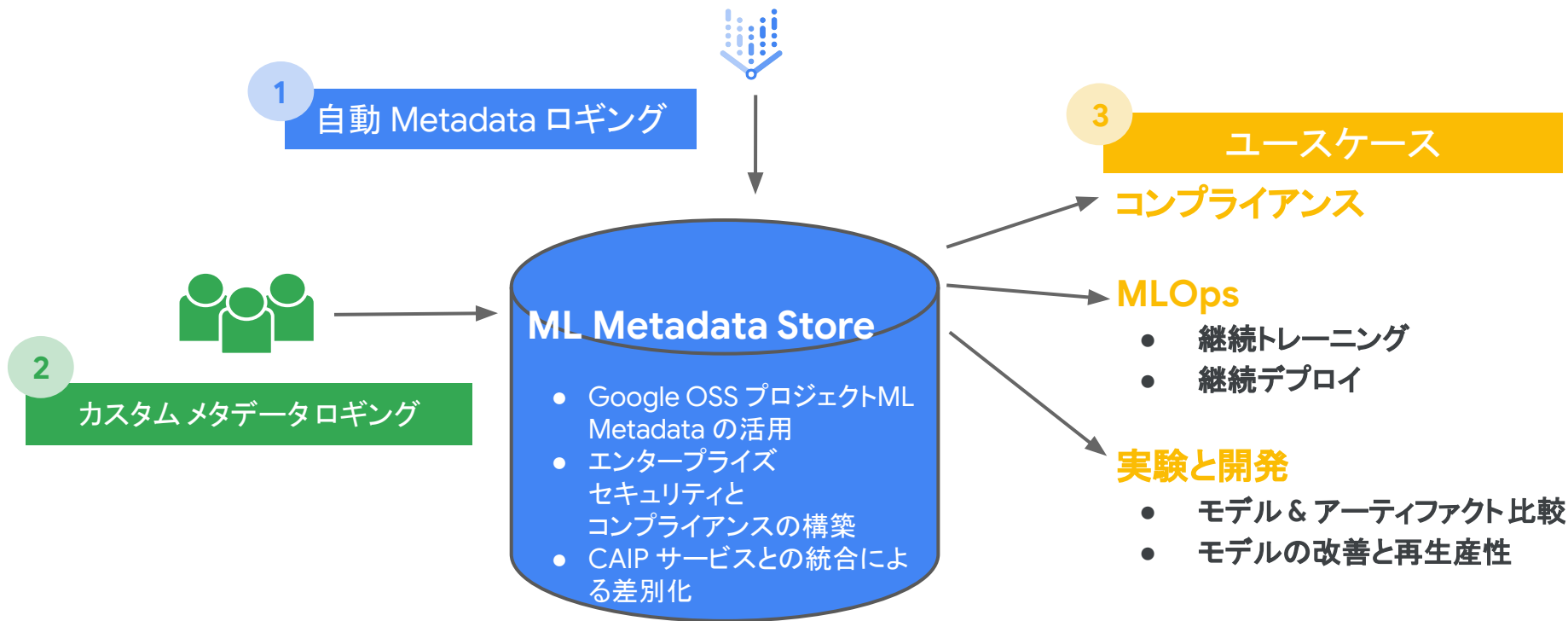
ML ワークフローの
アーティファクト、リネージ、
および実行のトラッキング

ML パイプライン内のすべての
コンポーネントとそのリネージへの
入出力を自動的にトラッキング

プロジェクトごとに
アーティファクトをグループ化して、組織
全体のデータセットと
モデルの使用状況をトラッキング



Vertex Metadata

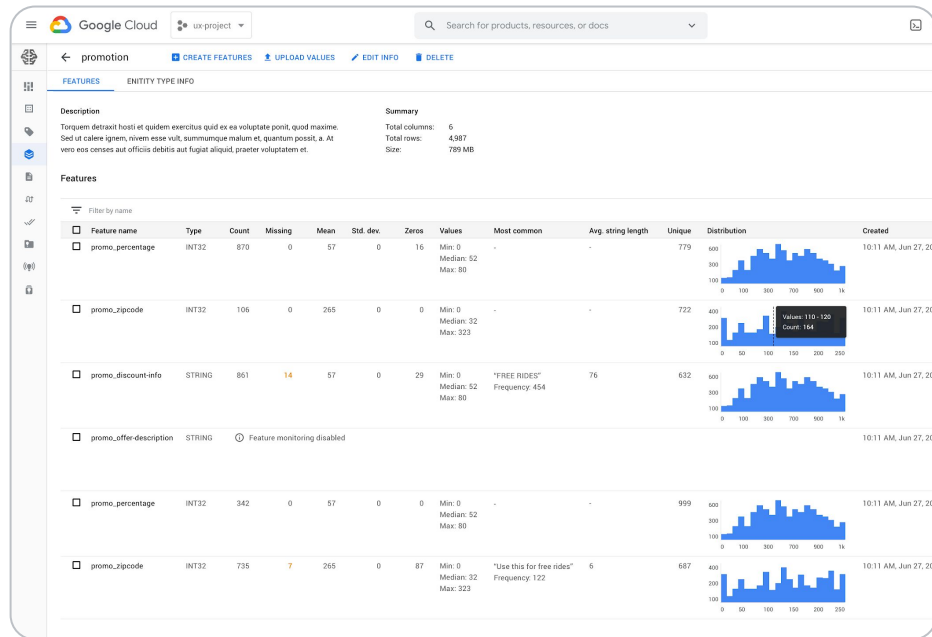


Vertex Feature Store

ユースケース間で ML 機能を
共有して再利用する

ML 機能を低レイテンシで
大規模に提供

トレーニング サービング
スキューを軽減する



特徴量にまつわる様々な課題

- 特徴量の共有と再利用
- 特徴量のオンライン サービング
- Training-Serving Skew



データサイエンティストやOps エンジニア、チーム間で
特徴量の計算方法がまちまちで、再利用できていない

ー 大手投資銀行



特徴量のオンラインサービングが難しく、多くのモデル
が使えない

ー モバイルアプリの開発会社



サービング環境を適切に実装できているか自信がな
い。学習時と推論時で同じ特徴量を異なる方法で
扱ってしまい、運用時に精度が出ない

ー 大手 e コマース企業

Vertex Feature Store

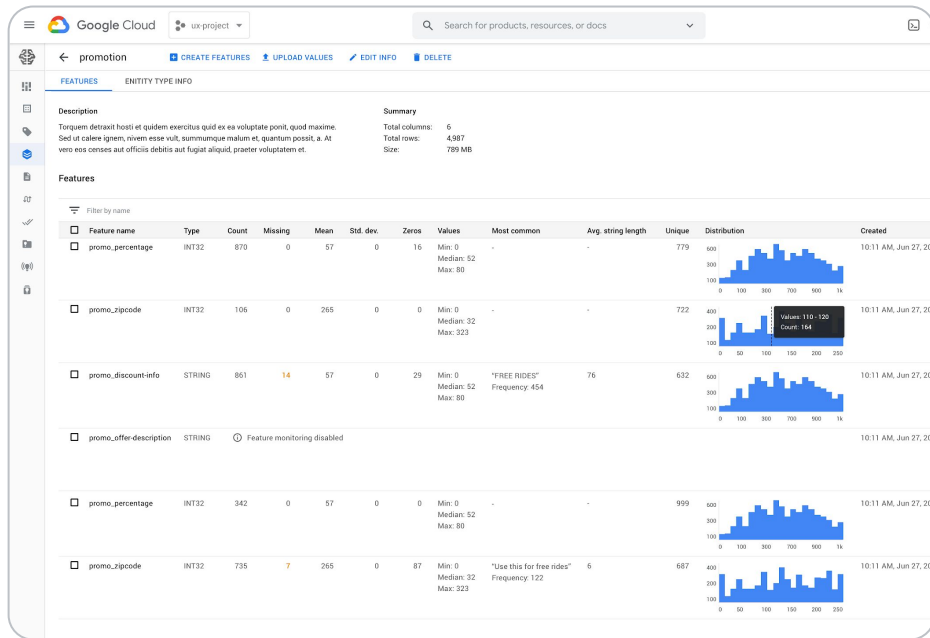
特徴量の共有と再利用

学習と推論で利用する特徴量を
使いやすいAPIで取得

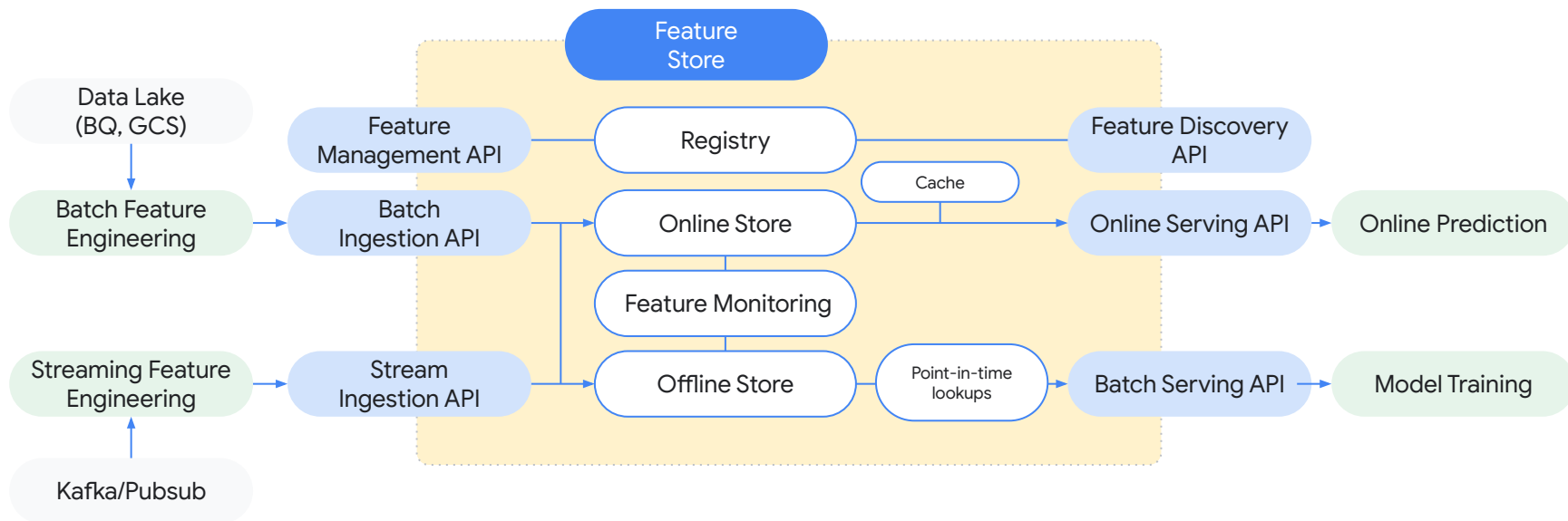
スケーラブル・低遅延で特徴量を
サービング特徴量のオンライン
サービングが容易に

Training-Serving Skew を軽減

特徴量を生成したら学習と推論で
再利用特徴量の分布変化を継続監視



Feature Store の API と SDK



