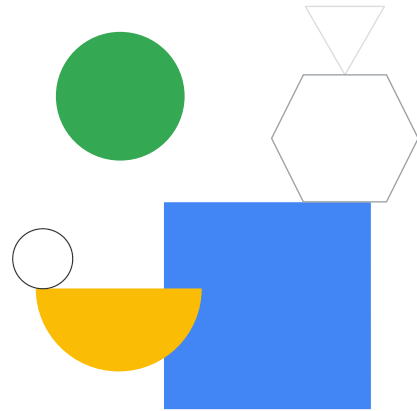




Production ML Pipelines



In a previous module, we leveraged pre-trained ML APIs to process natural text. These are great options for seeing if your use case can just use a model that's already created and trained on Google's data. But, you may want a more tailored model trained on your own data. For that we will need a custom model. Let's talk about the different ways of building custom models.



Module agenda



- 01 Ways to do ML on Google Cloud
- 02 Vertex AI Pipelines
- 03 AI Hub

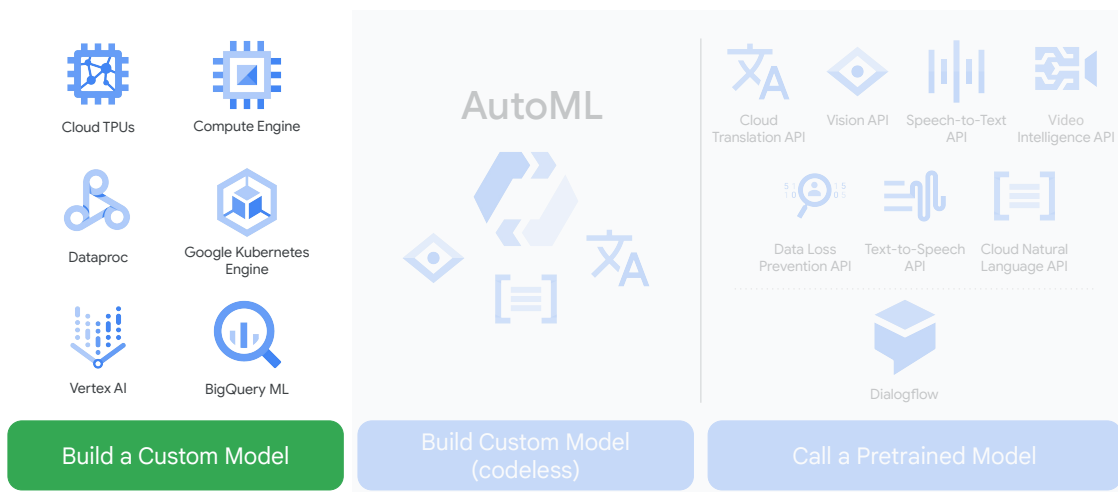
First, we will provide an overview of ways to do ML on Google Cloud. Then, we will talk about Vertex AI Pipelines, for deploying machine learning models in a production environment. Finally, we will discuss AI Hub, a repository of machine learning resources which can be made publicly available or available for only certain users.



Ways to do ML on Google Cloud

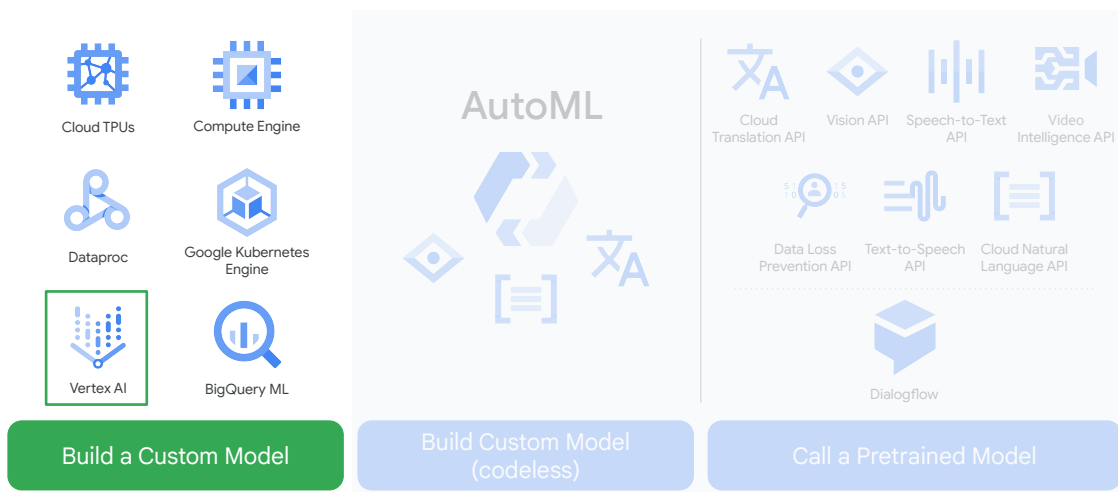
You've already learned that there are three ways you can do machine learning on Google Cloud.

