Ai-API™ Engine

Ai-API™ Engine

Overview

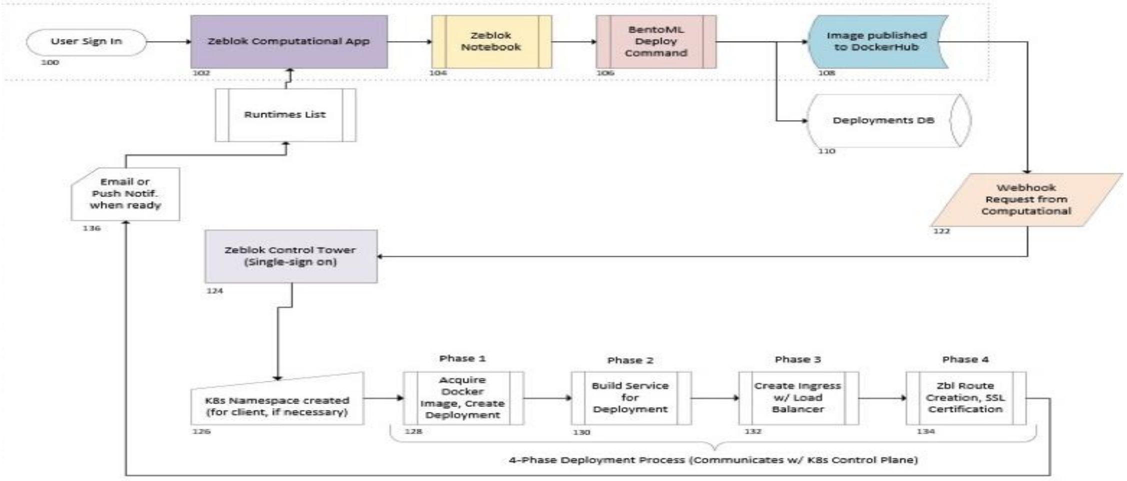
Getting machine learning models into production requires packaging the model as a docker container and translation into a standard application programming interface (API) – it is complex, with many steps. Data scientists are experts in developing and training AI/ML models, but not at building production services and DevOps best practices and they find it challenging to test and deploy trained models. Generally, they choose to hand the task over to a software engineering to avoid a time consuming and error-prone workflow.

Zeblok's Ai-API™ Engine is an Ai-MicroCloud™ framework, which translates a machine learning model into an API and then enables enterprises to deploy and manage each Ai-API™. It bridges the gap between data science and DevOps, enabling fast, repeatable and scalable delivery of prediction services.

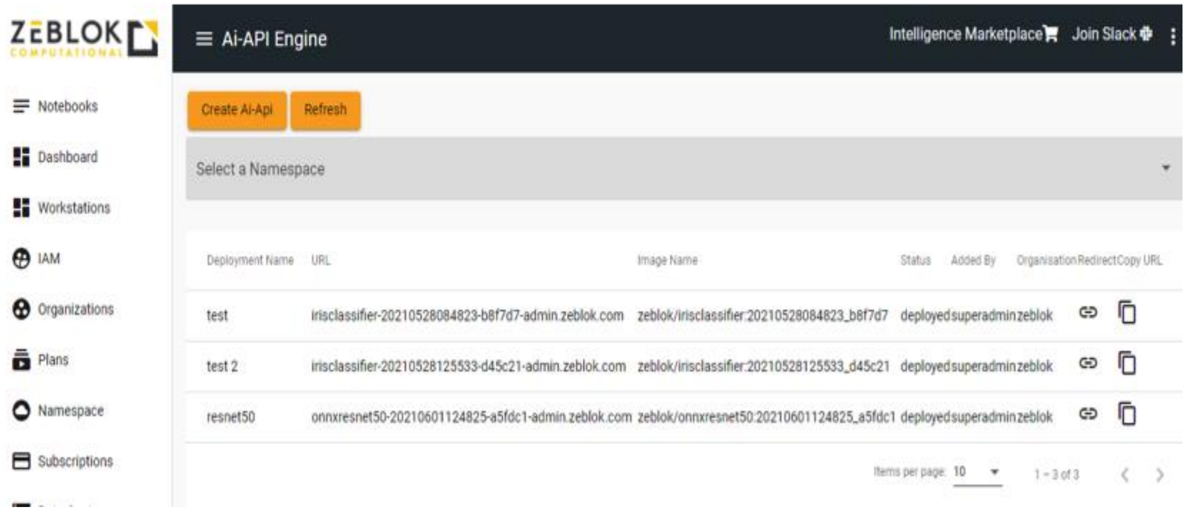
Zeblok's Ai-API™ Engine makes moving trained ML models to production easy, performing the following:

