avanade

Implantando modelos Deep Learning em cluster Kubernetes com GPU Ativada

Thaissa Bueno Sanches - Consultant at Avanade

Agenda

- Noções básicas de implantação (carga útil, lotes, HTTP, Web Service)
- Comparação de GPU / CPU para inferência
- Kubernetes
- Etapas comuns
- Implantação no Kubernetes usando o Kubectl
- Implantação no Kubernetes usando o AzureML
- Implantação no Kubernetes usando Kubeflow e "TF serving"

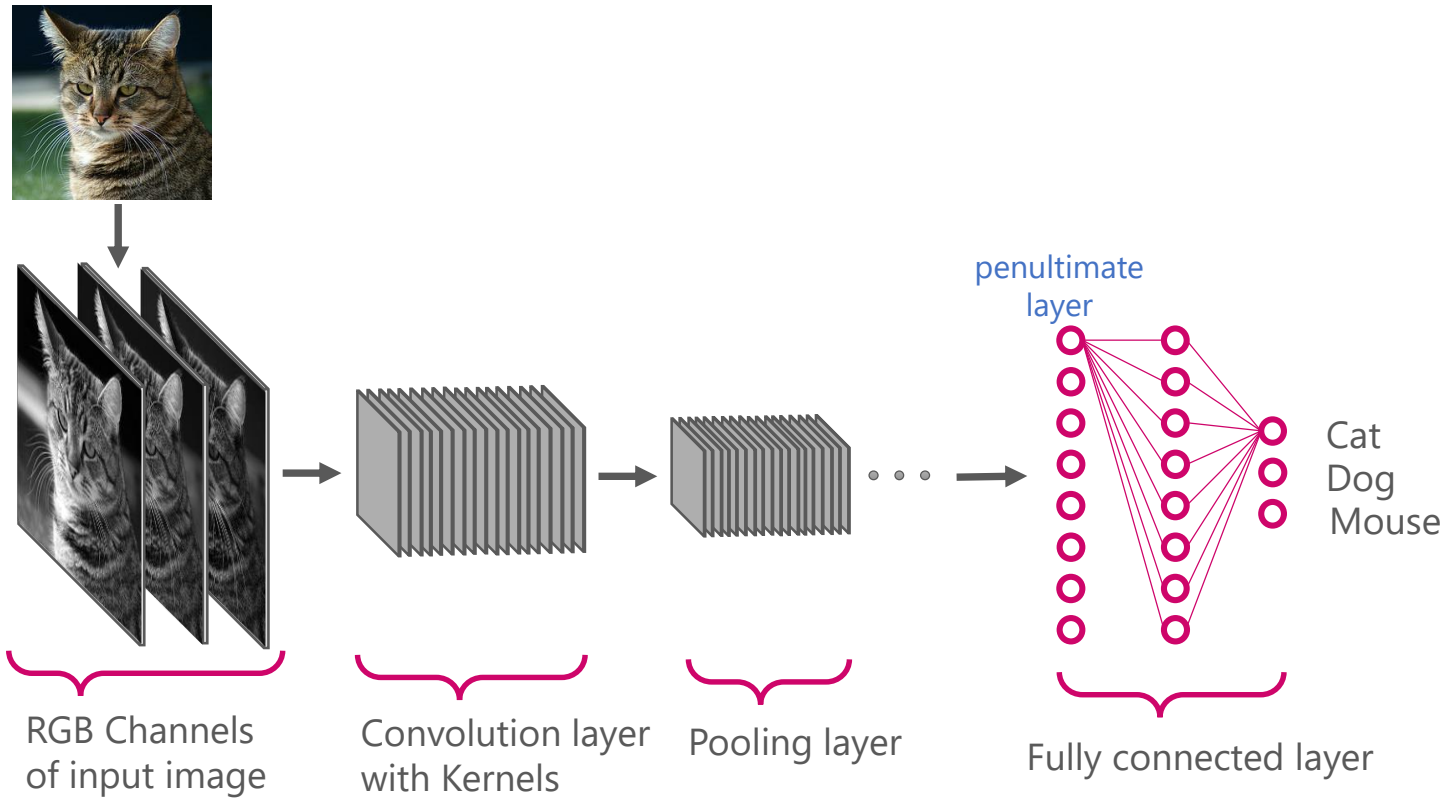
avanade

Noções básicas de implantação (payload, batching, HTTP, Web Service)

- O que é implantar?

Noções básicas de implantação (payload, batching, HTTP, Web Service)

- Payload



Noções básicas de implantação (payload, batching, HTTP, Web Service)