Create and deploy custom models with Vertex AI



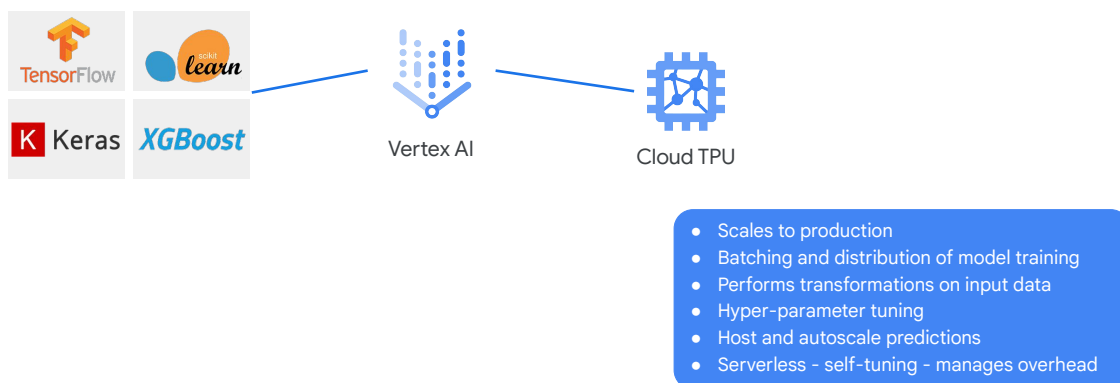
The pretrained models on the right have already been discussed. Now, we're going to visit the other side of the spectrum and build your own custom model and productionalize it on Google Cloud. There are a few ways of doing custom model development, training, and serving.

Create and deploy custom models with Vertex AI



Let's discuss Vertex AI.

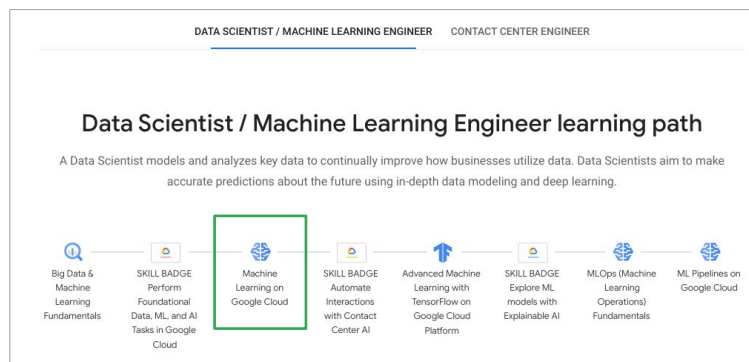
Vertex AI is a fully managed service for custom machine learning models



What is Vertex AI exactly? It's a fully managed service for custom machine learning models, both training and serving predictions. It can scale from the experimentation stage all the way to production. You can also, using the features of TensorFlow, include transformations on input data and perform hyperparameter tuning to choose the best model for your case. You can deploy your models to Vertex AI to serve predictions, which will autoscale to the demands of your clients.

Essentially, Vertex AI is the engine behind doing machine learning at scale on Google Cloud. A data scientist can train and deploy production models from Notebooks with just a few commands.