- Package models trained with ML framework and then containerize the model server for production deployment as an Ai-API™
- Deploy Ai-API™ anywhere via either online API serving endpoints or offline batch inference jobs
- High-performance API model server, with adaptive micro-batching support
- Ai-API™ server is able to handle high-volume without crashing, supports multi-model inference, API server Dockerization, Built-in Prometheus metric endpoint, Swagger/Open API endpoint for API Client library generation, serverless endpoint deployment etc.
- Central hub for managing models and deployment process via web UI and APIs
- Supports various ML frameworks including: Scikit-Learn, PyTorch, TensorFlow 2.0, Keras, FastAI v1 & v2, XGBoost, H2O, ONNX, Gluon and more
- Supports API input data types including: DataframeInput, JsonInput, TfTensorflowInput, ImageInput, FileInput, MultifileInput, StringInput, AnnotatedImageInput and more
- Supports API output adapters including: BaseOutputAdapter, DefaultOutput, DataframeOutput, TfTensorOutput and JsonOutput

Complete Workflow of Zeblok Ai-API™ Deployment


User Sign In
100
Zeblok Computational App
102
Zeblok Notebook
104
BentoML Deploy Command
106
Image published to DockerHub
108
Runtimes List
Deployments DB
110
Email or Push Notif. when ready
136
Webhook Request from Computational
122
Zeblok Control Tower (Single-sign on)
124
Phase 1
Phase 2
Phase 3
Phase 4
K8s Namespace created (for client, if necessary)
126
Acquire Docker Image, Create Deployment
128
Build Service for Deployment
130
Create Ingress w/ Load Balancer
132
Zbl Route Creation, SSL Certification
134
4-Phase Deployment Process (Communicates w/ K8s Control Plane)

At the conclusion of the process, each successfully deployed Ai-API™ is listed separately, with its corresponding URL as shown below.

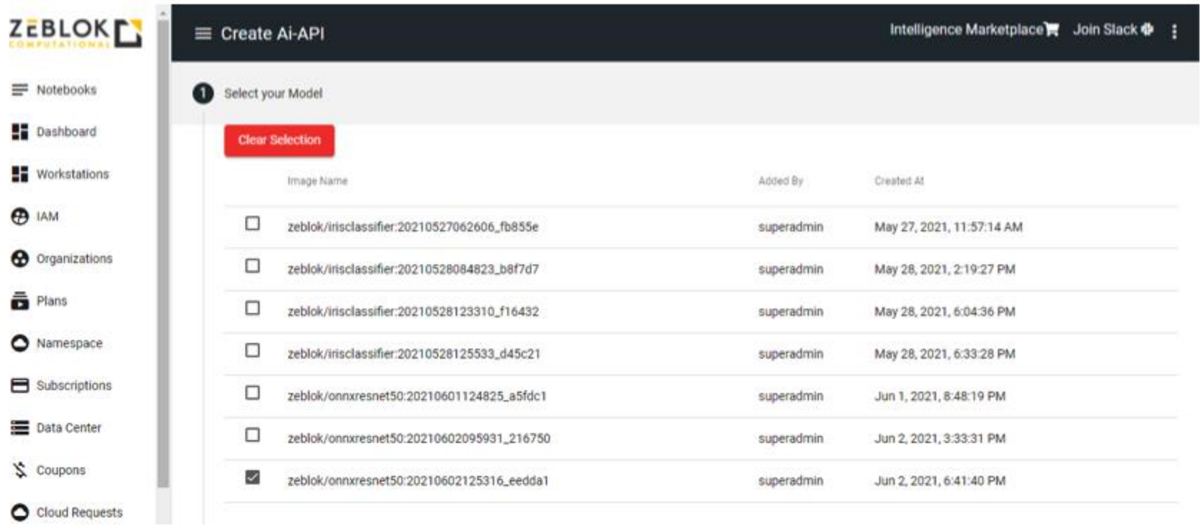


Deployment Name	URL	Image Name	Status	Added By	Organisation	Redirect	Copy URL
test	irisclassifier-20210528084823-b8f7d7-admin.zeblok.com	zeblok/irisclassifier:20210528084823_b8f7d7	deployed	superadmin	zeblok		
test 2	irisclassifier-20210528125533-d45c21-admin.zeblok.com	zeblok/irisclassifier:20210528125533_d45c21	deployed	superadmin	zeblok		
resnet50	onnxresnet50-20210601124825-a5fdc1-admin.zeblok.com	zeblok/onnxresnet50:20210601124825_a5fdc1	deployed	superadmin	zeblok		

