Oracle AI & Data Science Blog

Learn Data Science, Oracle AI

Model deployment for real-time predictions is now available in Oracle Cloud Infrastructure Data Science

Tzvi Keisar | March 22, 2021
Senior Principal Product Manager

On March 19, 2021, Oracle Cloud Infrastructure (OCI) Data Science released a new feature called Model Deployment to enable the serving of machine learning models as HTTP endpoints and provide real-time scoring of data. This feature is available in the OCI Software Development Kits (SDKs), OCI Command Line Interface (CLI), and OCI Console.

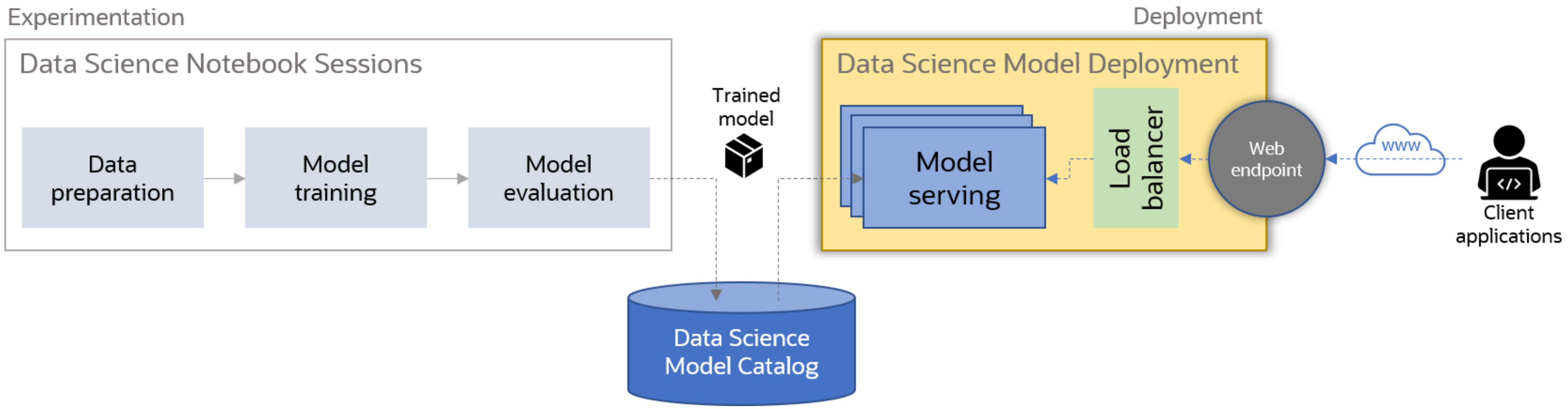
Machine learning, as complex as it may be, serves a business purpose, and that is using existing data to make predictions on future data. To meet that purpose, models that were trained on existing data need to be available to make predictions on fresh data that comes in.

We are excited to announce the release of model deployment, enabling models to be served as HTTP endpoints, receive requests, and send responses back with the model predictions in real time.

What does model deployment mean?

What does model deployment mean?

The machine learning journey does not end after training the model. Data science teams create models in order to generate predictions on fresh data. One of the most useful ways to do that is to wrap the model as a web service and enable it to accept HTTP requests with data to be scored. The web service then invokes the model using the request data and sends the output of the model prediction back to the requestor as HTTP response. This provides for data to be scored on demand **in real time.**



Deploying models in Oracle Cloud Infrastructure Data Science service

OCI Data Science provides a **managed solution for deploying models at scale**. The service comes with everything you can expect from a cloud service out-of-the-box:

Managed deployment

The service handles all the complex "wiring." There is no need to take care of provisioning the compute, installing the web service, or anything related to the infrastructure. Everything is handled and optimized by the service.

Secure

With Oracle Cloud Infrastructure Identity and Access Management built-in to the service, all that's left for you to do is define the groups and the policies for those groups to create and invoke deployment endpoints.