In this course, we don't cover writing TensorFlow models, only ways to operationalize them



Google Cloud Training - Machine Learning and AI

Since we're using Vertex AI, we will often be thinking about using TensorFlow models. However, this isn't the course to dive into the details of TensorFlow. You can learn more about this in the Machine Learning on Google Cloud course, which is part of the Machine Learning and AI learning path for Data Scientists and Machine Learning Engineers.



Vertex AI Pipelines

Where do Data Engineers come into the picture? Don't forget Data Engineers build data pipelines, and machine learning pipelines are no different. If we want to have a flexible pipeline for all stages of machine learning, Vertex AI Pipelines are a great option.

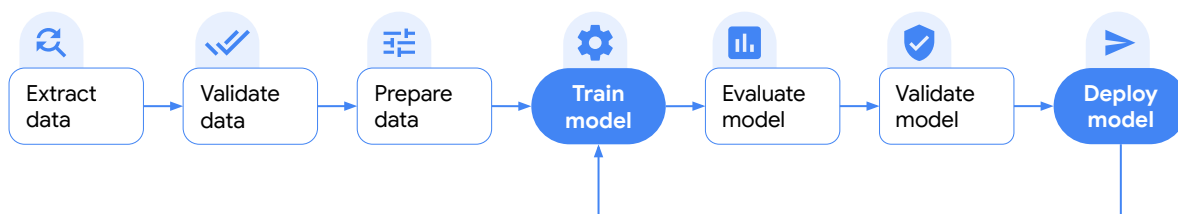
What are ML Pipelines?

What are ML Pipelines?

Building machine learning models require complex multi-step workflows. When building a model, you may have to clean and transform data, create features, train multiple models, and evaluate those models. Managing and executing these workflows can be difficult -- especially if you want to ensure that they are run in a reproducible, auditable, cost-effective, and scalable way.

Pipelines automate the training and deployment of models

A pipeline is a way of modeling a workflow as a set of connected steps. Each step takes as inputs the outputs of previous steps, performs some additional computations, and produces outputs that can be utilized by future components.

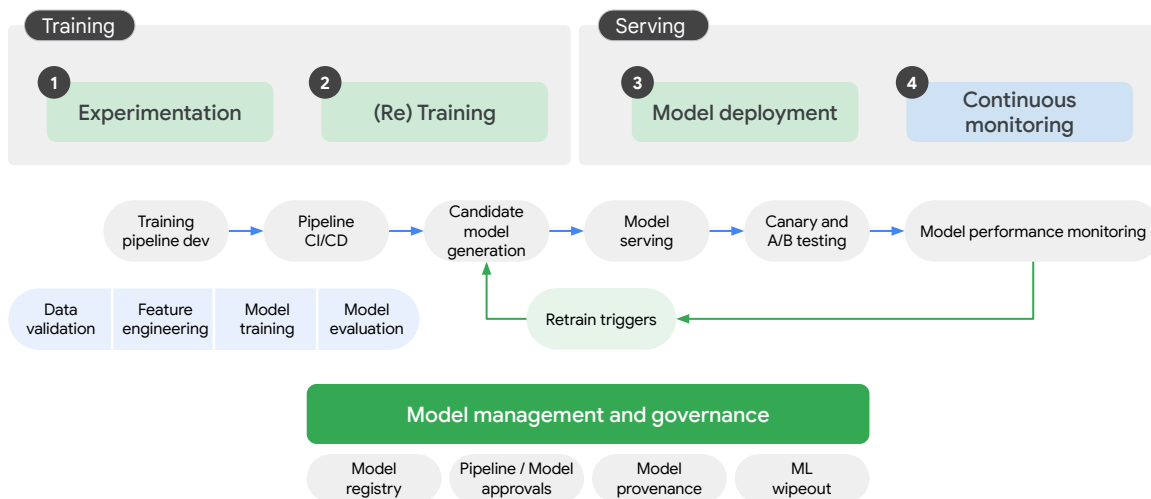


For example, a simple ML pipeline might do the following:

- Load a dataset from a comma-separated value file.
- Analyze the dataset to identify and remove outliers.
- Split the cleaned dataset into a training and evaluation dataset.
- Train a model on the training dataset.
- And evaluate the model against the evaluation dataset.

