Cloud AutoML Vision Object Detection

Deploying your model

Initial model deployment

After you have created (trained) a model, you must deploy the model before you can make online (or synchronous) calls to it.

You can now also update model deployment if you need additional online prediction capacity.

Deploying an Object Detection model incurs charges. For more information, see the pricing page (https://cloud.google.com/vision/automl/pricing).

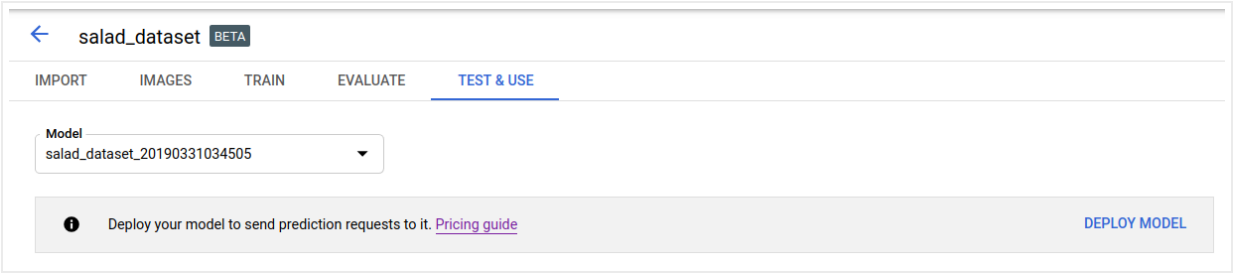
Node count will be subject to quota listed at the quota page; **by default, a user can get up to 10 nodes**.

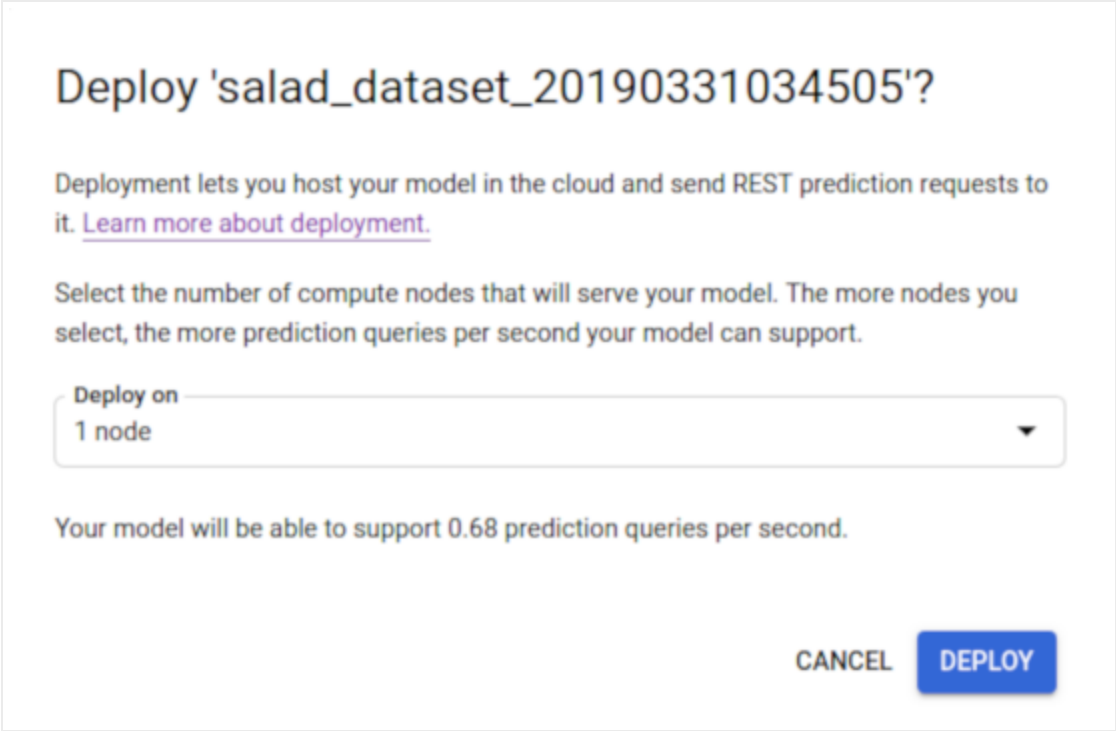
- The maximum lifespan for a custom model is 18 months as of the GA release. You must create and train a new model to continue classifying content after that amount of time.