Items per page: 10 1 – 3 of 3

Process

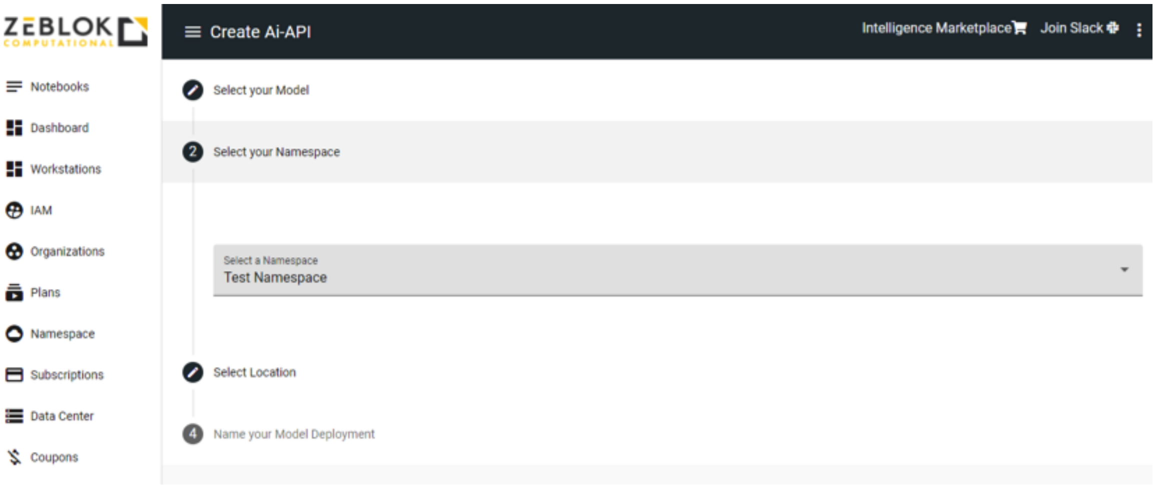
Step 1: Select the model to deploy as an Ai-API™



1 Select your Model

	Image Name	Added By	Created At
□	zeblok/irisclassifier:20210527062606_fb855e	superadmin	May 27, 2021, 11:57:14 AM
□	zeblok/irisclassifier:20210528084823_b8f7d7	superadmin	May 28, 2021, 2:19:27 PM
□	zeblok/irisclassifier:20210528123310_f16432	superadmin	May 28, 2021, 6:04:36 PM
□	zeblok/irisclassifier:20210528125533_d45c21	superadmin	May 28, 2021, 6:33:28 PM
□	zeblok/onnxresnet50:20210601124825_a5fdc1	superadmin	Jun 1, 2021, 8:48:19 PM
□	zeblok/onnxresnet50:20210602095931_216750	superadmin	Jun 2, 2021, 3:33:31 PM
☑	zeblok/onnxresnet50:20210602125316_eedda1	superadmin	Jun 2, 2021, 6:41:40 PM

