Deploying Mistral 7B Instruct NVIDIA NIM Microservice with OCI Data Science via OCI Marketplace

NOTE – At the time of writing this documentation April 2025, this is only available within the ***US East Ashburn*** Region.

Referenced Documentation

[1] - https://blogs.oracle.com/ai-and-datascience/post/nvidia-nim-on-oci-marketplace
[2] - https://cloudmarketplace.oracle.com/marketplace/en_US/listing/182674476

Description

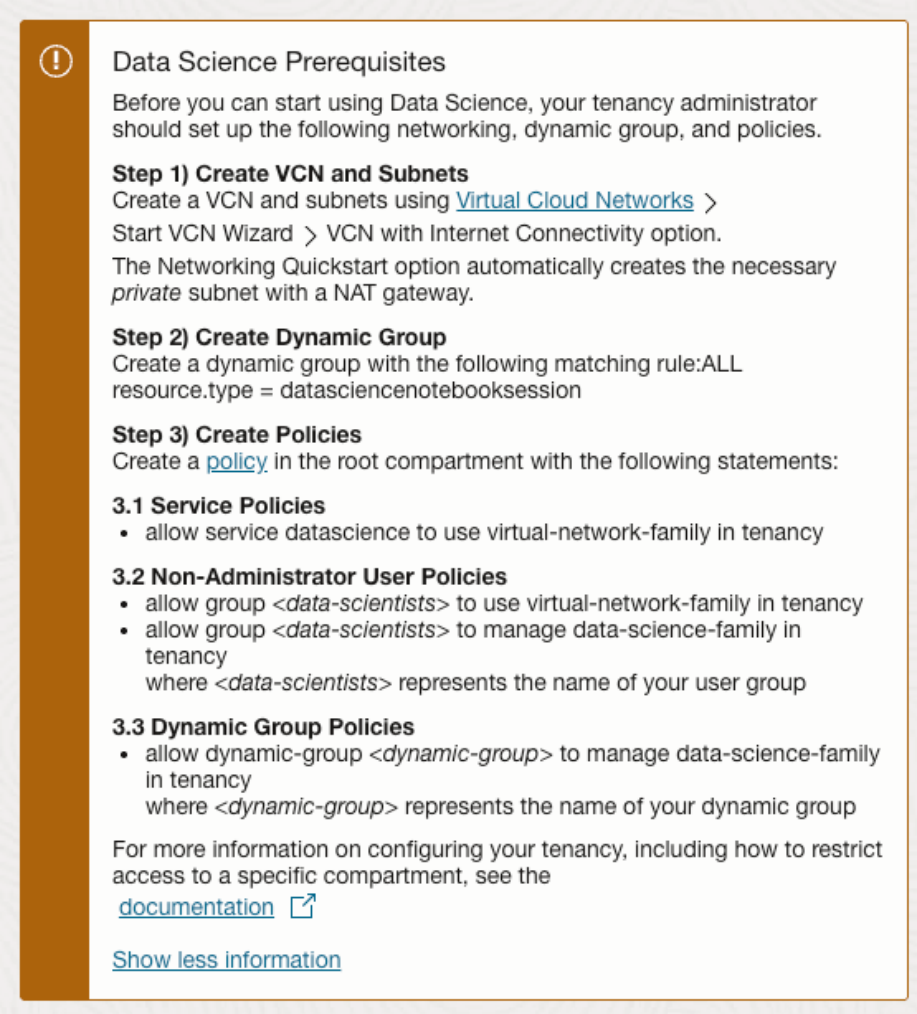
NVIDIA NIM™ provides prebuilt, optimized inference microservices that let you deploy the latest AI foundation models with security and stability on any NVIDIA-accelerated infrastructure, as OCI Data Science managed inference endpoints, for a code-free, scalable, secure inferencing. ***[1]***

NVIDIA NIM™, part of NVIDIA AI Enterprise, is a set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing. These prebuilt containers support a broad spectrum of AI models—from open-source community models to NVIDIA AI Foundation models, as well as custom AI models. NIM microservices are deployed with a single command for easy integration into enterprise-grade AI applications using standard APIs. Built on robust foundations, including inference engines like NVIDIA Triton Inference Server™, TensorRT™, TensorRT-LLM, and PyTorch, NIM is engineered to facilitate seamless AI inferencing at scale, ensuring that you can deploy AI applications with confidence. NIM is the fastest way to achieve accelerated generative AI inference at scale." ***[1]***