Vertex Training

カスタム コンテナにより
様々な ML フレームワークに対応

複数の GPU ノードを利用する
分散学習が可能

Google の最適化ツール Vizier による
ハイパー パラメータ チューニング

Train new model

- ✓ Define your model
- ✓ Choose training method
- 3 Container settings**
- 4 Hyperparameter tuning
- 5 Compute and pricing
- 6 Inference settings

START TRAINING CANCEL

Select a pre-built container or build a custom container using ML frameworks (as well as non-ML dependencies, libraries and binaries) that are not otherwise supported. [Learn more](#)

☒ Pre-built container
View the list of [supported runtimes](#) including TensorFlow, scikit-learn and PyTorch versions

☐ Custom container
Build a custom Docker container. Must be stored in [Container Registry](#)

Pre-built container settings

Before you begin, you need to package and upload your application code and dependencies to a Cloud Storage bucket. [Learn more](#)

In order to run in a pre-built container, your code needs to be in Python 3.7

Model framework *
TensorFlow

Model framework version *
2.1

 gs:// Package location (Cloud Storage path) * BROWSE

Learn how to [package and upload](#) your application code and dependencies

[+ ADD PACKAGE](#)

Python module *

 gs:// Model output directory BROWSE

Your model artifacts and other data needed for training will be stored on Cloud Storage. You should specify a path here if you do not set an output directory in your application code or arguments.

Vertex Prediction によるエンドポイント構築

統合 エンドポイント

**Models
trained with
AutoML**



Endpoints [+ CREATE ENDPOINT](#)

Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.

To create an endpoint, you need at least one machine learning model

Region
us-central1 (Iowa) ?

Filter endpoints

☐	Endpoint	ID	Models	Region
☐	v1	8619010077484908544	1	us-central1
☐	v2	2518180694769795072	1	us-central1