WEB UI REST & CMD LINE MORE

1. Open the Cloud AutoML Vision Object Detection UI (https://console.cloud.google.com/vision) and select the **Models** tab (with lightbulb icon) in the left navigation bar to display the available models.

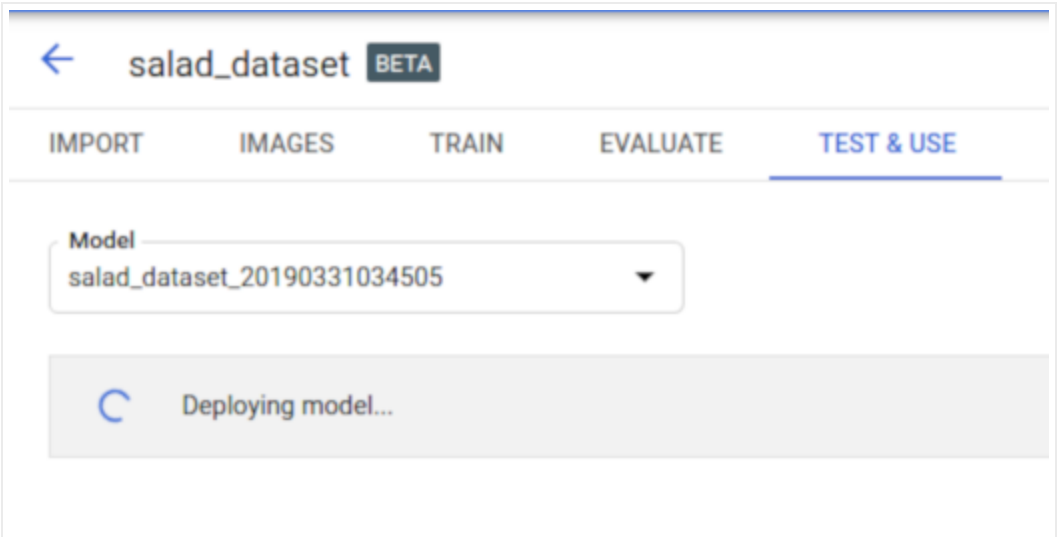
   To view the models for a different project, select the project from the drop-down list in the upper right of the title bar.
2. Select the row for the model you want to use to label your images.
3. Select the **Test & Use** tab just below the title bar.
4. Select **Deploy model** from the banner beneath your model name to open the model deployment option window.



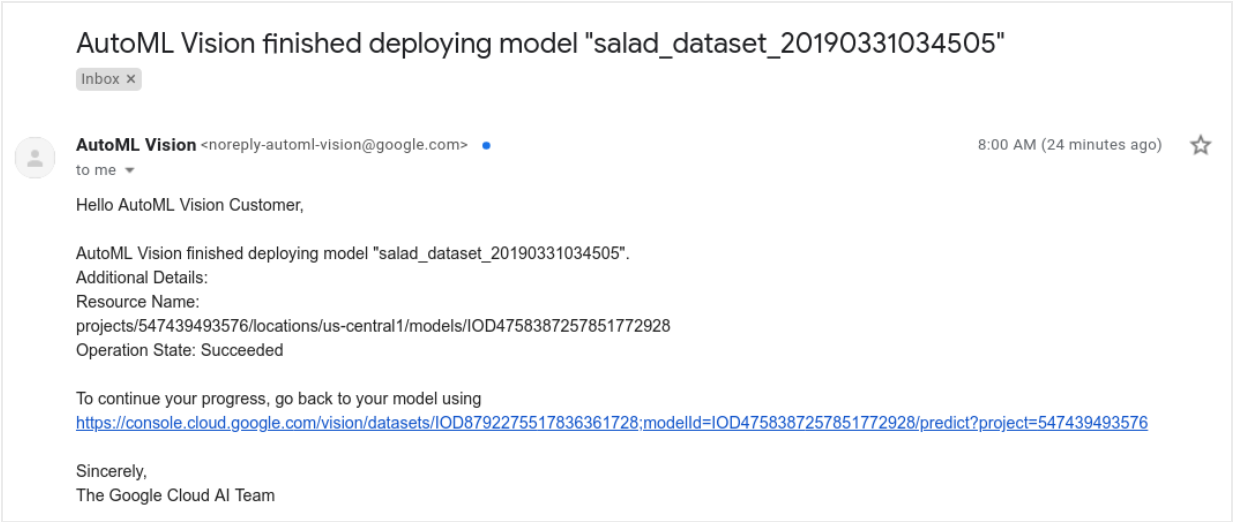


In this window you can select the number of nodes to deploy on and view the available prediction queries per second (QPS).