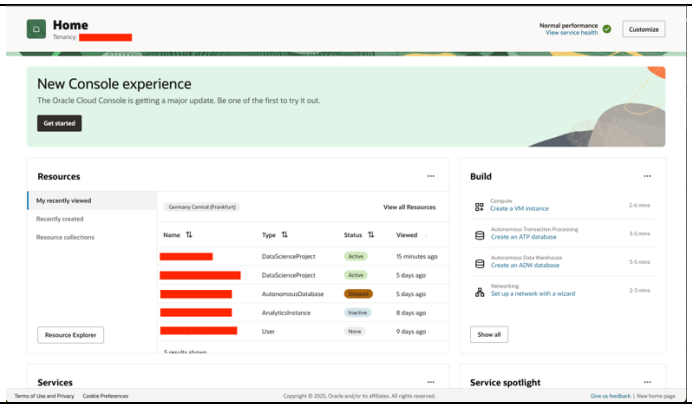
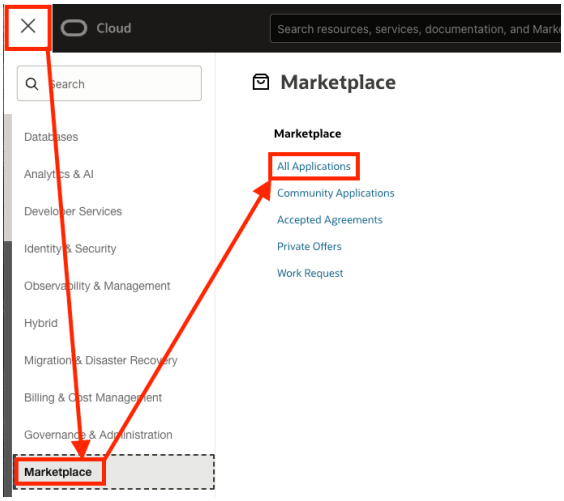
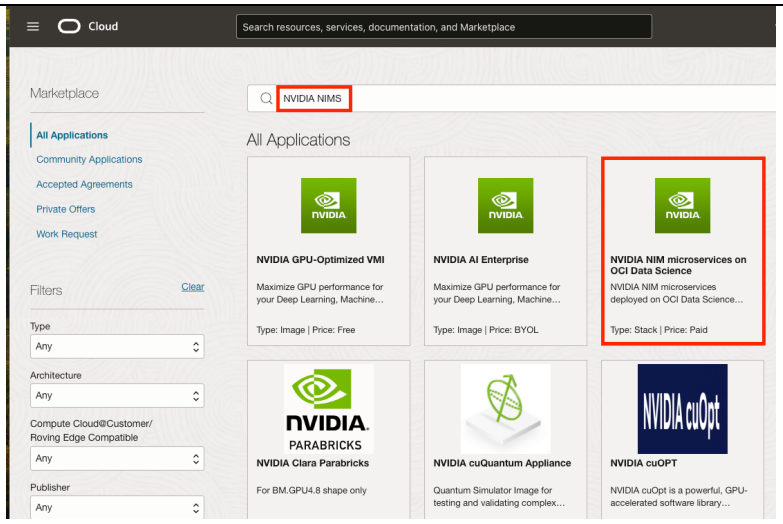
Models Available as NVIDIA NIM microservices *[1]*:

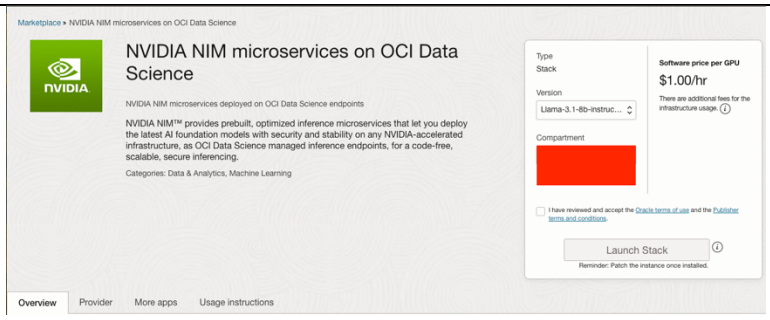
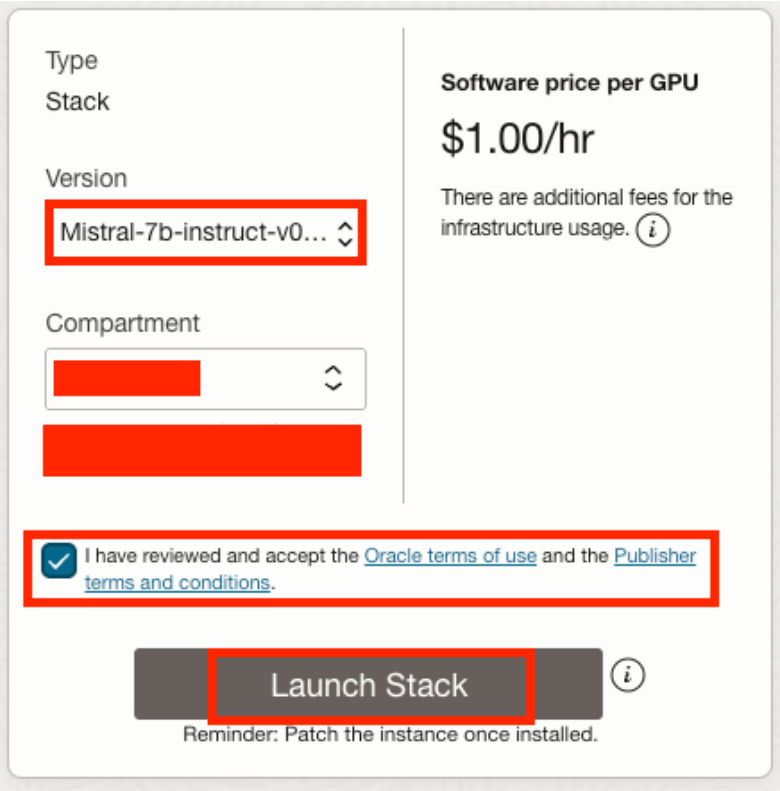
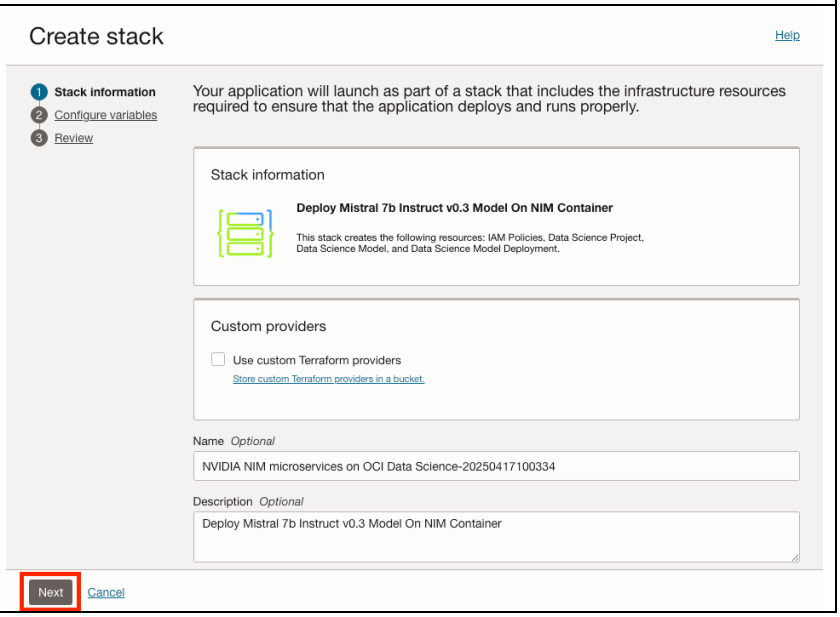
- **Meta Llama 3.1 8B Instruct** is an 8-billion-parameter multilingual large language model (LLM) pretrained and instruction tuned generative model. The Llama 3.1 instruction tuned text only model is optimized for multilingual dialogue use cases.
- **Meta Llama 3.1 70B Instruct** is an 70-billion-parameter multilingual large language model (LLM) pretrained and instruction tuned generative model. The Llama 3.1 instruction tuned text only model is optimized for multilingual dialogue use cases.
- **Mistral 7B Instruct v0.3** is a language model that can follow instructions, complete requests, and generate creative text formats. It is an instruct version of the Mistral-7B-v0.3 generative text model fine-tuned using a variety of publicly available conversation datasets.
- **Mixtral 8x7b Instruct v0.1**is a language model that can follow instructions, complete requests, and generate creative text formats. Mixtral 8x7B a high-quality sparse mixture of experts model (SMoE) with open weights.