**Models
trained with
custom code**

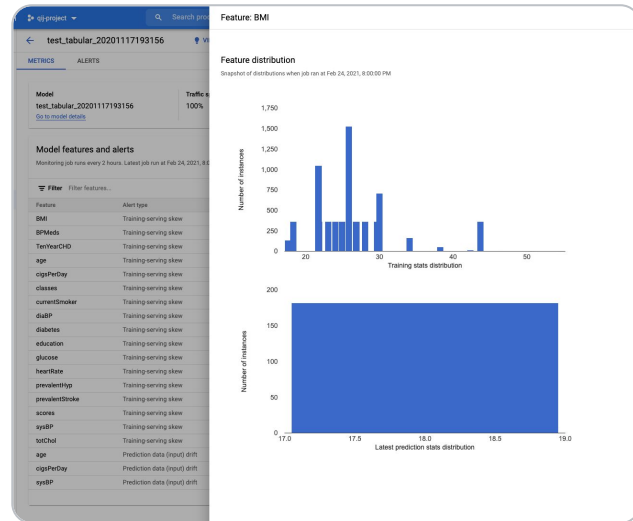
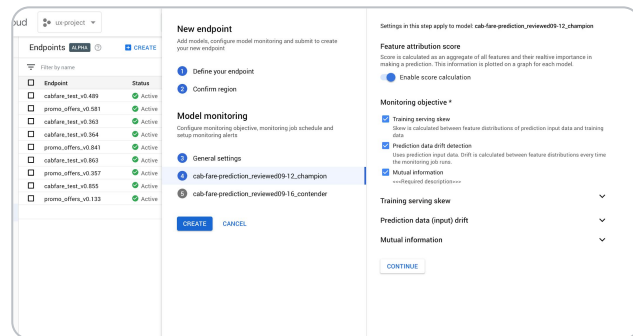
各種 ML フレームワークのモデルの推論サービスを提供
同一エンドポイントで複数モデルとバージョンを展開可能
推論サービスのインスタンス種別を選択可能