Step 2: Option to select the namespace



Select your Model

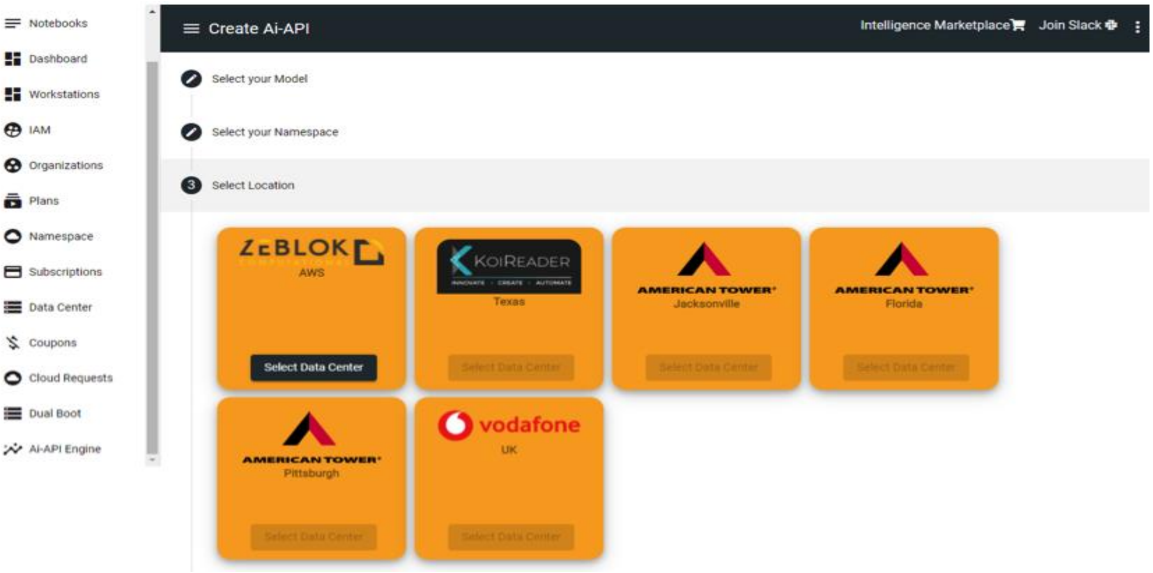
2 Select your Namespace

Select a Namespace: Test Namespace

Select Location

4 Name your Model Deployment

Step 3: Select the data center locations where model is to be deployed

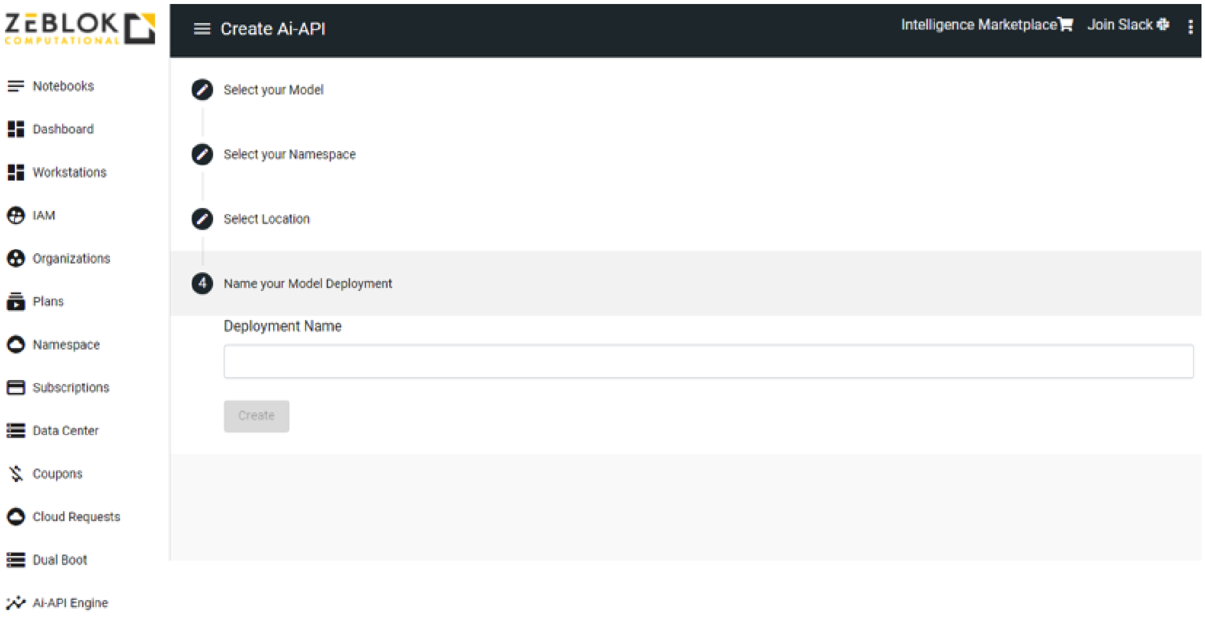


Select your Model

Select your Namespace

3 Select Location

Step 4: Enter a unique name for the Ai-API™ deployment



Select your Model

Select your Namespace

Select Location

4 Name your Model Deployment

Deployment Name

For more information: email Mouli Narayanan

Zeblok Computational Inc.
1500 Stony Brook Road
Stony Brook, NY 11794
www.zeblok.com
mouli.narayanan@zeblok.com
Phone: +1 (631) 223-8233