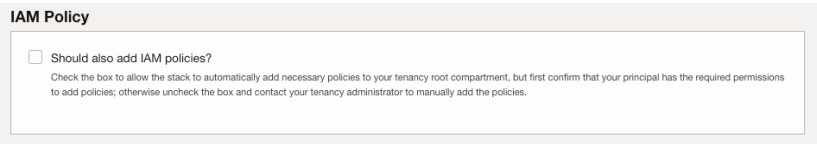
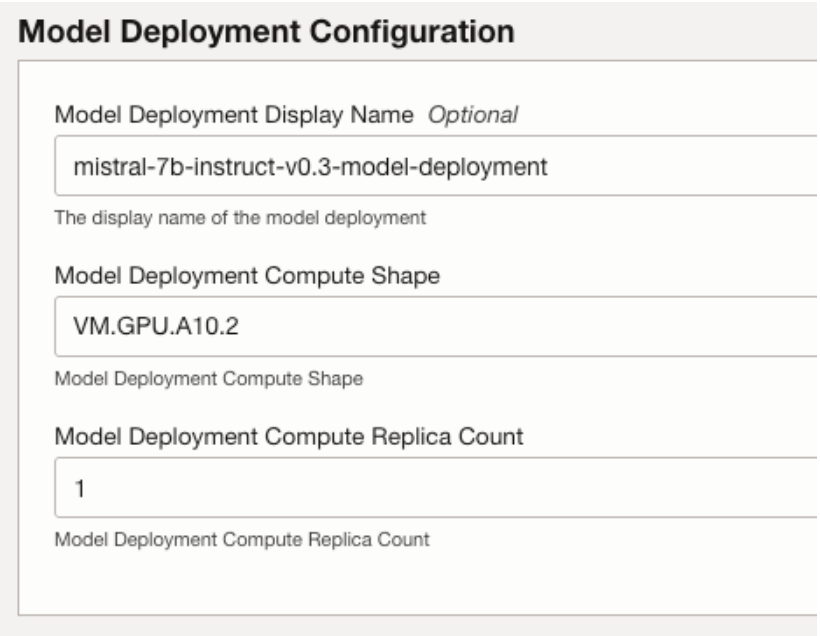
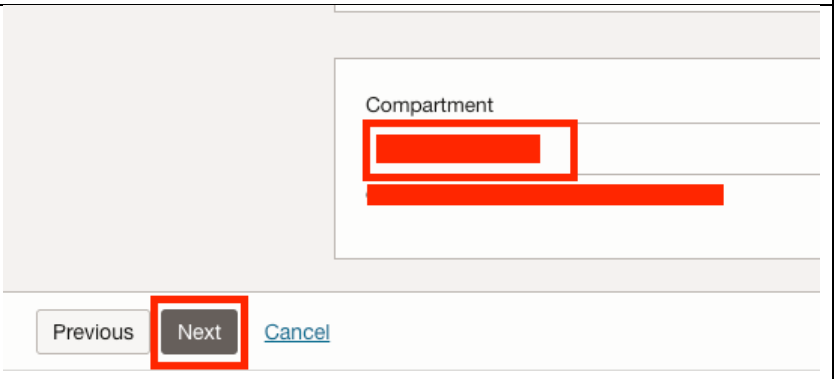
Pre-Requisites