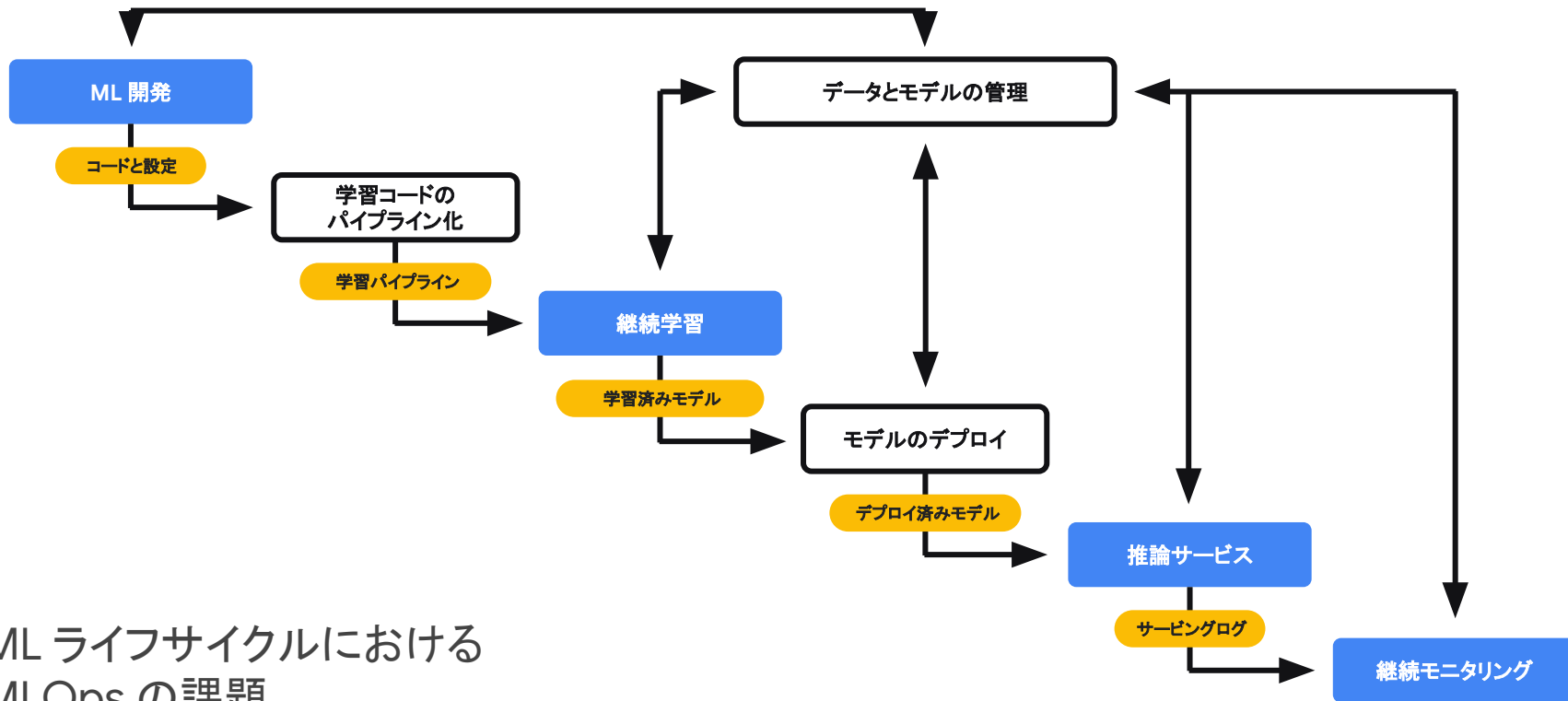
Vertex Model Monitoring

モデルのパフォーマンスが変更されたときに、
データサイエンティストと ML エンジニアに
自動的に警告

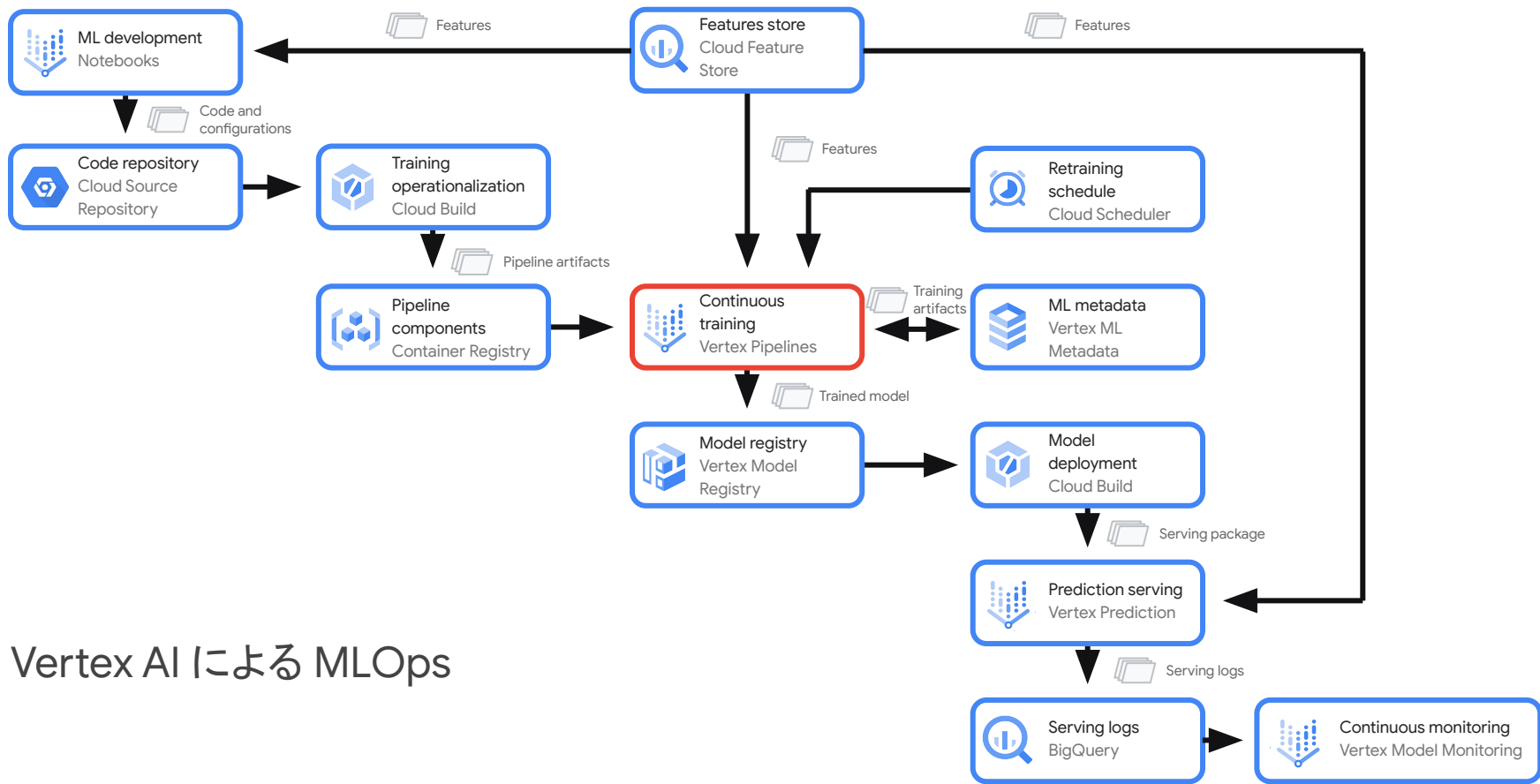
ドリフトとトレーニング サービング
スキューの検出

モデルの信頼性を高める





ML ライフサイクルにおける
MLOps の課題



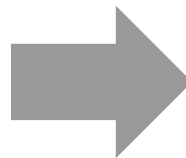
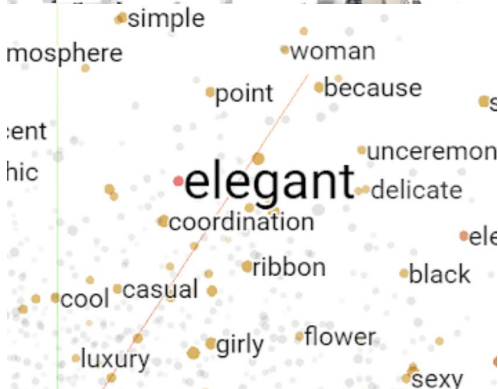
Vertex AI による MLOps

nissen, the largest mail-order service in Japan

Visual features
(shape, color, style)



Language features
(product description)



Analytics
with
Cloud AI



Computing done

This product is similar to...

1. BLK1017S0003



similarity : 0.53184

2. BLK1017S0004



similarity : 0.49584

6. BMZ0417M0009



7. AMR0118E0008



Recommendation with **insights of the sense and use cases**

CTR, CVR and session time has increased significantly