- Encode em base64

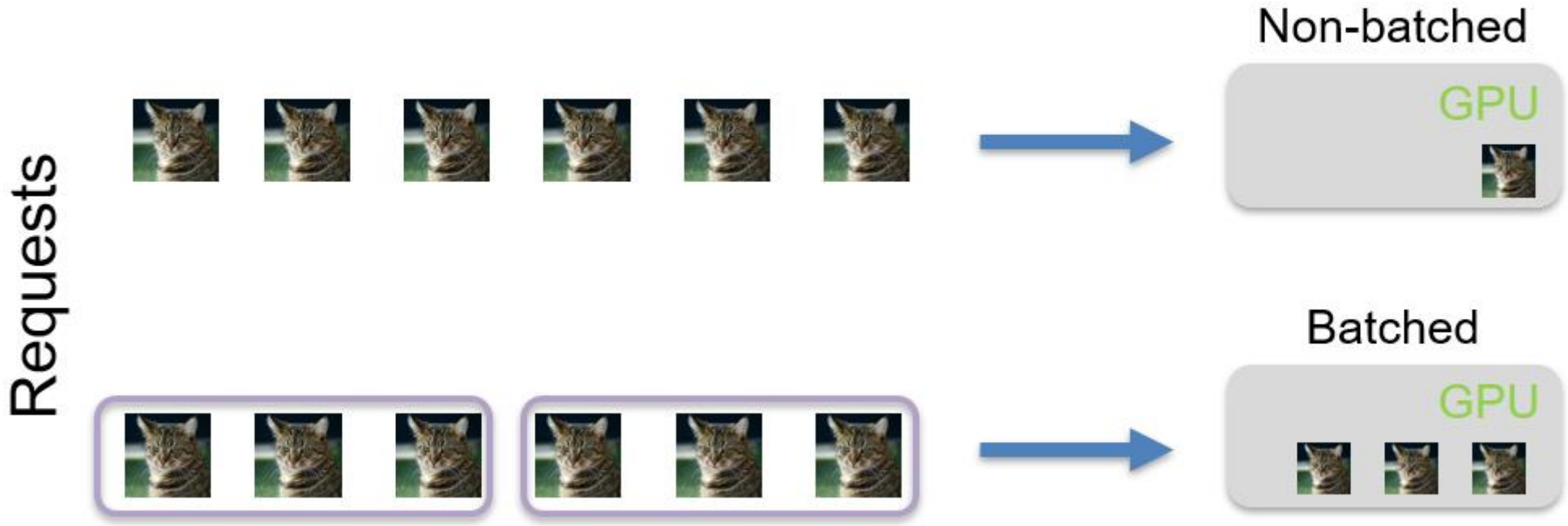
```
decoded_img = base64.b64decode(request.json["input"])
img_buffer = BytesIO(decoded_img)
pil_img = Image.open(img_buffer).convert("RGB")
```

- Transfere arquivo

```
Image.open(request.files['image']).convert("RGB")
```

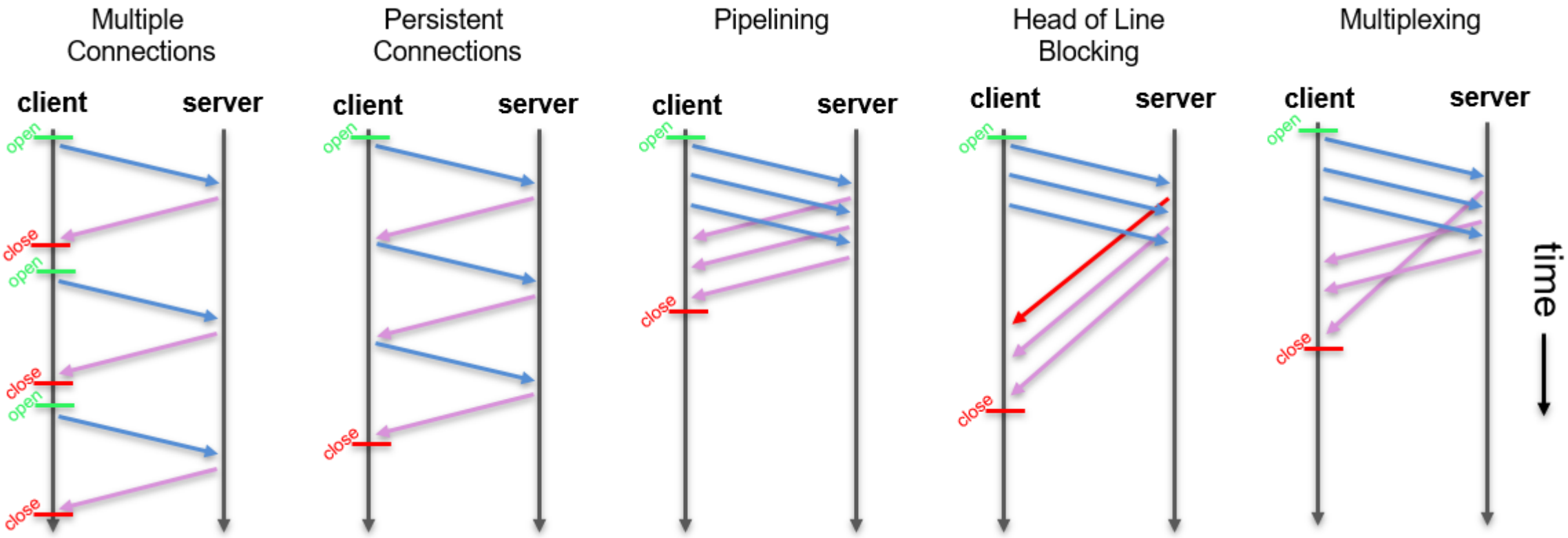
Noções básicas de implantação (payload, batching, HTTP, Web Service)

- Batching



Noções básicas de implantação (payload, batching, HTTP, Web Service)

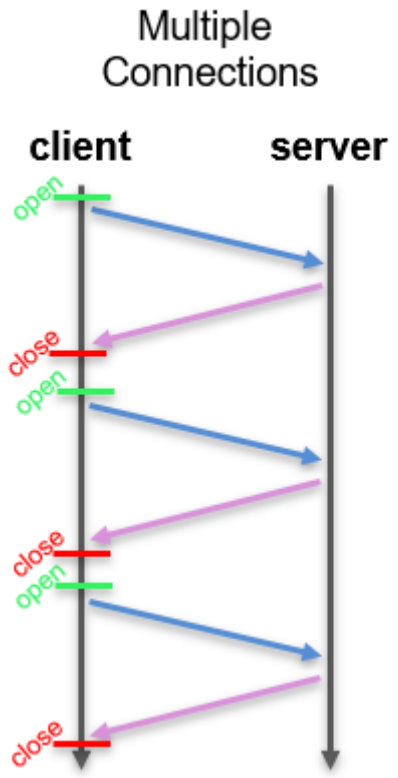
- HTTP



Noções básicas de implantação (payload, batching, HTTP, Web Service)

