AI in Business, Best Practices, Learn Data Science, Oracle AI

Simplifying AI Integration: Bringing Hugging Face models to AI Quick Actions in OCI Data Science

September 10, 2024 | 4 minute read


Wendy Yip
Senior Product Manager, OCI Data Science


Tzvi Keisar
Director of Product Management


Julien Lehmann
Product Marketing Director, Oracle Modern Data Platform

In the latest release of AI Quick Actions, we're supporting bring-your-own-model from Hugging Face. Hugging Face is a popular AI model repository that hosts many state-of-the-art large language models (LLMs), including Meta's Llama models and Mistral. In the previous release of AI Quick Actions, we supported bring your own model from Oracle Cloud Infrastructure (OCI) Object Storage and opened up the option for customers to bring any model to be deployed and fine-tuned, provided the model artifacts are first saved in Object Storage. With this latest release, we take it one step further by allowing you to bring a model from Hugging Face into AI Quick Actions, without the need to download any model artifacts. This option opens the door for you to choose from thousands of models that Hugging Face provides and lets you utilize AI Quick Actions' no-code experience to work with these models in a simplified process. In this blog post, we will cover how to bring a model from Hugging Face into AI Quick Actions.

Register a model from Hugging Face

When you first navigate to AI Quick Actions inside the Data Science notebook, you come to the model explorer. Under my models, you can choose to import a new model from either Hugging Face or Object Storage.

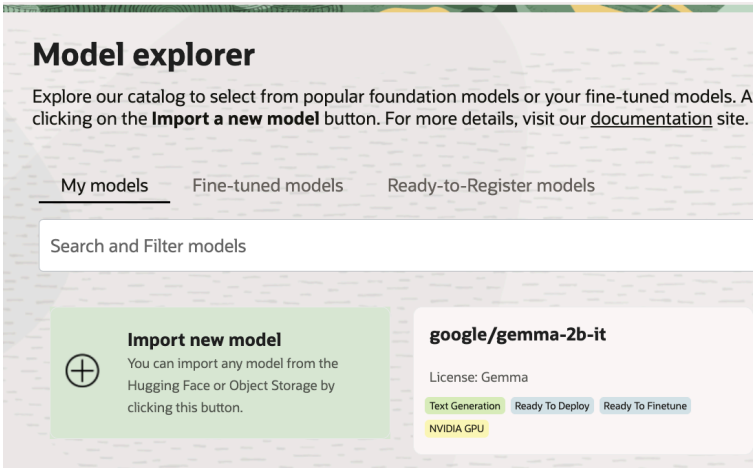


Figure 1: OCI Data Science model explorer

You can choose to register a model from Hugging Face or Object Storage from the menu. Model registration is a necessary process for a model to be brought into AI Quick Actions.

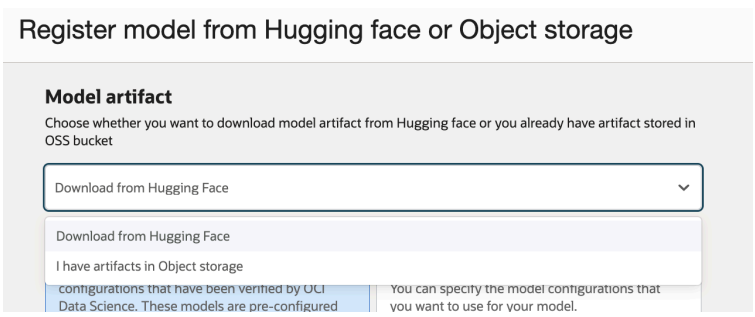


Figure 2: Register a model from Hugging Face or Object Storage in Oracle Cloud Console

You can register either a verified or unverified model. A verified model is a model whose configuration the OCI Data Science service has tested already; an unverified model is one that the OCI Data Science service hasn't tested.

You can navigate to "register a verified model" and see the list of models that our service has tested in the drop down menu, as shown in Figure 3. If the model you are interested in is on the list, select the model. You can specify the Object Storage location where you want to save the model artifacts and download them from Hugging Face, as shown in Figure 4.