Product recommender > Learning language similarity



Tokenized sentences
Cloud Storage

Tokenize
Compute Engine

Word Embedding
Cloud Machine Learning

GPU

TensorFlow™

Product features
BigQuery

Measuring similarity
Cloud Machine Learning

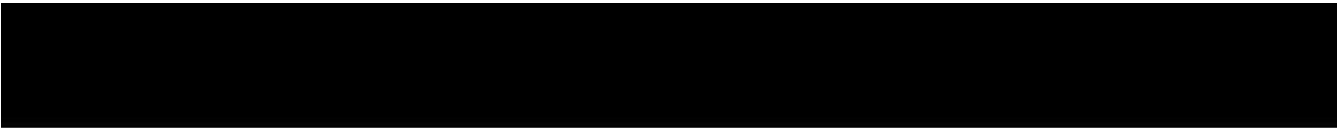
Similarity result set
BigQuery

Website

nissan

Product description

Customer Review



An Odyssey of Deep Learning underwear classification

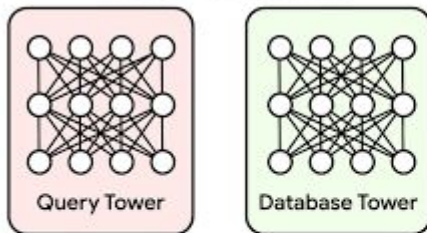


Matching Engine

Vertex Matching Engine を使用すると、

- ベクトル類似性検索を行えるため、大量のデータに対して効率的かつ正確な検索を行うことができます。
- ML モデルでは、テキストや画像などのデータ入力を、高次元ベクトルを表すエンベディングに変換します。
Matching Engine を使用すると、クエリを実行したデータに類似したデータ内で、完全一致と意味的に類似した一致(エンベディング)の両方のデータに対してクエリを実行できます。
- ベクトルをミリ秒単位で指定すると、Matching Engine は、ベクトルの大きなコーパスから最も類似したベクトルを見つけるのに役立ちます