- Ensure you have your OCI Data Science GPU service limits raised for the GPU Shapes you plan to use. This can be done from OCI Console by your OCI Admin.
- Permissions to create and manage resources within an OCI Compartment. Please speak with your OCI Admin to gain this access.
- Have your OCI Admin create the below Policies within the Root Compartment:
    - `define tenancy containerTenancy as 'ocid1.tenancy.oc1..aaaaaaaakjncznwynlcsjpxrqub3jskbzmz3qlkgoffiv7yjmyrfqqgy7gaq'`
    - `define tenancy modelTenancy as 'ocid1.tenancy.oc1..aaaaaaaawhegebhtosat4uy2xjmdvgqreelfxrf4zlquo6bcyjsp6dh5rjwq'`
    - `endorse any-user to read repos in tenancy containerTenancy where all {request.principal.type = 'datasciencemodeldeployment'}`
    - `endorse any-user to read objects in tenancy modelTenancy where all {request.principal.type = 'datasciencemodeldeployment'}`
    - `allow any-user to manage data-science-family in Tenancy`
- ***[OPTIONAL]*** If this is your first time using OCI Data Science, please get your OCI Admin to follow these instructions to enable you to use OCI Data Science Notebook Sessions to utilise the JupyterLab Notebook Environment to later call our Deployed Model.
    - 

Guide

Step	Screenshot
Login to the Cloud Console. ***cloud.oracle.com***	
Navigate to the OCI Marketplace where we can deploy the NVIDIA NIMs Microservices to OCI Data Science. Navigate to ***OCI Menu > Marketplace > All Applications.***	
Search for ***NVIDIA NIM***. Select the ***NVIDIA NIM microservices on OCI Data Science*** Application.	

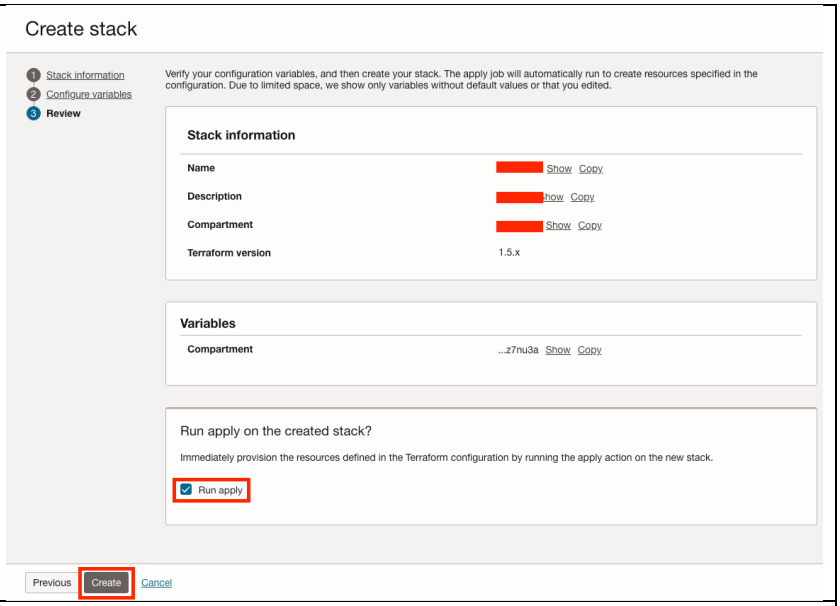
Please read through the entire page to be familiar with what you are deploying and the costs associated with deploying this service.	
Select the Version, I have gone for ***Mistral 7B Instruct v0.3*** Ensure you select the right ***Compartment*** you have permissions to Manage resources within. ***Please review and accept the T&Cs.*** Click ***Launch Stack***. A ***Stack*** is a feature of Oracle Cloud that empowers users to automate the provisioning of multiple cloud services as a single unit.	
You can leave all the options on the first page as default. Click ***Next***. ***Please Note*** – This stack will create an OCI Data Science Project, Model and Model Deployment behind the scenes to host the NIM Microservice.	

Leave the IAM Policy option ***unticked***. This is because your OCI Admin should have done this manually as part of the pre-requisites.	
Here you can set your ***Model Deployment Name***. I have left as default. Select the ***Model Deployment Compute Shape***. I have left as the default VM.GPU.A10.2 which is 2 A10 GPUs. Select the ***Model Deployment Compute Replica Count***. This is the number of Nodes you want sitting behind the deployment. I have left as the default, 1.	
Confirm you still have your correct ***Compartment*** select which you have permissions to manage resources in. Click ***Next***.	