In this example, each step (except for the first one) takes as input the output of a previous step; and produces an output that can be utilized by a subsequent step.

Pipelines are the backbone of production ML systems



Pipelines form the core of production ML systems and are used in all levels of development, testing, and deployment.

Pipelines allow the completion of routine steps to allow experimentation to move more rapidly.

They are critical in training and re-training because they allow the isolation of steps and stages to pinpoint areas for optimization.

Pipelines are also the most durable way to deploy models in production particularly at scale.

Finally, pipelines provide a variety of control points for continuous monitoring and optimization of production assets.

Pipelines product portfolio



Kubeflow

Kubeflow Pipelines

- Kubernetes-native.
- Open source.
- The industry standard for running ML Pipelines.



AI Platform Pipelines - Hosted ^{Beta}

- Kubeflow pipelines running on Google Cloud.
- Optimized for GKE.
- Integrated with Google Cloud services.

Vertex Pipelines - Managed ^{PREVIEW}

- Fully managed and serverless.
- Allows users to focus on building their pipelines, scale easily, and pay only for the resources they use.

Due to the size of the challenge to enable ML pipelines, numerous solutions have been explored and expanded.

Kubeflow Pipelines is a Kubernetes-native, open source product, that has grown into the industry standard for running ML pipelines over the years.

The next solution was **AI Platform Pipelines**, a GKE-optimized service aimed at making it easy to deploy Kubeflow Pipelines to Google Cloud resources.

With **Vertex Pipelines**, the moving of resource management away from users continued, with a reduction in the day-to-day inconvenience of managing configuration files. For example, you are no longer required to create a dedicated Kubernetes cluster using GKE to run your Pipelines. Instead, Vertex AI manages the Kubernetes clusters and the Pods running on them behind the scenes.

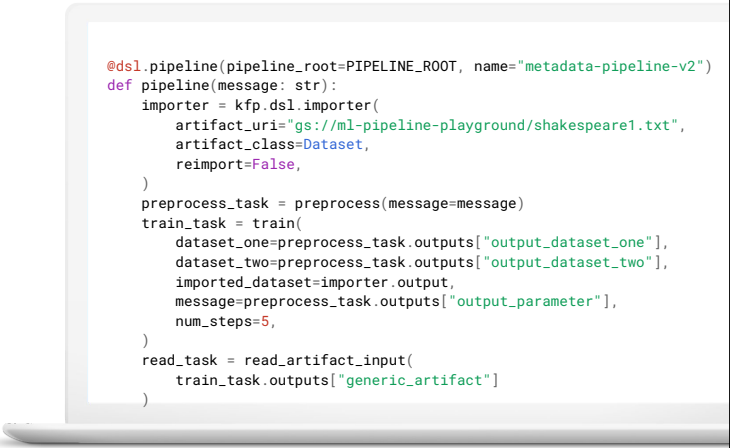
Write your pipeline

Easy to use Python SDKs

Build pipelines using Data Scientist friendly SDKs like TensorFlow Extended and Kubeflow Pipelines.

Rich, scalable pre-built components

We provide a rich set of pre-built components for common ML tasks, which leverage Google Cloud services.