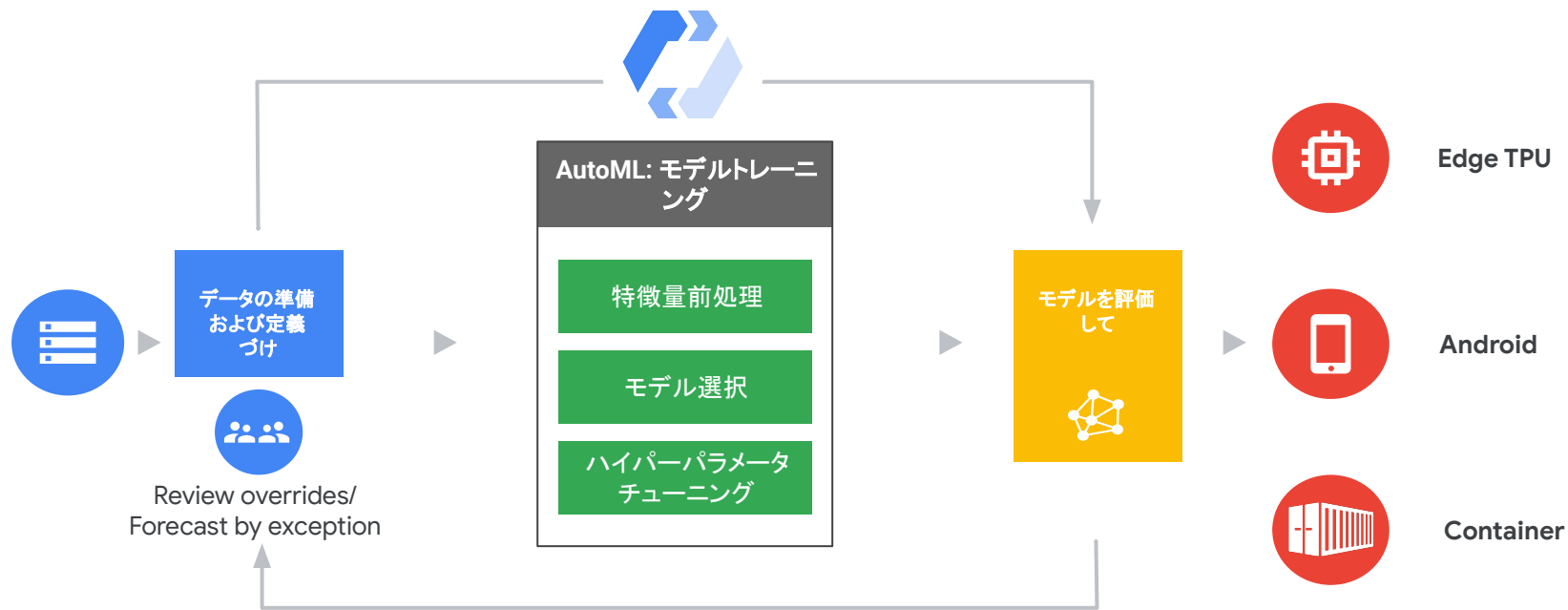
Model Architecture



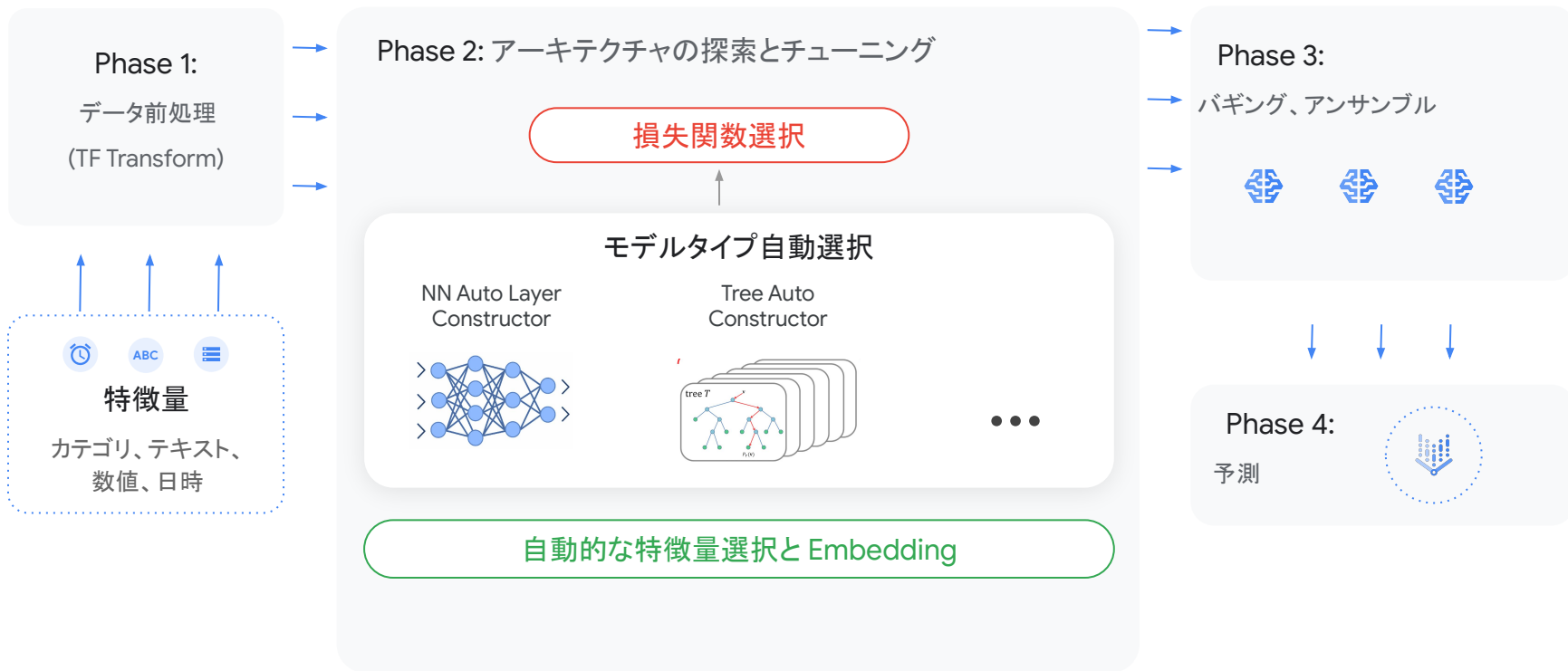
AutoML Forecast

02

AutoML Forecasting で時系列分析を自動化

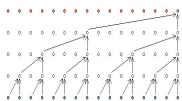


Google Brain による最新の研究成果を活用

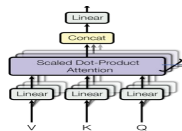


自動化されたモデルアーキテクチャ探索

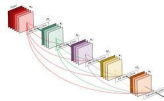
Google の最適な予測モデルアーキテクチャを評価



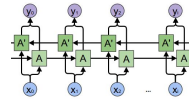
Convolutions



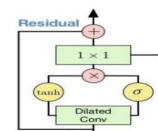
Attention



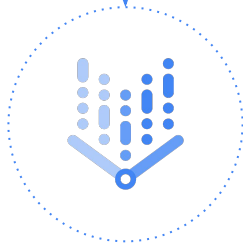
Skip Connections



LSTM



Gating



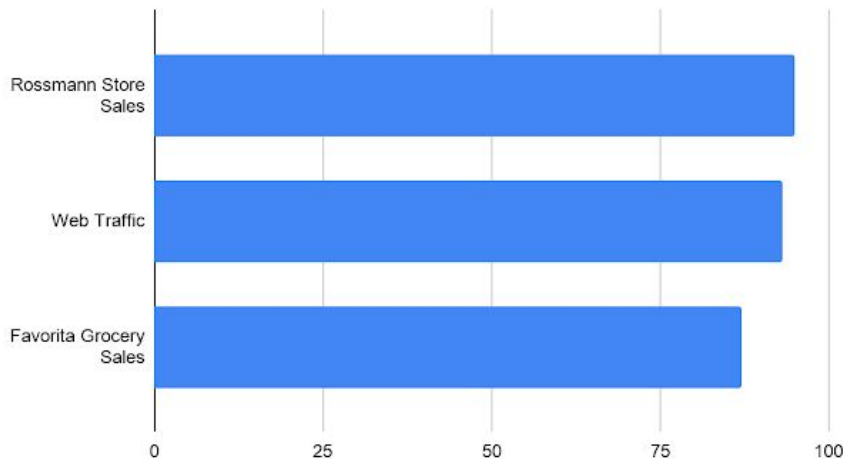
Kaggle データセットにおける成果