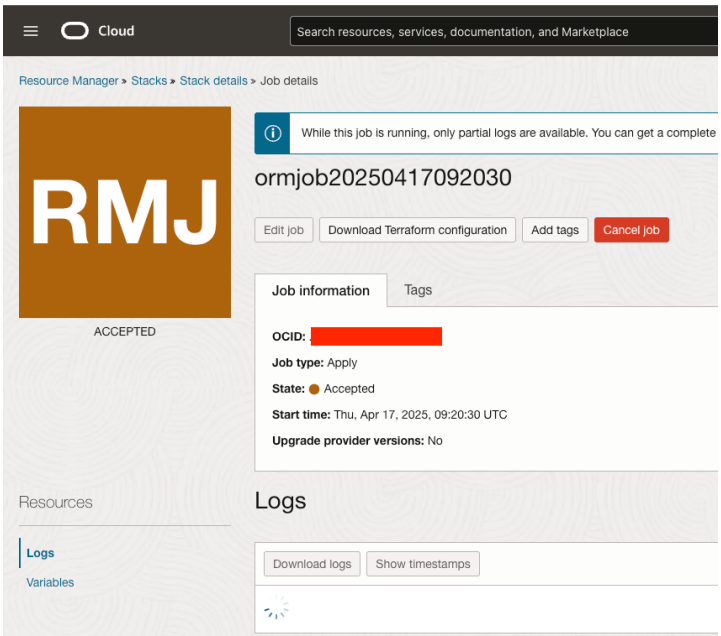
Review all the information on the page.

Ensure the ***Run apply*** option has been ticket so the provisioning of resources can happen automatically.

Click ***Create***.

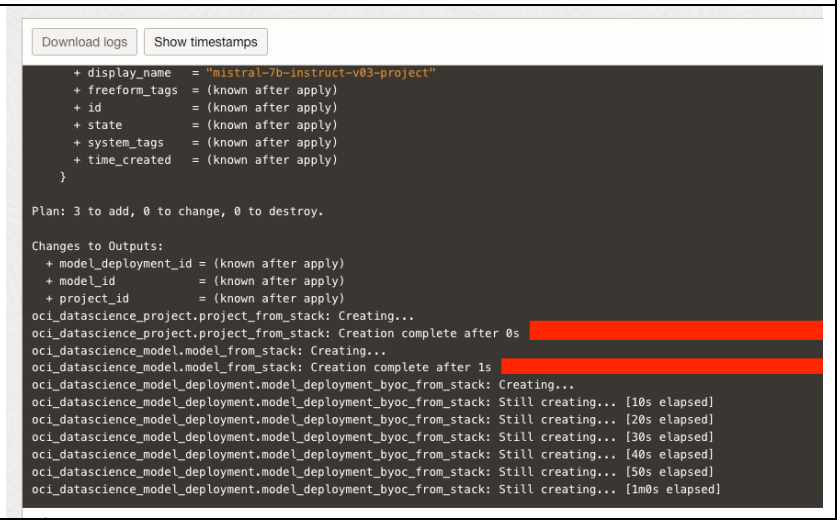


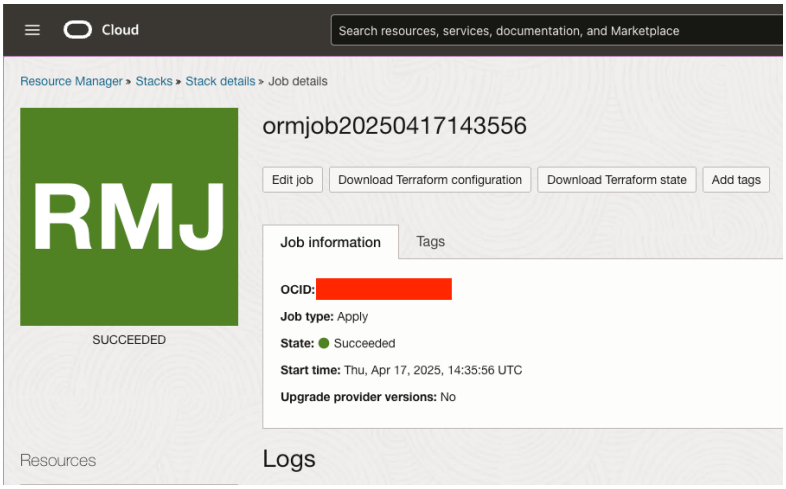
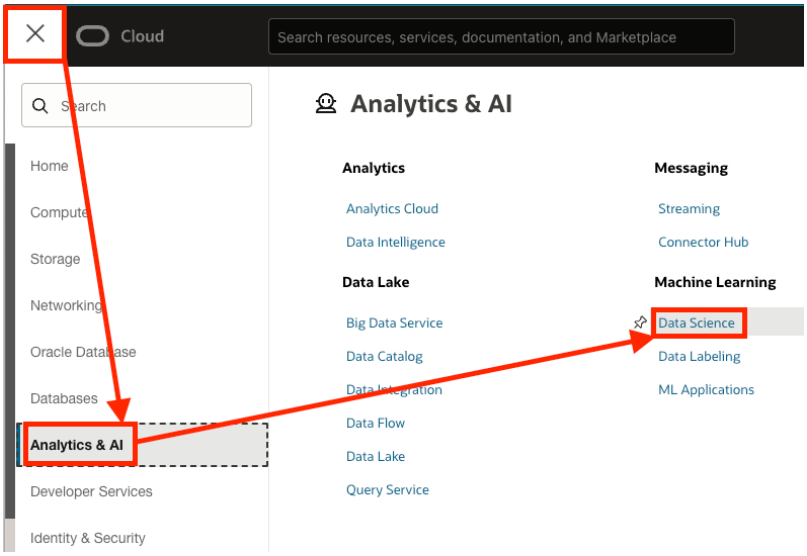
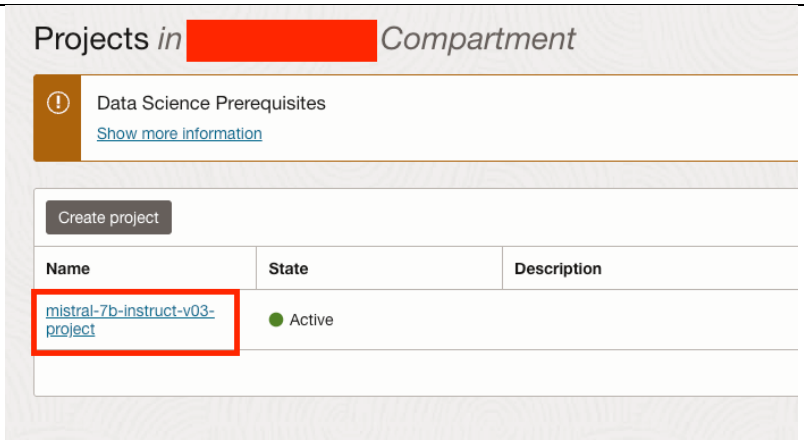
The Job will then go into the ***ACCEPTED*** State, then transition to the ***IN PROGRESS*** state.