5. Select **Deploy** to begin model deployment.



6. You will receive an email when model deployment has completed.



AutoML Vision finished deploying model "salad_dataset_20190331034505"

Inbox ×

AutoML Vision <noreply-automl-vision@google.com> 8:00 AM (24 minutes ago)

to me

Hello AutoML Vision Customer,

AutoML Vision finished deploying model "salad_dataset_20190331034505".
Additional Details:
Resource Name:
projects/547439493576/locations/us-central1/models/IOD4758387257851772928
Operation State: Succeeded

To continue your progress, go back to your model using
https://console.cloud.google.com/vision/datasets/IOD8792275517836361728;modelId=IOD4758387257851772928/predict?project=547439493576

Sincerely,
The Google Cloud AI Team

Update a model's node number

Once you have a trained deployed model you can update the number of nodes the model is deployed on to respond to your specific amount of traffic. For example, if you experience a higher amount of queries per second (QPS) than expected.

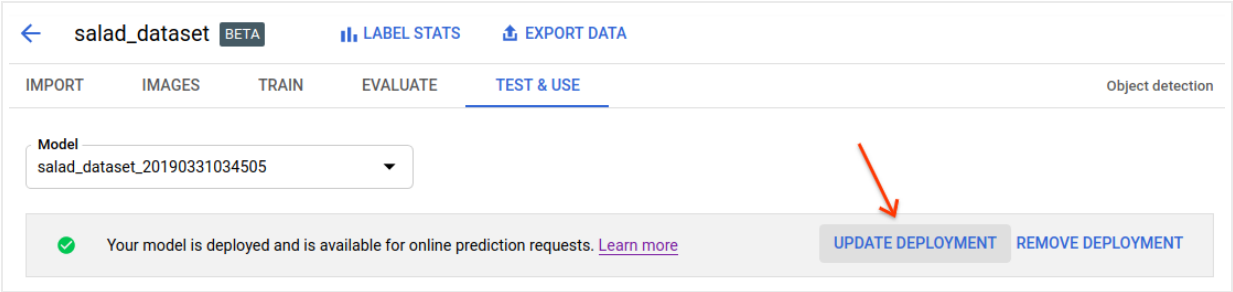
You can change this node number *without* first having to undeploy the model. Updating deployment will change the node number without interrupting your served prediction traffic.

Node count will be subject to quota listed at the quota page; **by default, a user can get up to 10 nodes**.