世界のトップ予測コンテストで上位 **2.5%** (5558 チーム中 138 チーム)

[M5: Estimate the sales of Walmart retail goods.](#)

さまざまな業界のさまざまなデータセットで一貫して上位 20% にランクイン

Better than % of Kaggle Participants



小売における基本的な用途: 需要予測

商品メタデータ



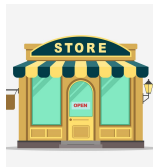
- サマードレス
- 模様
- 長袖Vネック
- 55%リネン、綿 45%
- 洗濯機洗い可能

需要関連要素



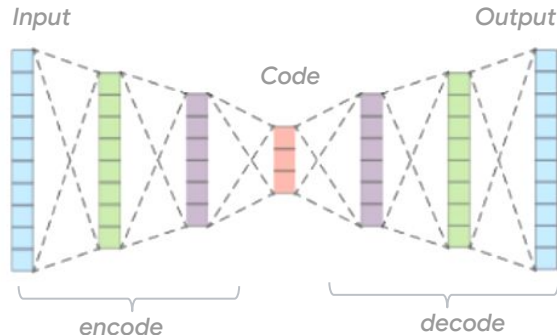
- 販売
- 価格
- 競合他社の価格
- プロモーション/イベント
- 休日

店舗データ



- 店舗の説明: 大、小、専門
- 店の所在地
- アクセス

AutoML Forecast



機械学習モデルが、豊富なメタデータ、製品間の関係、価格設定、競争、製品ライフサイクルの相乗効果を理解

長期的な予測区間 12-16 ヶ月

- シーズン前の計画
- 購入/注文計画
- 新規およびコールドスタートアイテム

短期的な予測区間 0-8 週

- シーズン中の計画
- 補充、在庫
- 価格設定、割り当て
- SKU レベル