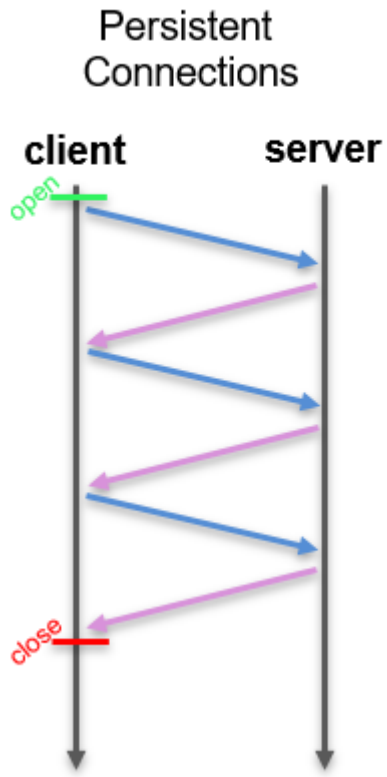
- Múltiplas conexões:

Uma conexão é aberta e somente é aberta uma nova conexão quando a anterior for fechada



Noções básicas de implantação (payload, batching, HTTP, Web Service)

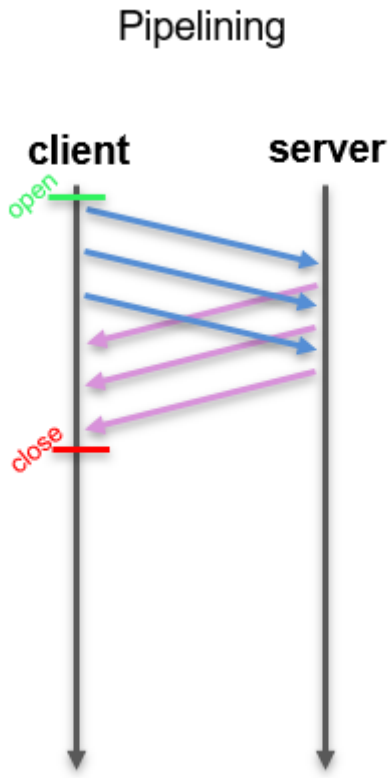
- Conexões persistentes:
- Uma conexão é aberta e tem seu estado persistido até o termino de sua utilização.



Noções básicas de implantação (payload, batching, HTTP, Web Service)

- Pipelining:

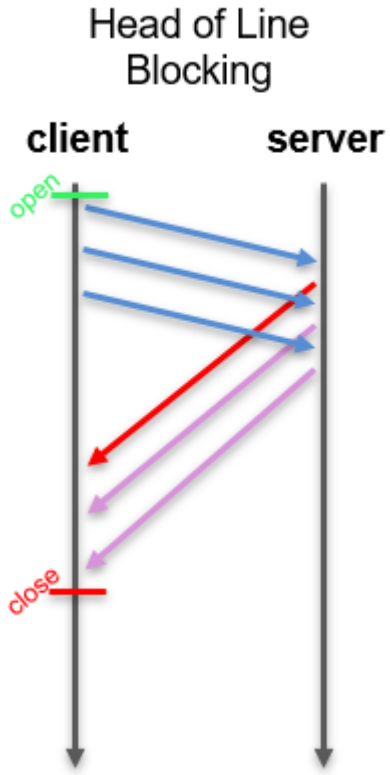
Em apenas uma única solicitação TCP varias conexões HTTP são enviadas ao mesmo tempo.



Noções básicas de implantação (payload, batching, HTTP, Web Service)

- Head of line blocking:

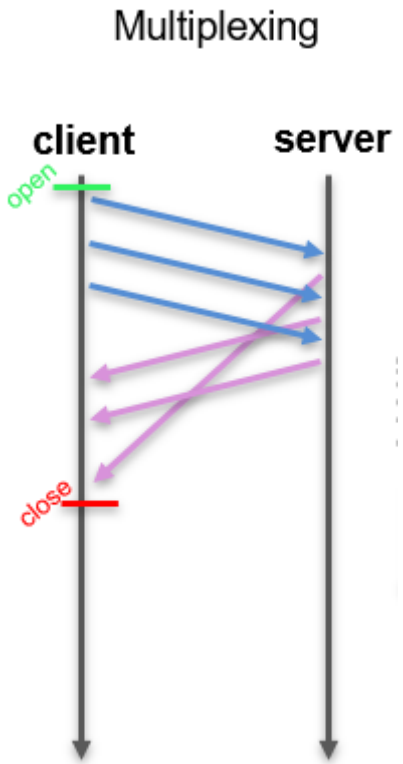
Em apenas uma única solicitação TCP varias conexões HTTP são enviadas ao mesmo tempo porem uma ou mais solicitações podem ser bloqueada para aguardar o termino da anterior por ter excedido o limite de requests paralelos.



Noções básicas de implantação (payload, batching, HTTP, Web Service)

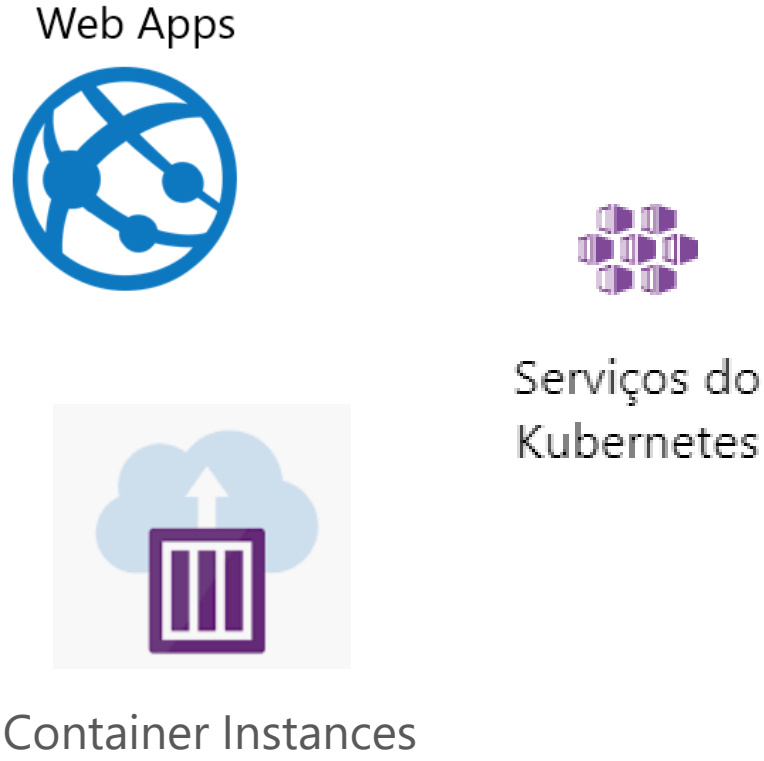
- Multiplexing:

A multiplexação é um método no HTTP / 2 pelo qual várias solicitações HTTP podem ser enviadas e as respostas podem ser recebidas de forma assíncrona por meio de uma única conexão TCP. A multiplexação é o coração do protocolo HTTP / 2



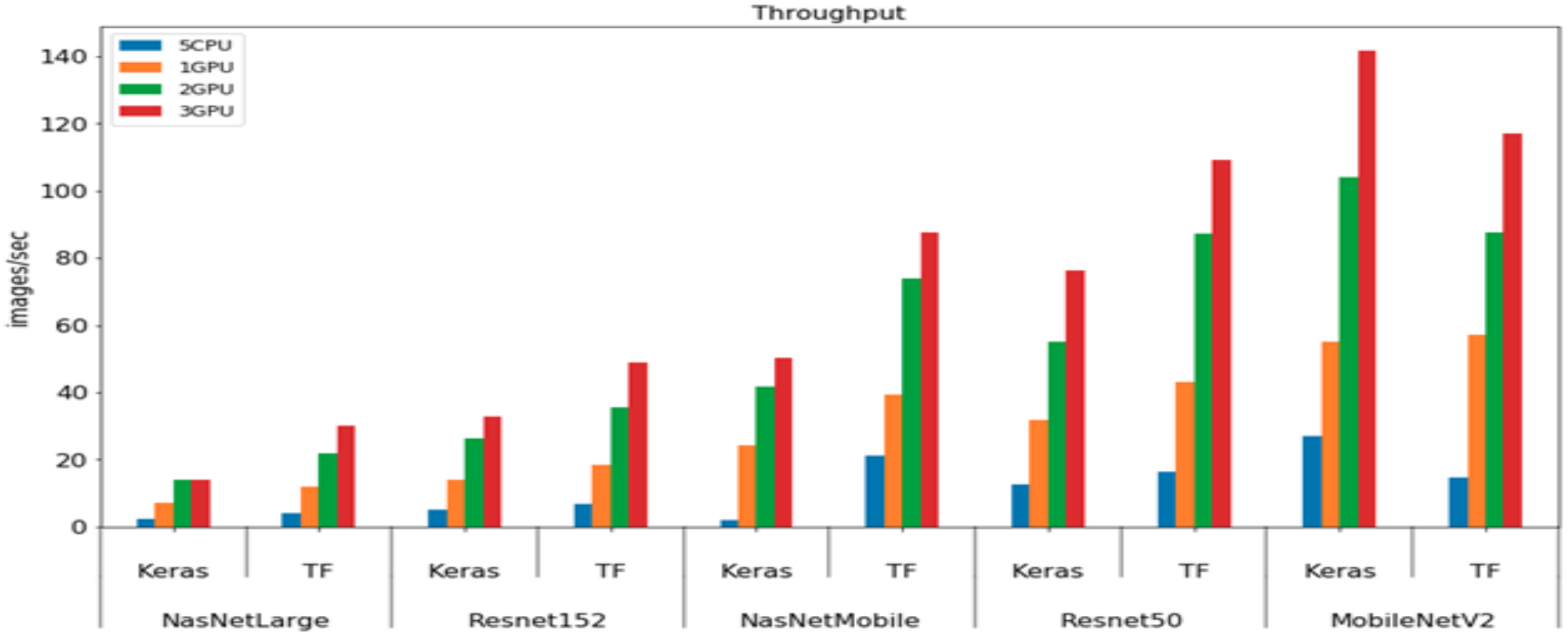
Noções básicas de implantação (payload, batching, HTTP, Web Service)

- Web Service:
- É uma solução para integração de sistemas e comunicação com aplicações diferentes, podendo utilizar o protocolo de comunicação SOAP (Simple Object Access Protocol, em português Protocolo Simples de Acesso a Objetos) ou a arquitetura REST(Representational State Transfer em português Transferência de Estado Representacional).