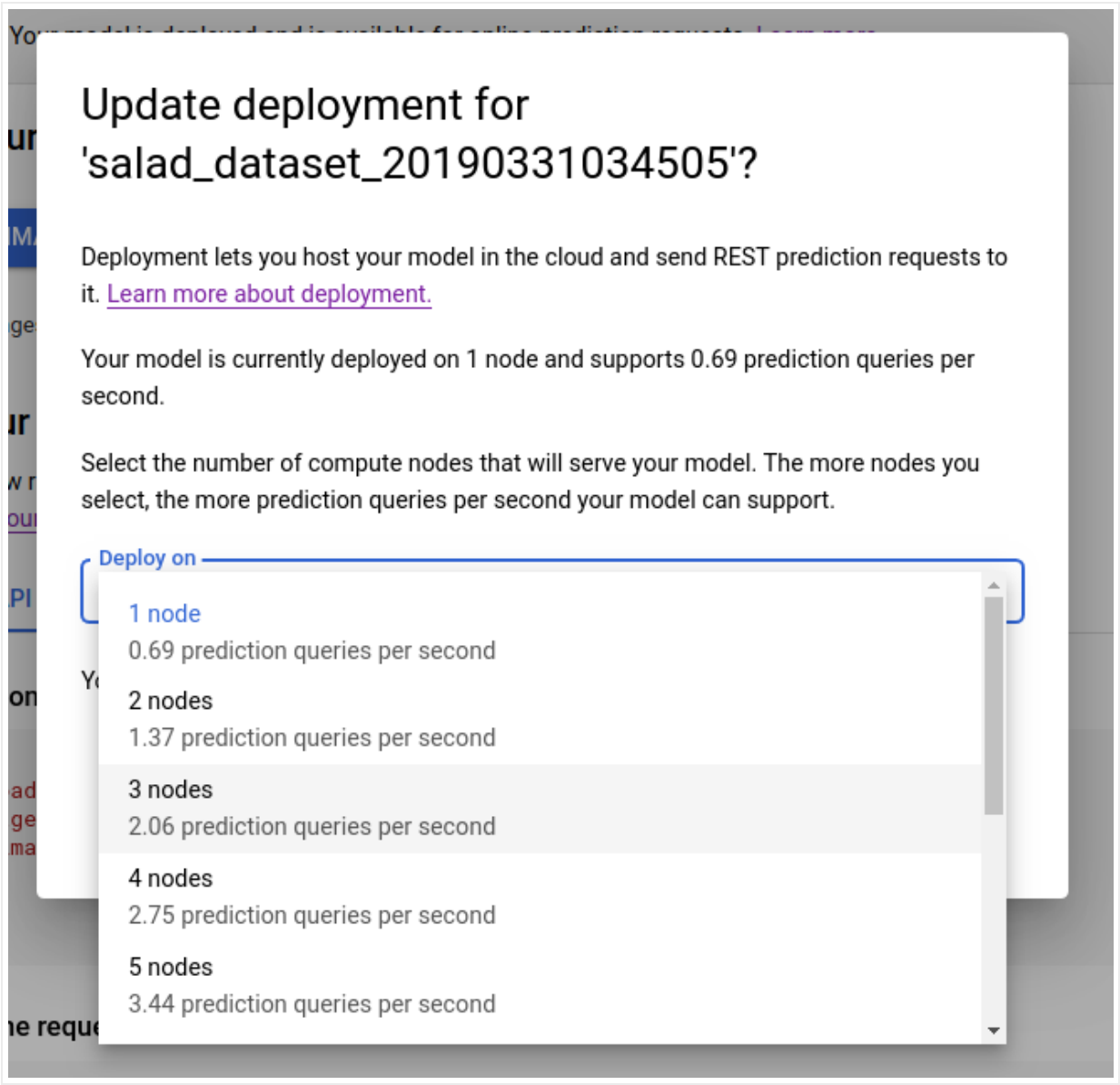
WEB UI REST & CMD LINE MORE

1. In the Cloud AutoML Vision Object Detection UI (https://console.cloud.google.com/vision) and select the **Models** tab (with lightbulb icon) in the left navigation bar to display the available models.

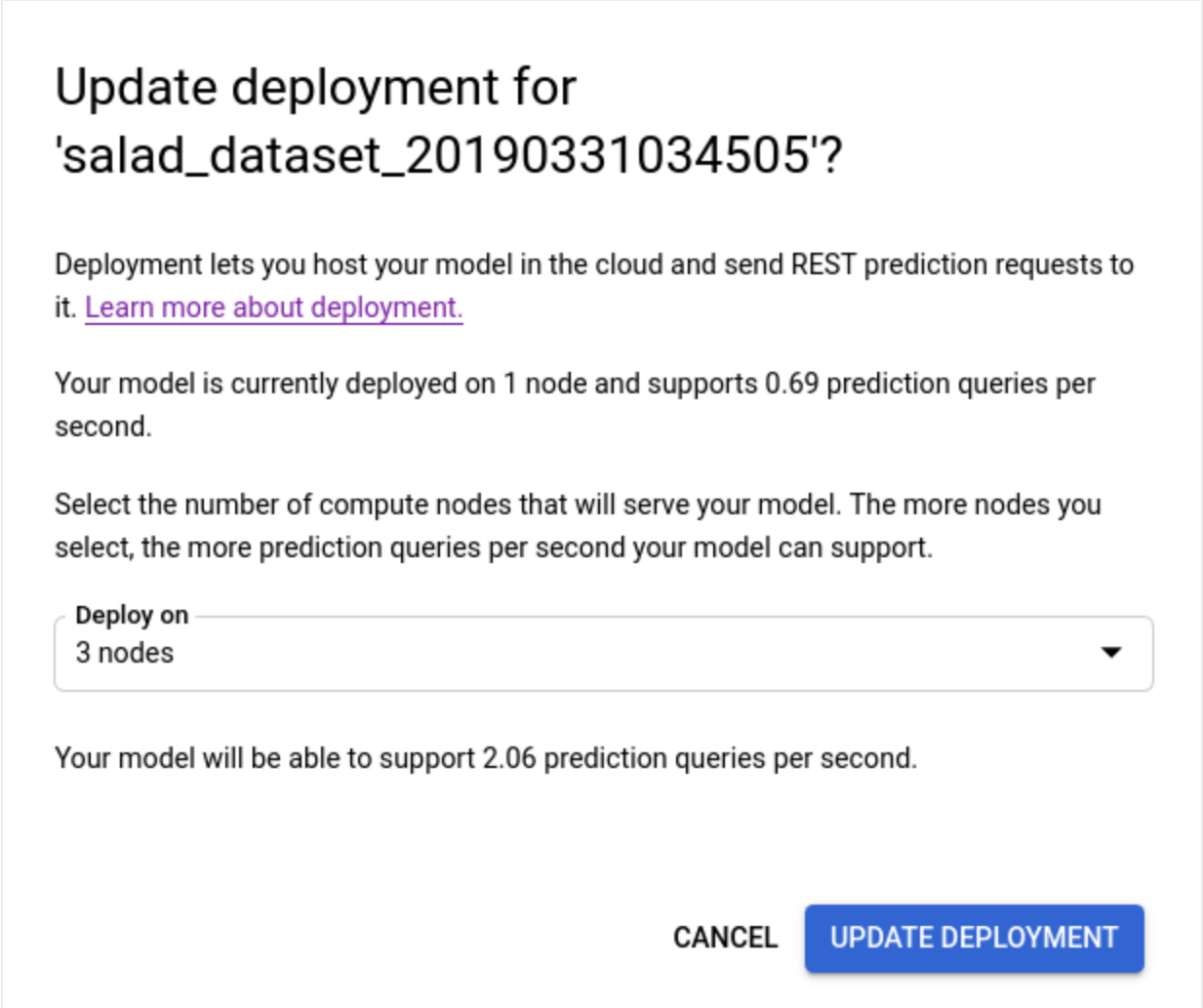
   To view the models for a different project, select the project from the drop-down list in the upper right of the title bar.
2. Select your trained model that has been deployed.
3. Select the **Test & Use** tab just below the title bar.
4. A message is displayed in a box at the top of the page that says "Your model is deployed and is available for online prediction requests". Select the **Update deployment** option to the side of this text.