Easy and streamlined

By pulling models from the model catalog, along with their inference, you can have a deployed model ready with a few lines of code (from the SDK/CLI) or a few clicks of a button (from the OCI Console). The built-in load balancer will handle the web traffic automatically and efficiently.

Flexible

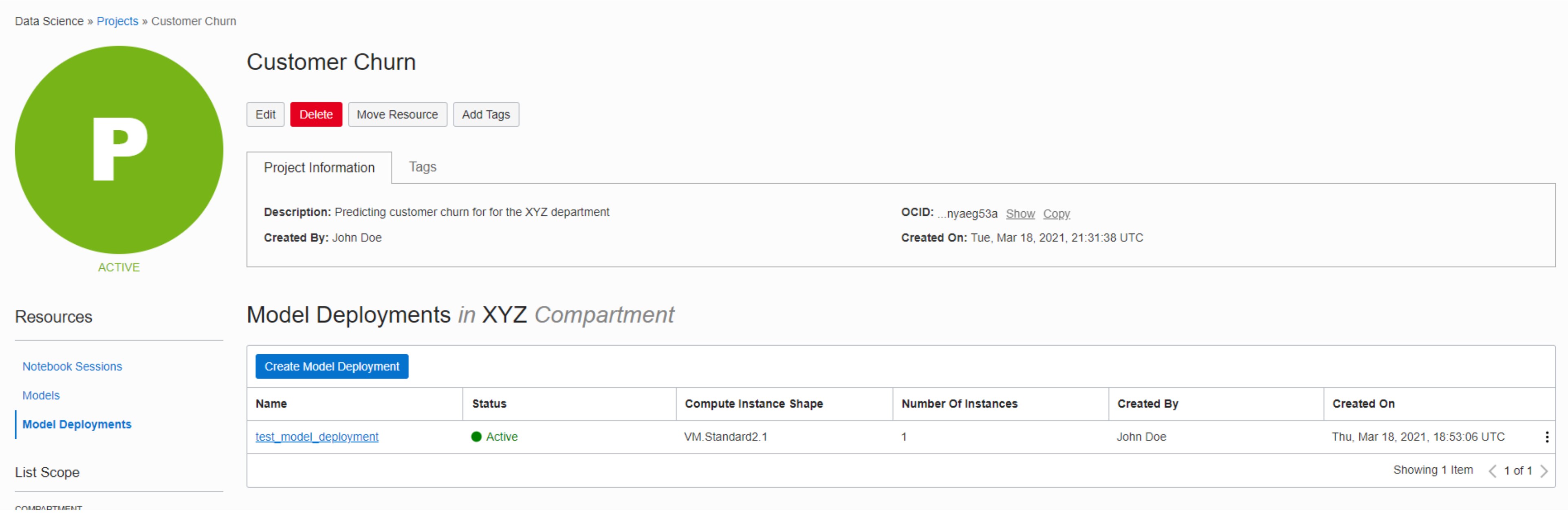
Choose from any of the available compute shapes to deploy on, all are fully managed so you don't have to worry about provisioning. Need to change the compute? You can scale up or down without downtime.

Traceable

Analyze the logs to gain insights into what's happening with the model's scoring or debug issues with the deployment.

Deploy from code or from the OCI Console

To deploy models with a few clicks directly from the OCI Console, all you need is a registered model in the model catalog.