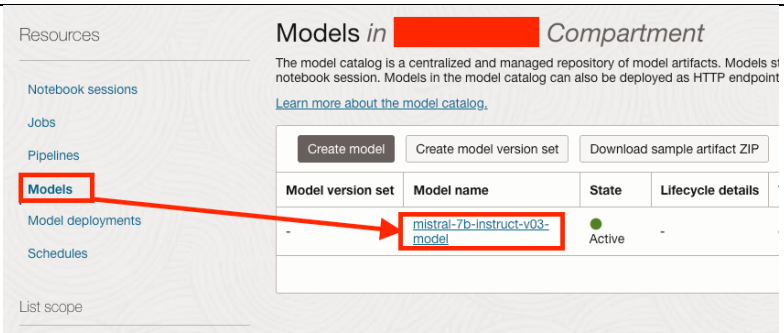
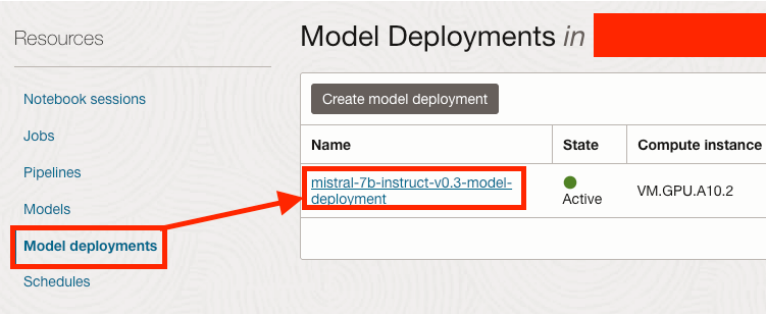
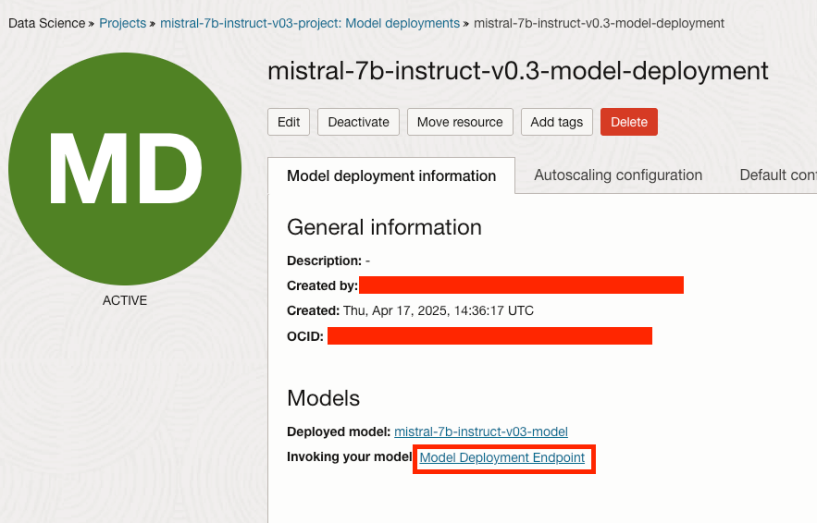