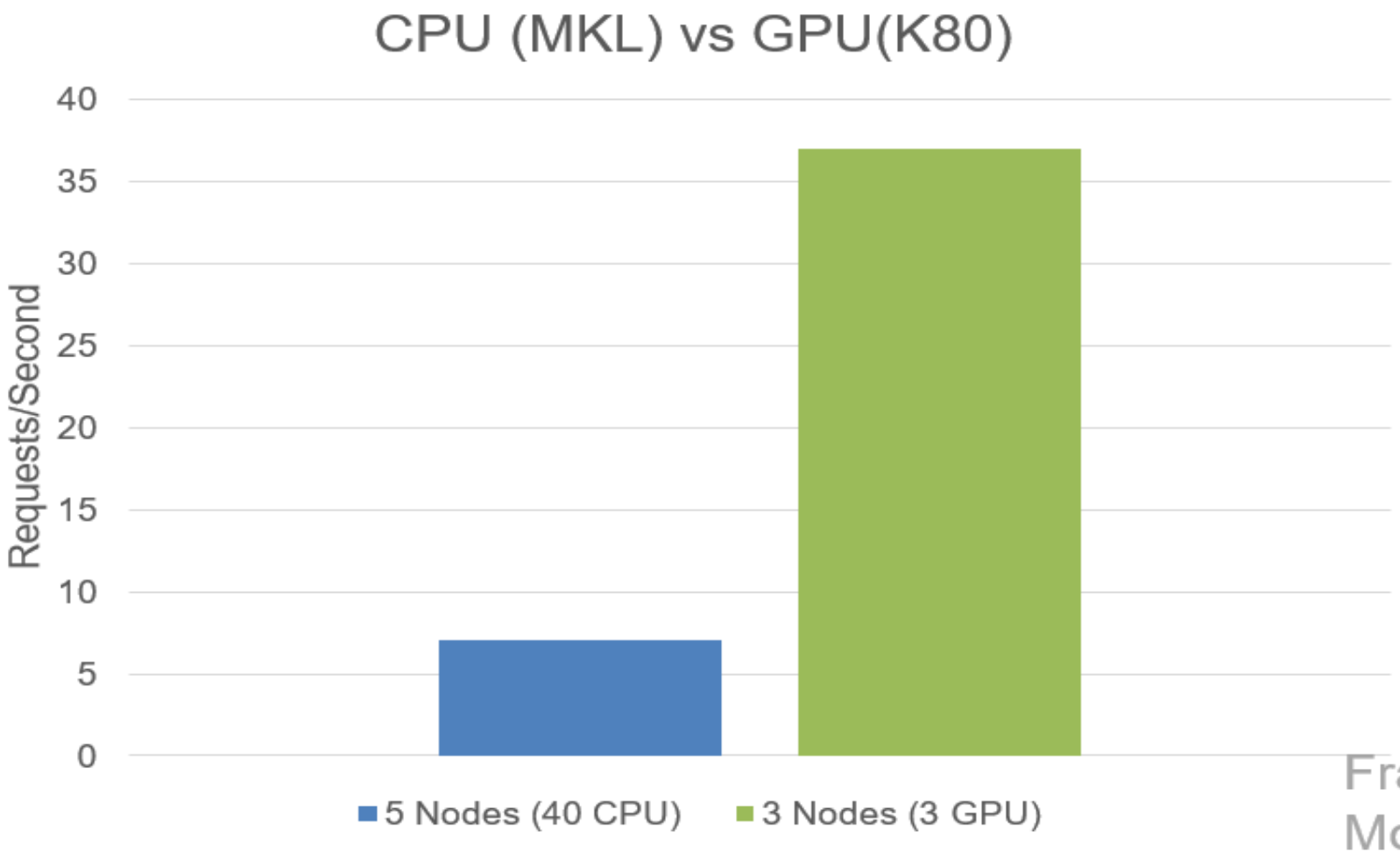


Comparação de GPU / CPU para inferência

- CPU VS GPU



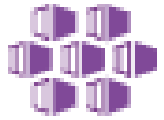
Comparação de GPU / CPU para inferência



Framework: Keras
Model: ResNet152

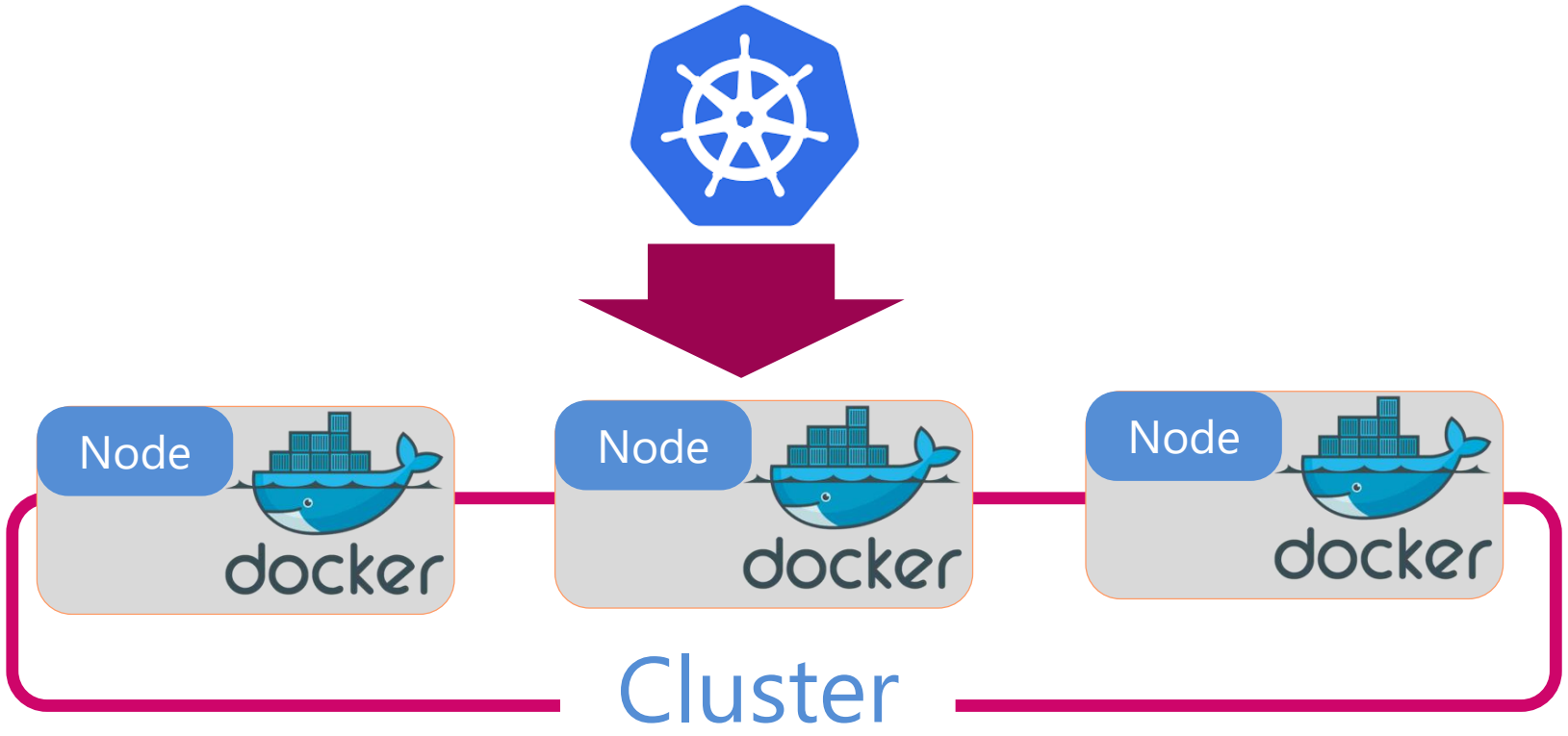
Kubernetes

- Kubernetes é um sistema de orquestração de contêiners open-source que automatiza a implantação, o dimensionamento e a gestão de aplicações em contêiners.



Serviços do Kubernetes

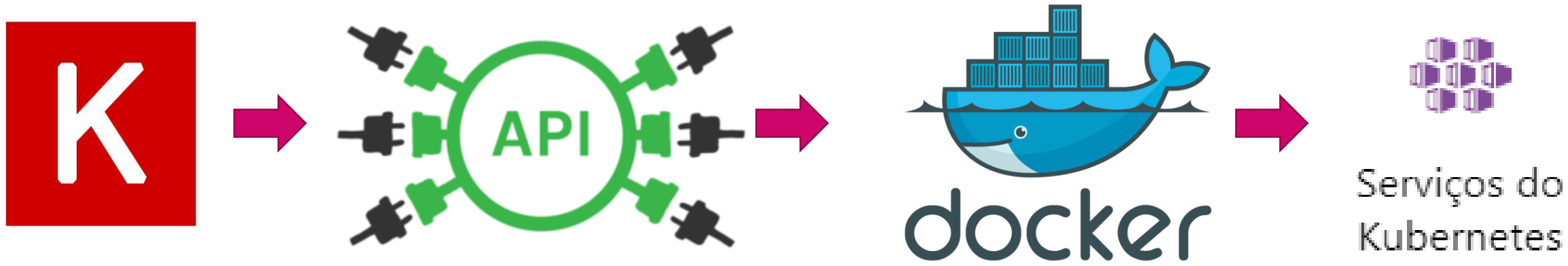
Kubernetes



avanade

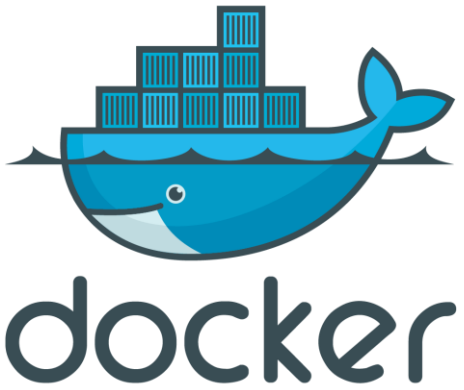
Etapas comuns

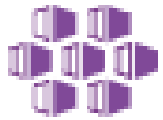
- Desenvolver modelo
- Desenvolver API de modelo
- Preparar contêiner docker para o serviço da web
- Implantar no Kubernetes



Implantação no Kubernetes usando o Kubectl

- Desenvolver modelo, modelo API, Flask App
- Crie uma imagem de contêiner com modelo, modelo API e Flask App
- Teste localmente e envie a imagem para o Docker Hub
- Provisionar o Cluster Kubernetes
- Conecte-se ao Kubernetes com kubectl
- Implantar aplicativo usando o manifesto (.yaml)