Data Science » Projects » Customer Churn

Customer Churn

Description: Predicting customer churn for for the XYZ department

Created By: John Doe

OCID: ...nyaeg53a Show Copy

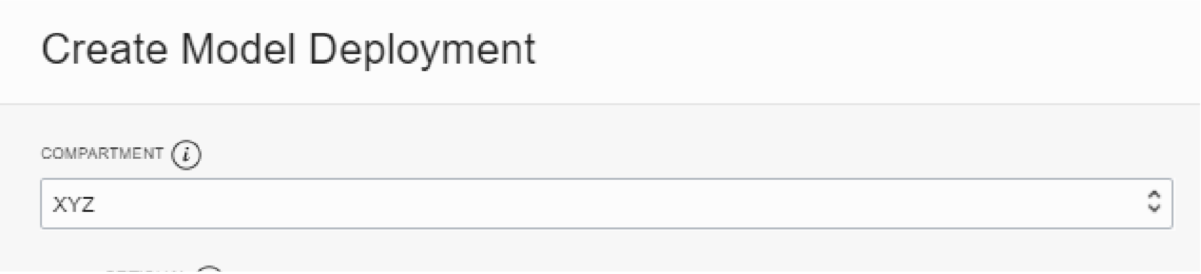
Created On: Tue, Mar 18, 2021, 21:31:38 UTC

Model Deployments *in* XYZ *Compartment*

Name	Status	Compute Instance Shape	Number Of Instances	Created By	Created On
test_model_deployment	● Active	VM.Standard2.1	1	John Doe	Thu, Mar 18, 2021, 18:53:06 UTC