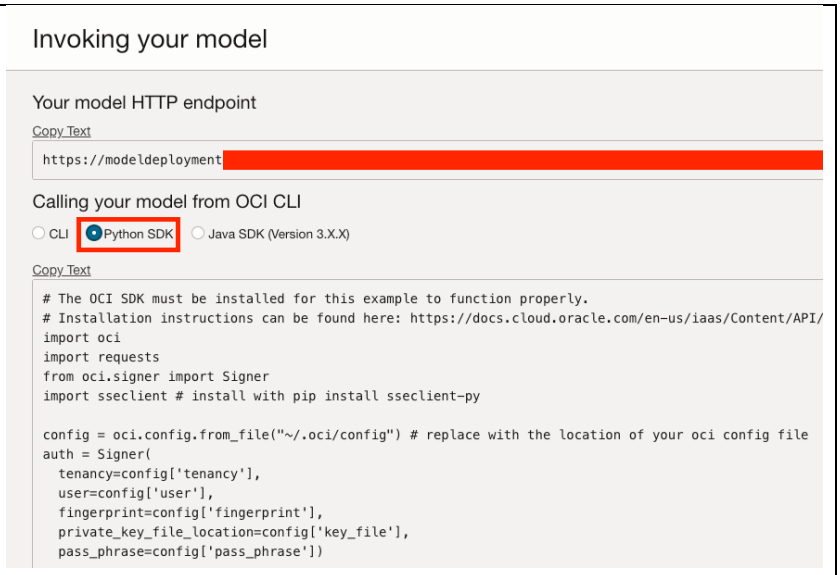
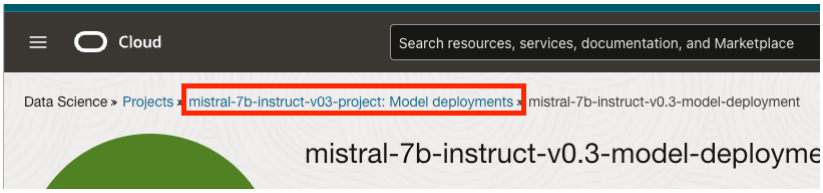
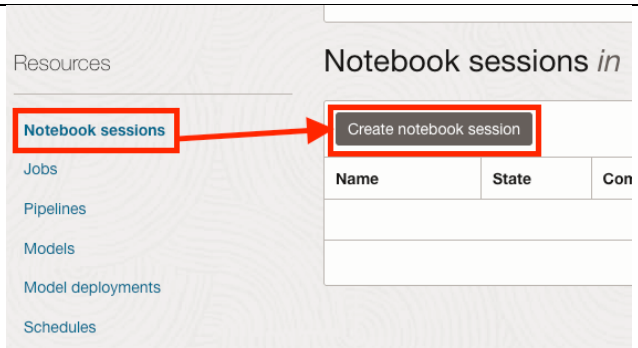
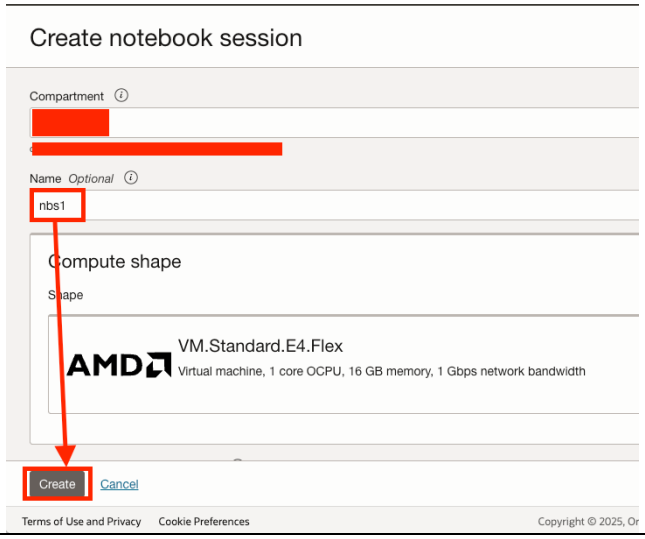
If you scroll down to the Logs, you will be able to see the detailed progress of:

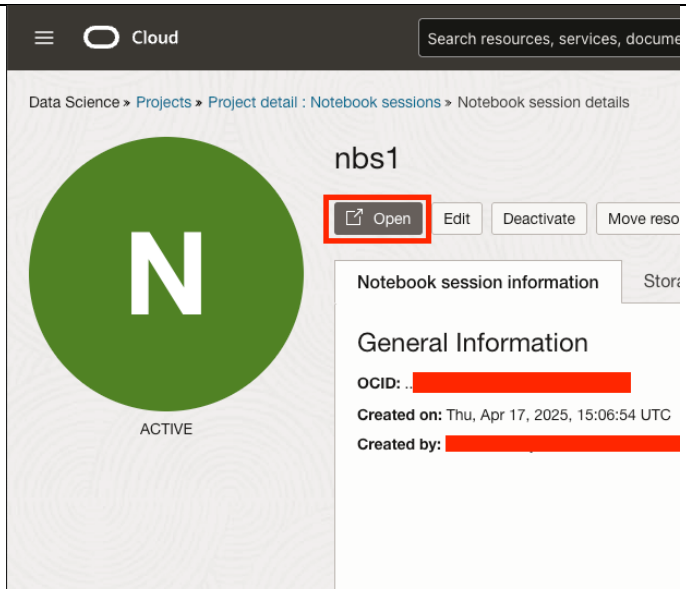
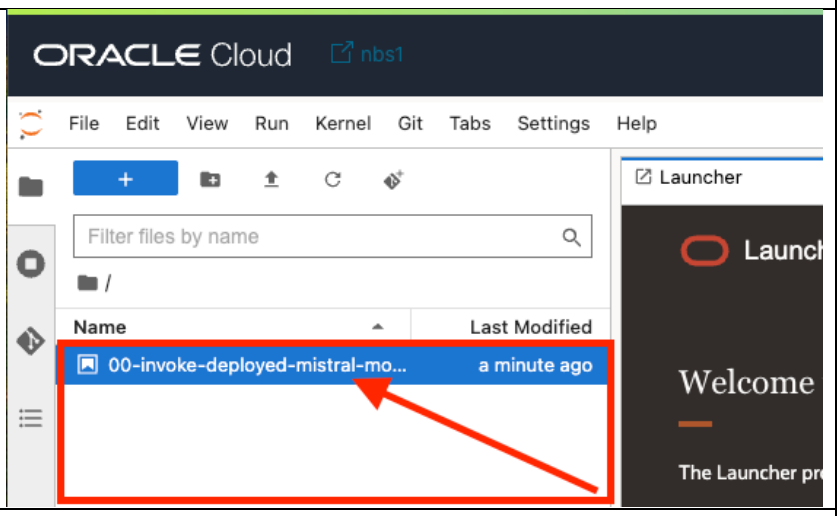
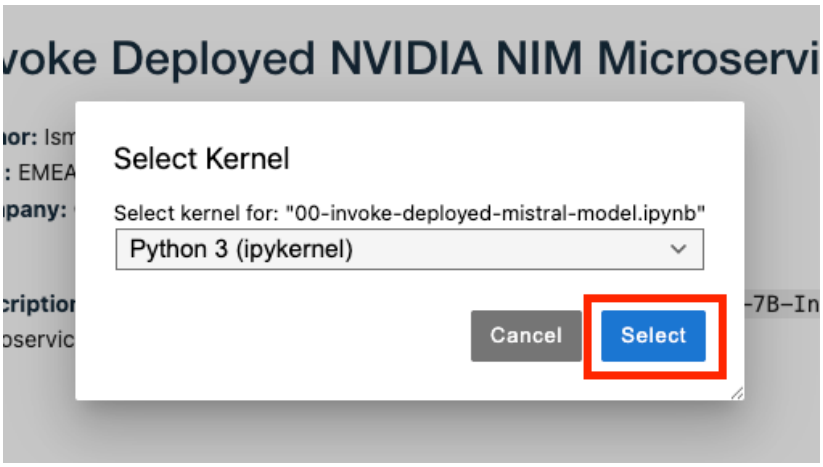
Creating Project
Creating Model
Creating Deployment