5. In the **Update deployment** window that opens select the new node number to deploy your model on from the list. Node numbers display their estimated prediction queries per second (QPS).

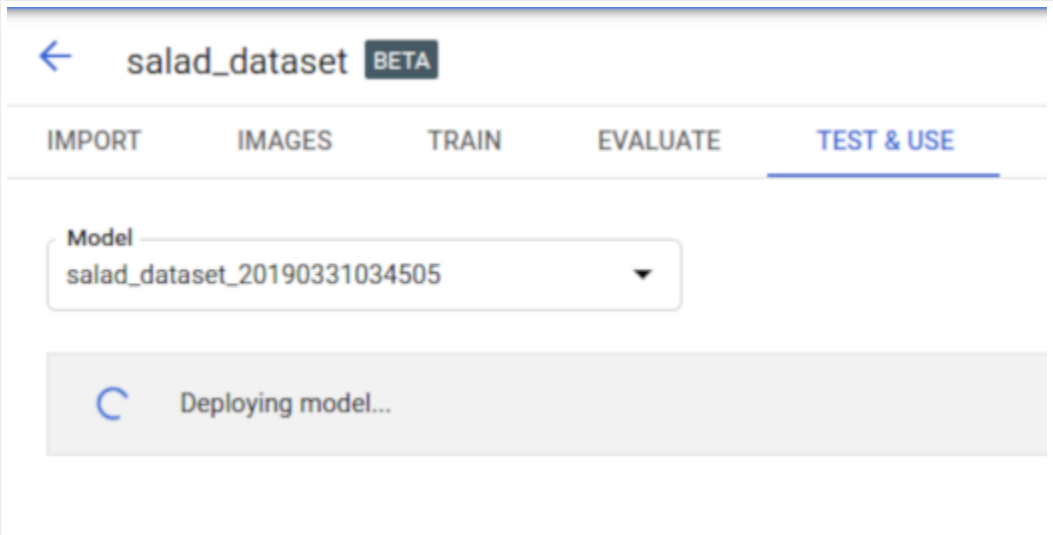


6. After selecting a new node number from the list select **Update deployment** to update the node number the model is deployed on.



7. You will be returned to the **Test & Use** window where you see the text box now displaying "Deploying model...".

★ **Note:** This is similar to the initial deployment of your model.



8. After your model has successfully deployed on the new node number you will receive an email at the address associated with your project.

Last updated November 20, 2019.