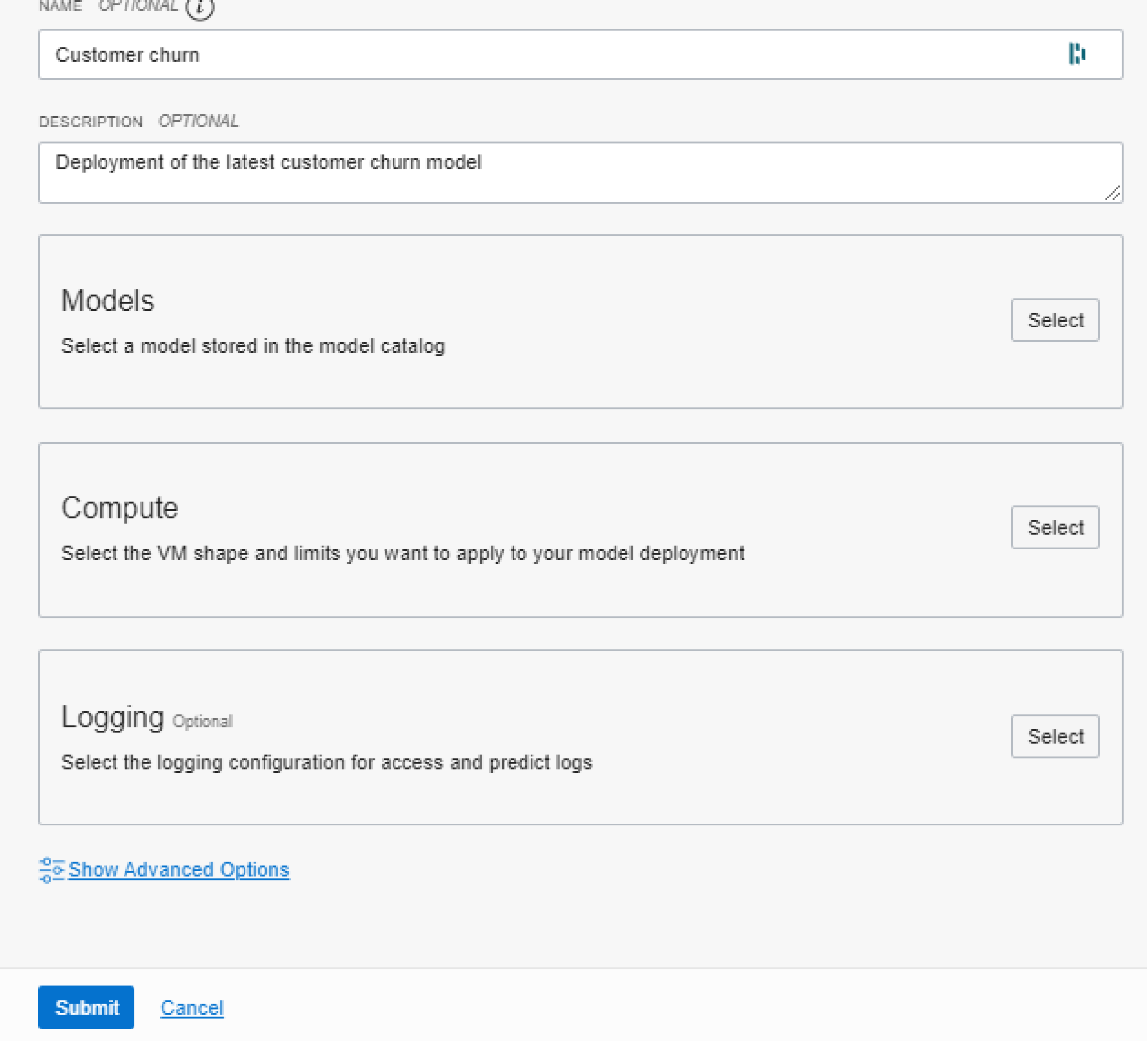
Showing 1 Item ‹ 1 of 1 ›

Figure 1 The new model deployment resource page with "Create Model Deployment" button



Create Model Deployment

COMPARTMENT: XYZ



NAME OPTIONAL

Customer churn

DESCRIPTION OPTIONAL

Deployment of the latest customer churn model

Models

Select a model stored in the model catalog

Compute

Select the VM shape and limits you want to apply to your model deployment

Logging Optional

Select the logging configuration for access and predict logs

Figure 2 Create model deployment page. Select the model from the model catalog, select compute and logging, and deploy.

If you prefer code, you can use the OCI SDK to deploy a model with a few lines of code. You'll also be able to customize the logging configuration to your preference.

Deploy model (this is a short version for illustration purpose, the full code is available in the service documentation)**:**

```
# Create the model deploy configuration:
model_deploy_configuration = CreateModelDeploymentDetails(display_name=model_deployment_name,
                                                          description=description,
                                                          project_id=project_ocid,
                                                          compartment_id=compartment_ocid,
                                                          model_deployment_configuration_details=single_model_config
```

```
                                                           model_deployment_configuration_details=single_model_config,
                                                           category_log_details=logs_configuration_details_object)

# Creating a model deployment. This action takes a few minutes to complete.
data_science_composite = DataScienceClientCompositeOperations(data_science)
model_deployment = data_science_composite.create_model_deployment_and_wait_for_state(model_deploy_configuration,
                                                                                     wait_for_states=["SUCCEEDED", "FAII
```

Copy code snippet

Automatic handling of HTTP traffic with load balancing

The managed service also handles the load balancing of incoming web traffic to make sure no single endpoint is overwhelmed with requests. The load balancer receives each request and routes it to the next available endpoint. This way, you get a continuous, reliable predictions service without the hassle of managing API gateways by yourself.

Manage the model deployment lifecycle

You are in control of the model deployment state. Once the deployment is ready, switch it to an active state from the OCI Console, the SDK, or the CLI.

When you want to pause the predictions service, simply deactivate it. The deployment is still ready but it no longer provides service to incoming requests. This also stops the billing for this deployment to save you costs.

Call the scoring service from anywhere

Once you have an active endpoint, you can send requests from any HTTP client.

Utilizing OCI Request Signatures for secure connection, you can use web clients such as Postman, Fiddler, or cUrl, or your own client application, for testing purposes.

Learn more about model deployment and how to call the endpoint in the service documentation.

Keep in touch!

- Visit the OCI Data Science webpage
- Visit our service documentation
- Visit our YouTube Playlist
- Visit our LiveLabs Hands-on Lab

Tzvi Keisar
Senior Principal Product Manager

Previous Post

6 tips for easier AI adoption

Kristen Lee | 6 min read

Next Post

A deepfake example on Oracle Cloud Infrastructure

Maria Patelkou | 9 min read