Once complete the Stack will display **SUCCEED**.	
We can now navigate to OCI Data Science. ***OCI Menu > Analytics & AI > Data Science***	
We will then see the OCI Data Science Project created. ***Select our OCI Data Science Project.***	

If we scroll down and select ***Models*** under the Resources section, we can see the Model that was saved under the Model Catalog.	
If we then select ***Model Deployments*** under the Resources section, we can see the Model Deployment that was created from our Model. ***Click on the Model Deployment*** created.	
Here we can find information on the Model Deployment. Click on ***Model Deployment Endpoint*** to see how we can invoke the model.	

You will then see different options on how to invoke the model including the Model Endpoint. ***Copy the Model Endpoint. We will need this later.*** Let's take a look at how you can call the model using Python. Click **Close**.	
Navigate back to your project.	
Navigate to ***Notebook Sessions*** under Resources. Click ***Create notebook session.***	
Give the Notebook Session a ***Name***. Leave all other options as ***Default***. Click ***Create***.	

Once the Notebook Session is up and running, click ***Open***.	
Drag and drop the provided Notebook (*00-invoke-deployed-mistral-model.ipynb*) as part of this Guide into the File Directory. ***Double Click*** on the uploaded Notebook.	
Select the Default ***Python 3*** Kernel.	

Scroll down and ***edit the endpoint parameter*** with the one you copied earlier from our Model Deployment.

Define Request Parameters

```
[8]: # Define Endpoint
     endpoint = '[redacted]

     # Define Header
     headers= {'Content-Type':'application/json', 'Accept': 'text/event-stream'}

     # Define Message Body
     body = {
         "model": "odsc-llm",
         "prompt": "What is Artificial Intelligence?",
         "max_tokens": 250,
         "temperature": 0.7,
         "top_p": 0.8,
     }
```

Then we can scroll back up to the top, ***click on the first cell*** and ***click the >> (Restart Kernel Run all cells) button***.

When prompted, ***Click Restart.***

Launcher | 00-invoke-deployed-mistra | Code

OCI Data Science Notebook

Invoke Deployed NVIDIA NIM Microservice Mistral 7B Instruct

Author: Ismail Syed
Title: EMEA Specialist Leader - Data Science
Company: Oracle

Description: In this notebook we will make a REST API call to our `Mistral-7B-Instruct-v0.3` LLM which we deployed using microservice.

Imports

```
[ ]: import oci
     import json
     import requests
```

We can then scroll down and view the outputs of the run cells and eventually the response of the LLM.

Send Request to Deployed LLM

```
[4]: %%time

     # Send Request
     response = requests.post(endpoint, json=body, auth=auth, headers=headers)

     CPU times: user 12.9 ms, sys: 2.95 ms, total: 15.8 ms
     Wall time: 5.86 s
```

Parse Response

```
[5]: # Decode Response
     decoded_response = response.content.decode('utf-8')

     # Extract Answer
     answer = json.loads(decoded_response)['choices'][0]['text']

     # Print Answer
     print(answer)

     Artificial Intelligence (AI) is a field of computer science that aims to create intelligent machines that can perform tasks th
     at would normally require human intelligence. AI systems can analyze large amounts of data, learn from it, and make decisions
     based on that learning. AI can be used in a variety of applications, including speech recognition, natural language processin
     g, image recognition, and decision-making.

     There are two main types of AI: narrow AI and general AI. Narrow AI is designed to perform a specific task, such as voice reco
     gnition or driving a car. General AI, on the other hand, is designed to perform any intellectual task that a human can do. Gen
     eral AI is still a goal that has not been achieved yet, but researchers are working towards it.

     AI has the potential to revolutionize many industries, including healthcare, finance, and transportation. It can help doctors
     diagnose diseases, financial analysts make investment decisions, and self-driving cars navigate roads. However, it also raises
     ethical and societal concerns, such as job displacement and privacy issues.

     AI is a rapidly evolving field, and new developments are being made all the time. As AI becomes more sophisticated, it has the
     potential to
```

End of Notebook