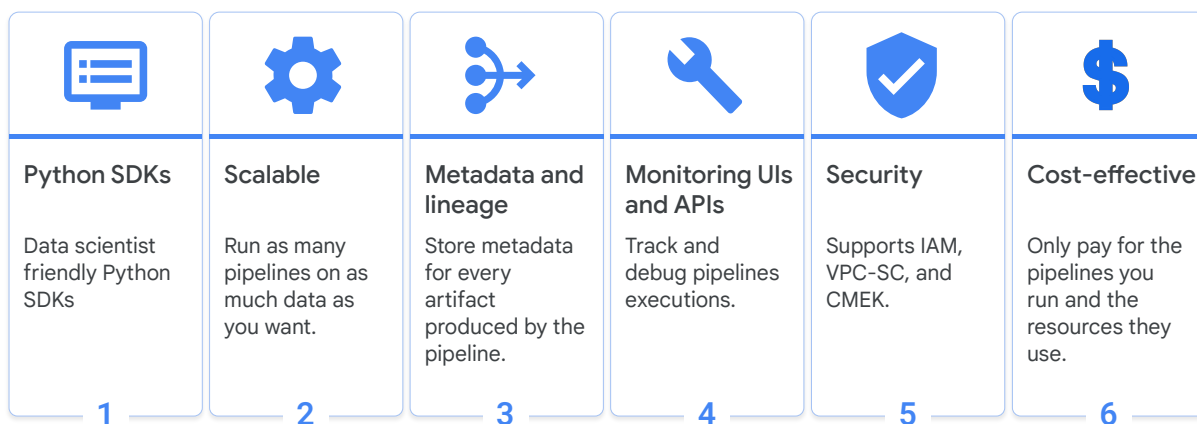


```
@dsl.pipeline(pipeline_root=PIPELINE_ROOT, name="metadata-pipeline-v2")
def pipeline(message: str):
    importer = kfp.dsl.importer(
        artifact_uri="gs://ml-pipeline-playground/shakespeare1.txt",
        artifact_class=Dataset,
        reimport=False,
    )
    preprocess_task = preprocess(message=message)
    train_task = train(
        dataset_one=preprocess_task.outputs["output_dataset_one"],
        dataset_two=preprocess_task.outputs["output_dataset_two"],
        imported_dataset=importer.output,
        message=preprocess_task.outputs["output_parameter"],
        num_steps=5,
    )
    read_task = read_artifact_input(
        train_task.outputs["generic_artifact"]
    )
```

Pipelines can be easily developed using flexible Python SDKs.

This allows rapid iteration and faster development to deployment cycles by likely leveraging skills that exist in your organization.

Key capabilities



So to summarize, managed pipelines:

1. Are built with easy-to-use, data scientist-friendly, Python SDKs.
2. Are scalable because they leverage Google Cloud's best of breed managed services.
3. Automatically store metadata for every artifact produced by the pipeline.
4. Have robust tools for managing pipelines.
5. Are secure.
6. Will be extremely cost-effective because resources are allocated on a per pipeline-step process.

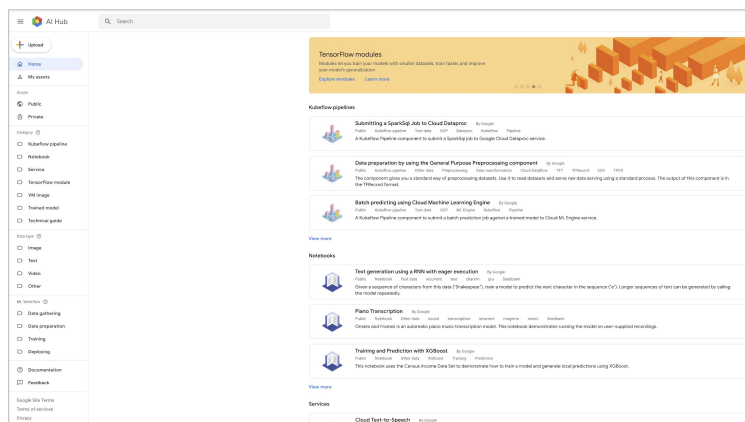


AI Hub

Vertex AI pipelines can be packaged and shared with other users. This leads us to a discussion of AI Hub.

AI Hub is a repository for AI assets

Don't reinvent the wheel! Find and deploy ML pipelines.



Google Cloud

AI Hub is a repository for ML components. Don't reinvent the wheel! Avoid building some component when someone else has already built it, and most likely, has already optimized it. You can find and deploy not just containerized applications for machine learning, but full ML pipelines on AI Hub.

AI Hub stores various asset types

- Kubeflow pipelines and components
- Jupyter notebooks
- TensorFlow modules
- Trained models
- Services
- VM images

What asset types can we find on AI Hub? Among the assets stored on AI Hub are entire Kubeflow pipelines, Jupyter notebooks, TensorFlow modules, fully trained models, services, and VM images.

This is what a typical asset looks like

The screenshot displays the Google Cloud AI Hub interface for a pipeline asset titled "Deploying a trained model to Cloud Machine Learning Engine". The interface includes a sidebar with filters for Scope (Public), Version (1), Category (Kubeflow pipeline), Publisher (Google), Data type (Text), and Labels (GCP, ML Engine, Kubeflow, Pipeline). The main content area provides documentation, intended use, runtime arguments, and output information.

