Regional endpoints

To use online prediction, you can interact with the AI Platform Training and Prediction API through its global endpoint (`ml.googleapis.com`) or through one of its regional endpoints (***`REGION`***`-ml.googleapis.com`). Using a regional endpoint for online prediction provides additional protection for your model against outages in other regions, because it isolates your model and version resources from other regions.

AI Platform Prediction currently supports the following regional endpoints:

- `us-central1`
- `europe-west4`
- `asia-east1`

This guide compares the benefits and limitations of using regional endpoints versus the global endpoint. The guide also walks through using a regional endpoint for online prediction.

Understanding regional endpoints

Regional endpoints have several key differences from the global endpoint:

- *Regional endpoints only support Compute Engine (N1) machine types*. You cannot use legacy (MLS1) machine types on regional endpoints. This means that all the benefits and limitations of using Compute Engine (N1) machine types (/ai-platform/prediction/docs/machine-types-online-prediction#differences) apply. For example, you can use GPUs on regional endpoints, but you cannot currently enable stream (console) logging.

  To use a Compute Engine (N1) machine type, you must use a regional endpoint.

- *Regional endpoints only support online prediction and AI Explanations* (/ai-platform/prediction/docs/ai-explanations/overview). Models deployed to regional endpoints do not support batch prediction (/ai-platform/prediction/docs/batch-predict).

  AI Platform Prediction shares the AI Platform Training and Prediction API with AI Platform Training (/ai-platform/training/docs) and AI Platform Optimizer (/ai-platform/optimizer/docs).

> Note that regional endpoints do not currently support AI Platform Training. Only the `us-central1` endpoint supports AI Platform Optimizer.
>
> See the API reference (/ai-platform/prediction/docs/reference/rest#service-endpoint) for more details about which API methods are available on which endpoints.

AI Platform Prediction resource names are unique for your Google Cloud project on any given endpoint, but they can be duplicated on various endpoints. For example, you can create a model named "hello-world" on the `europe-west4` endpoint and another model named "hello-world" on the `us-central1` endpoint.

When you list models on a regional endpoint, you only see models created on that endpoint. Similarly, when you list models on the global endpoint, you only see models created on the global endpoint.

Regional endpoints versus global endpoint regions

When you create a model resource (/ai-platform/prediction/docs/deploying-models#create_a_model_resource) on the global endpoint, you can specify a region for your model. When you create versions within this model and serve predictions, the prediction nodes (/ai-platform/prediction/docs/overview#node-allocation) run in the specified region.

When you use a regional endpoint, AI Platform Prediction runs your prediction nodes in the endpoint's region. However, in this case AI Platform Prediction provides additional isolation by running all AI Platform Prediction infrastructure in that region.

For example, if you use the `us-east1` region on the global endpoint, your prediction nodes run in `us-east1`. But the AI Platform Prediction infrastructure managing your resources (routing requests; handling model and version creation, updates, and deletion; etc.) does not necessarily run in `us-east1`. On the other hand, if you use the `europe-west4` regional endpoint, your prediction nodes and all AI Platform Prediction infrastructure run in `europe-west4`.

Using regional endpoints

To use a regional endpoint, you must first create a model on the regional endpoint. Then perform all actions related to that model (like creating a model version and sending prediction requests) on the same endpoint.

If you are using the Google Cloud Console, make sure to select the **Use regional endpoint** checkbox when you create your model. Perform all other Cloud Console actions like you would on the global endpoint.

If you are using the `gcloud` *command-line tool*, `--region` flag to the region of your endpoint on every command that interacts with your model and its child resources. This includes the following:

- Every command in the `gcloud ai-platform models` command group (/sdk/gcloud/reference/ai-platform/models).
- Every command in the `gcloud ai-platform versions` command group (/sdk/gcloud/reference/ai-platform/versions).
- Every command in the `gcloud ai-platform operations` command group (/sdk/gcloud/reference/ai-platform/operations) when interacting with long-running operations associated with a version of the model.
- The `gcloud ai-platform predict` command (/sdk/gcloud/reference/ai-platform/predict).
- The `gcloud ai-platform explain` command (/sdk/gcloud/reference/ai-platform/explain).

Do not confuse the `--region` flag with the `--regions` flag of the `gcloud ai-platform models create`
and. The latter is used to specify a region when you create a model on the global endpoint (#global-endpoint-r
ot permitted when you use a regional endpoint.

If you are interacting directly with the AI Platform Training and Prediction API (for example, by using the Google APIs Client Library for Python (/ai-platform/prediction/docs/python-client-library)), make all API requests like you would to the global endpoint, but use the regional endpoint instead. See the API reference (/ai-platform/prediction/docs/reference/rest#service-endpoint) for more details about which API methods are available on regional endpoints.

The following examples demonstrate how to use a regional endpoint to create a model, create a version, and send an online prediction request. To use the examples, replace ***REGION*** wherever it appears with one of the regions where regional endpoints are available:

- `us-central1`
- `europe-west4`
- `asia-east1`

Creating a model

Cloud Consolegcloud (#gcloud)Python (#python)

1. In the Cloud Console, go to the **Create model** page and select your Google Cloud project:

   Go to the Create model page (https://console.cloud.google.com/ai-platform/create-model?projec

2. Name your model, select the **Use regional endpoint** checkbox, and select the region of the endpoint that you want to use from the **Region** drop-down list.

3. Click the **Create** button.

Learn more about creating a model (/ai-platform/prediction/docs/deploying-models#create_a_model_resource).

Creating a model version

This example assumes that you have already uploaded compatible model artifacts (/ai-platform/prediction/docs/exporting-for-prediction) to Cloud Storage.

Cloud Consolegcloud (#gcloud)Python (#python)

Using the model that you created in the previous section, follow the guide to creating a model version (/ai-platform/prediction/docs/deploying-models#create_a_model_version) in the Cloud Console.

Learn more about creating a model version (/ai-platform/prediction/docs/deploying-models#create_a_model_version).

Sending an online prediction request

Cloud Consolegcloud (#gcloud)Python (#python)

1. In the Cloud Console, go to the **Models** page:

   Go to the Models page (https://console.cloud.google.com/ai-platform/models)

2. In the **Region** drop-down list, select the region of the endpoint that your model uses. Click the name of the model that you created in a previous section to navigate to its **Model Details** page.
3. Click the name of the version that you created in a previous section to navigate to its **Version Details** page.
4. Click the **Test & use** tab. Enter one or more instances of input data and click the **Test** button to send an online prediction request.

Learn more about getting online predictions (/ai-platform/prediction/docs/online-predict).

What's next

- See differences in regional availability (/ai-platform/prediction/docs/regions) for regional endpoints and the global endpoint. This includes differences in GPU availability.
- Learn more about Compute Engine (N1) machine types (/ai-platform/prediction/docs/machine-types-online-prediction), which are required for regional endpoints.
- Read about other additional options that you can configure when you create models and versions (/ai-platform/prediction/docs/deploying-models).

Last updated 2020-08-19 UTC.