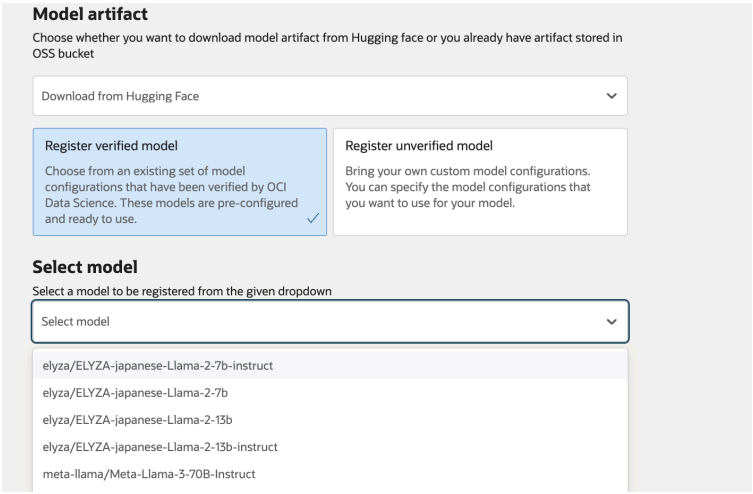


Figure 3: Register a verified model in the Oracle Cloud Console

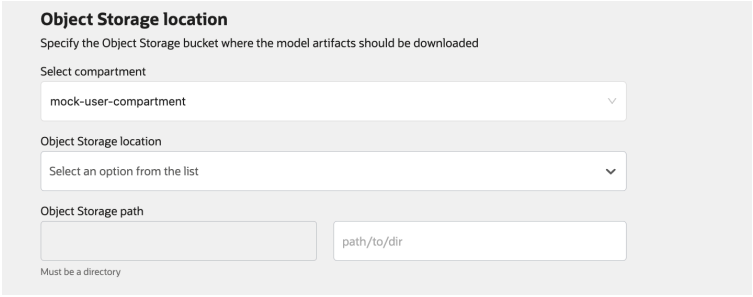


Figure 4: Specify an Object Storage location in the Console

Hugging Face offers certain gated models that require the acceptance of a user agreement, such as Meta's Llama models. If you choose to bring a gated model from Hugging Face into AI Quick Actions, you need to log into Hugging Face with the Hugging Face CLI and your Hugging Face token to verify your access to the model. You can use the terminal launched inside the notebook. For information on how to log in with the Hugging Face CLI, see the Hugging Face CLI page. If you don't have a Hugging Face token, refer to this page to generate one.

If the model you want to bring in from Hugging Face isn't a verified model, navigate to "Register unverified model" and enter the name of the model as it appears on the Hugging Face website to search for it, as shown in figure 5.

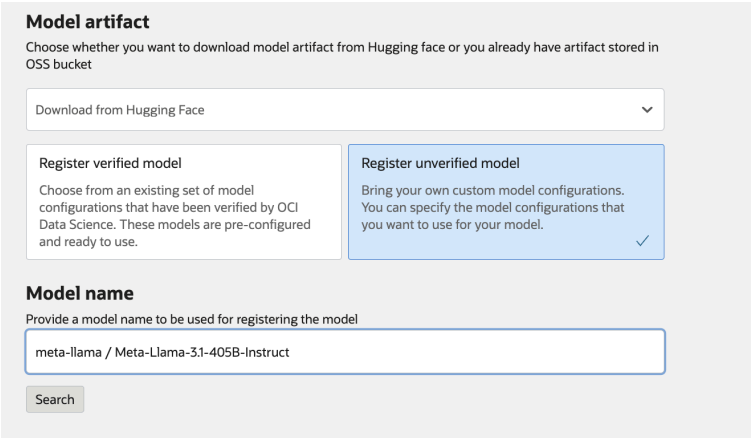
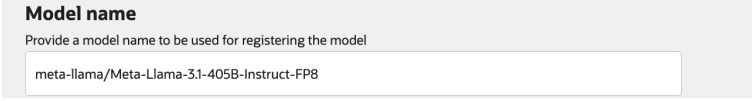


Figure 5: Register an unverified model from Hugging Face in the Oracle Cloud Console

The search will return basic download statistics and information about the model. You can choose a service-managed inference container for working with the model, one for a model compatible with inferencing engine vLLM or one compatible with TGI, as shown in Figure 6. For models in GGUF format, you can choose to use a service-managed inference container with llama.cpp.



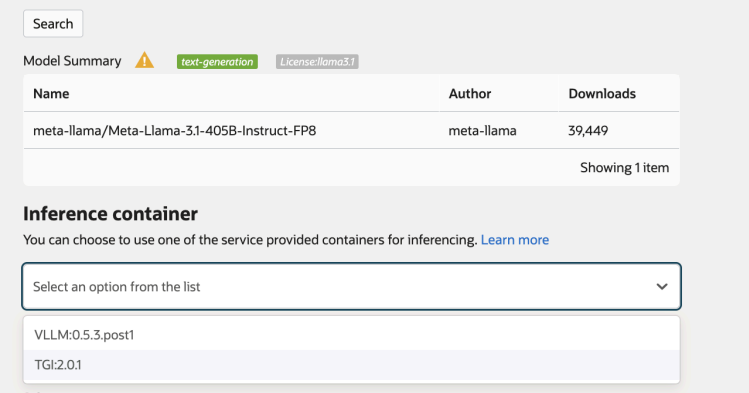


Figure 6: Specify the inference container

After the model registration is complete, you can see the name of the model under “My models” in the model explorer. You can deploy and fine-tune this model using AI Quick Actions’ no-code user environment.

Key takeaway

Our latest feature of bringing your own model from Hugging Face enables you to harness the power of open source models and use them in a simplified workflow inside AI Quick Actions.

Try Oracle Cloud Free Trial! A 30-day trial with US$300 in free credits that gives you access to the Oracle Cloud Infrastructure Data Science service. For more information, see the following resources:

- Learn about AI Quick Actions in OCI Data Science.
- Full sample, including all files in OCI Data Science sample repository on GitHub.
- Visit our service documentation.
- Watch our tutorials on our YouTube playlist.
- Try one of our LiveLabs. Search for “data science.”
- Question? Reach out to us at ask-oci-data-science_grp@oracle.com.

Wendy Yip
Senior Product Manager, OCI Data Science

Tzvi Keisar
Director of Product Management

Product Manager in OCI Data Science

Julien Lehmann
Product Marketing Director, Oracle Modern Data Platform

In Oracle since 2018, Julien is a subject matter expert as cloud and cybersecurity/CDN solutions architect, product director and successful global sales. He's a certified architect with OCI, AWS and Azure. Julien
Show more