Documentation

Deploying a trained model to Cloud Machine Learning Engine

A Kubeflow Pipeline component to deploy a trained model from a Cloud Storage path to a Cloud Machine Learning Engine service.

Intended use

Use the component to deploy a trained model to Cloud Machine Learning Engine service. The deployed model can serve online or batch predictions in a KFP pipeline.

Runtime arguments:

Name	Description	Type	Optional	Default
model_uri	The Cloud Storage URI which contains a model file. The commonly used TF model search path (export/exporter) will be used.	GCSPath	No	
project_id	The ID of the parent project of the serving model.	GCPProjectID	No	
model_id	The user-specified name of the model. If it is not provided, the operation uses a random name.	String	Yes	
version_id	The user-specified name of the version. If it is not provided, the operation uses a random name.	String	Yes	
runtime_version	The Cloud Machine Learning Engine's runtime version to use for this deployment. If it is not set, the Cloud ML Engine uses the default stable version, 1.0.	String	Yes	
python_version	The version of Python used in the prediction. If it is not set, the default version is 2.7. Python 3.5 is available when the runtime_version is set to 1.4 and above. Python 2.7 works with all supported runtime versions.	String	Yes	
version	The JSON payload of the new Version.	Dict	Yes	
replace_existing_version	A Boolean flag indicates whether to replace existing version in case of conflict.	Bool	Yes	False
set_default	A Boolean flag indicates whether to set the new version as default version in the model.	Bool	Yes	False
wait_interval	A time-interval to wait for in case the operation has a long run time.	Integer	Yes	30

Output:

Use this asset

[Download](#)

Create a Kubeflow Cluster to use this pipeline
Learn more about how to use pipelines

[Feedback](#)

[f](#) [in](#)

One-click deployment of ML pipelines via Kubeflow on Google Cloud as platform for AI, or on premise.

Here you see what a typical asset looks like. You can see information about the pipeline, such as inputs and outputs, and download options.

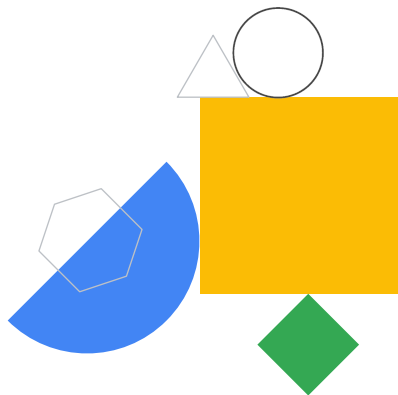
Assets on AI Hub are collected in two scopes: public assets and restricted assets

- Public scope are available to all AI Hub users.
- Restricted scope contains AI components that you have uploaded and assets that have been shared with you.

The assets on AI Hub are collected into two scopes: public assets and restricted assets. Public assets are available to all AI Hub users. Restricted scope assets contain AI components you have uploaded and those that have been shared with you. For example, you could have assets only available to people within your organization or teams.

Lab Intro

Running Pipelines on Vertex AI



To get a better understanding of how Vertex AI Pipelines works, let's dive into a lab.

In this lab, you learn how to utilize Vertex AI Pipelines to execute a simple Kubeflow Pipeline SDK-derived ML Pipeline.

To start, you set up the project environment. Then you configure and inspect the Pipeline code. Lastly, you execute the AI Pipeline.

Summary

- Use ML on Google Cloud using either:
 - Vertex AI (your model, your data)
 - AutoML (our models, your data)
- Use Vertex AI Pipelines to deploy end-to-end ML pipelines.
- Don't reinvent the wheel for your ML pipeline! Leverage pipelines on AI Hub.

To summarize:

- Google Cloud has several options to suit your machine-learning needs. Depending on the time and resources you have available, you have the option to use Vertex AI or AutoML.
- You can use Vertex AI Pipelines to deploy end-to-end ML pipelines.
- And remember don't reinvent the wheel for your ML pipeline, leverage pipelines on AI Hub.