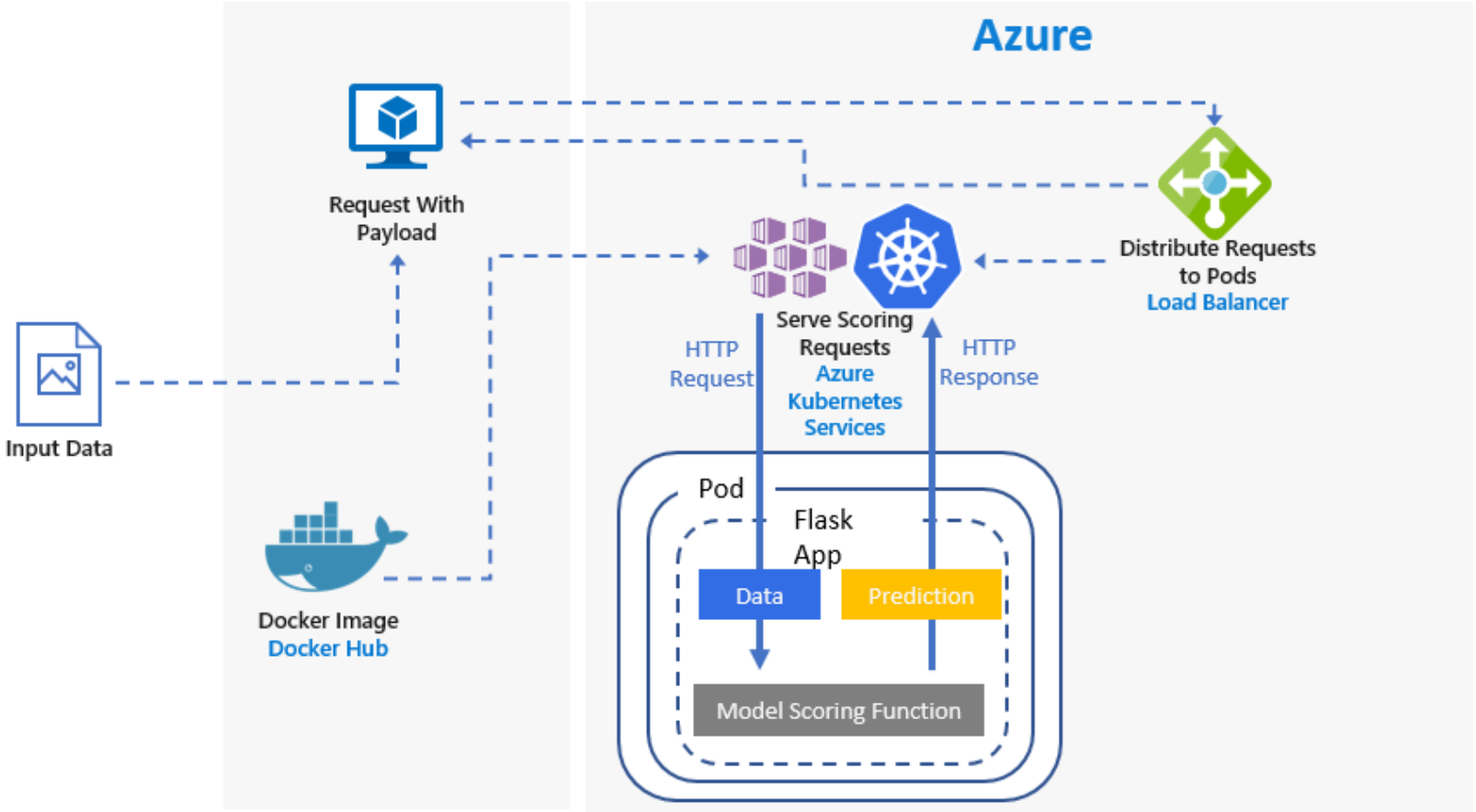




Serviços do Kubernetes

avanade

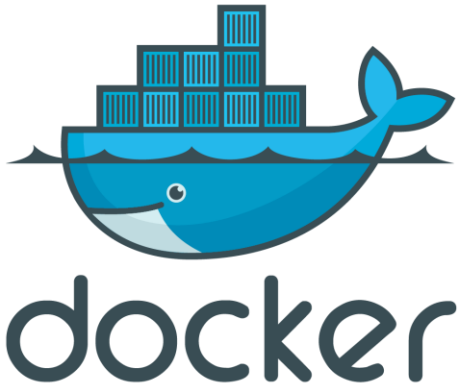
Implantação no Kubernetes usando o Kubectl

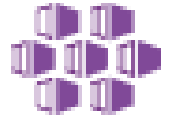


https://github.com/Microsoft/AKSDeploymentTutorial.git

Implantação no Kubernetes usando o AzureML

- Configuração do AzureML, modelo de desenvolvimento, registro do modelo no espaço de trabalho do AzureML, desenvolvimento da API do modelo (script de pontuação)
- Crie imagens com o AzureML usando dependências do conda, requisitos de pip e outras dependências
- Capturar imagem do Azure Container Registry (ACR) e testar localmente
- Provisione o cluster do Kubernetes e implemente o serviço da Web com o AzureML.

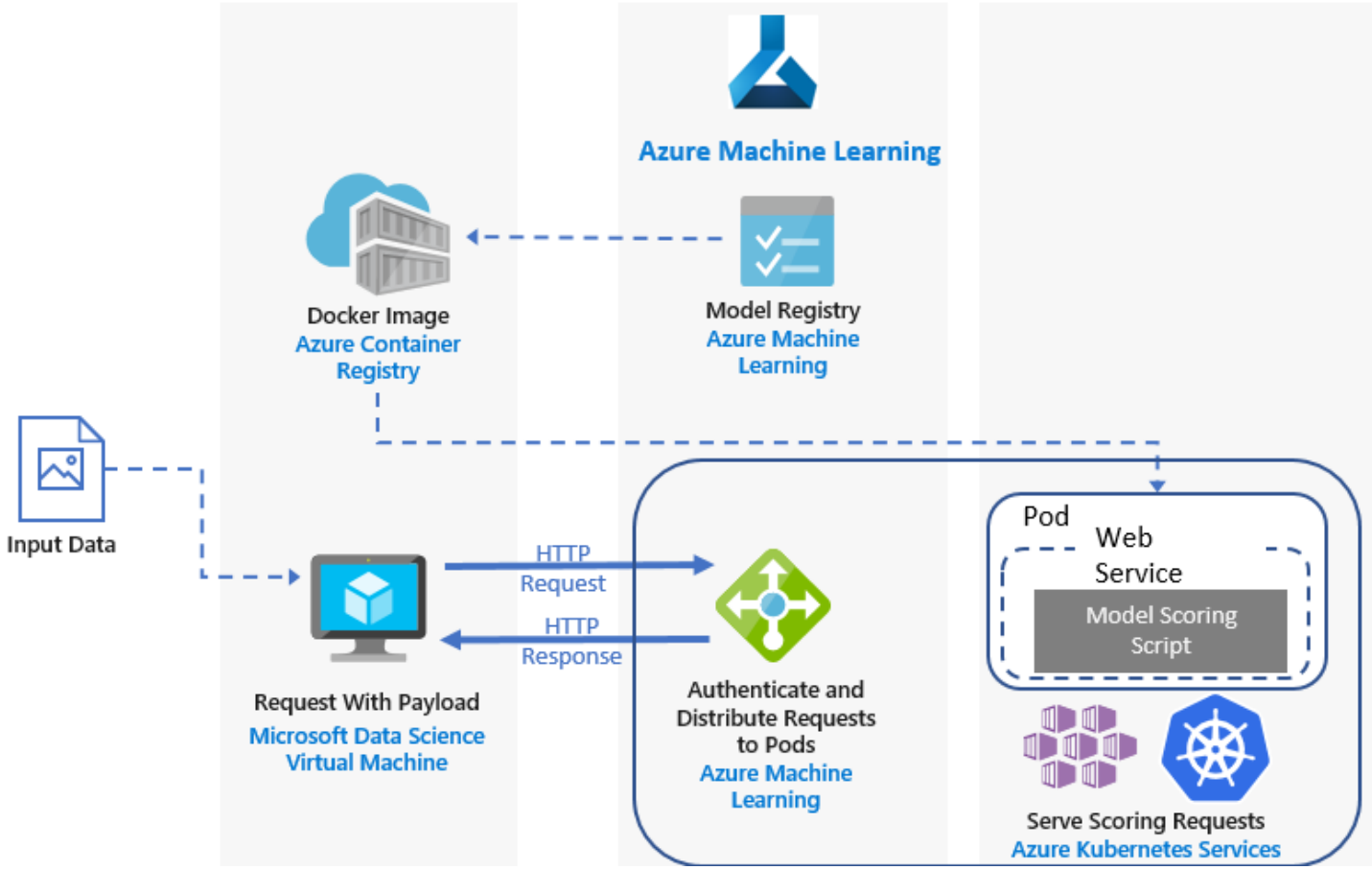




Serviços do Kubernetes

avanade

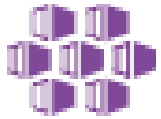
Implantação no Kubernetes usando o AzureML



https://github.com/Microsoft/AKSDeploymentTutorialAML.git

Implantação no Kubernetes usando o AzureML

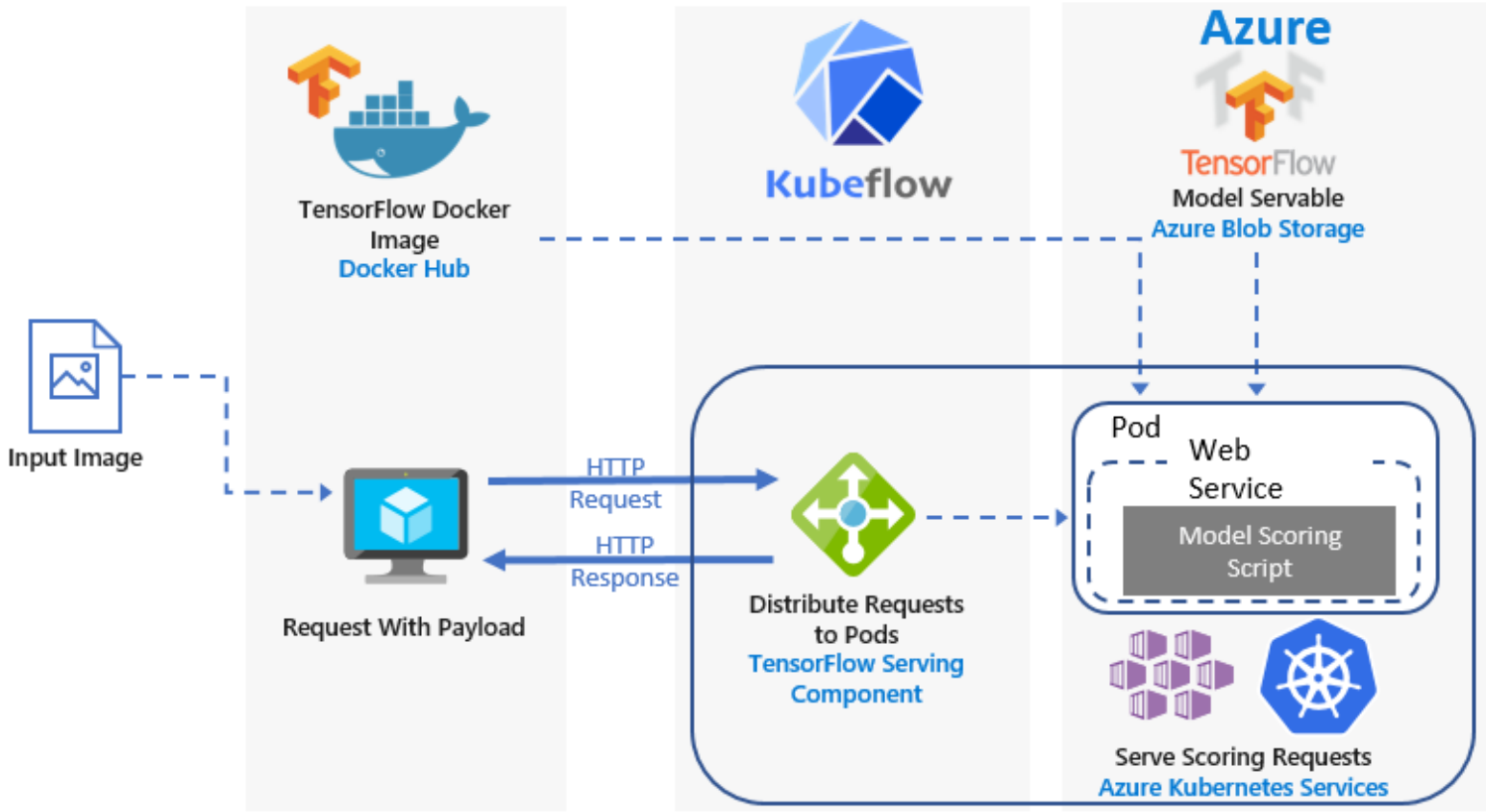
- Desenvolva modelo e salve como TensorFlow servable
- Puxe a imagem TensorFlow Serving do Docker hub, monte o caminho do modelo, abra a porta da API REST, teste localmente
- Crie um Kubernetes cluster, anexe o armazenamento blob no AKS, copie o modelo servable
- Instale o Ksonnet, Kubeflow e implante o webservice usando o componente Kubeflow TensorFlow serving usando "ksonnet template".



Serviços do Kubernetes

avanade

Implantação no Kubernetes usando o AzureML



https://github.com/Microsoft/DeployDLKubeflowAKS.git

Participe de um treinamento **GRATUITO** de Azure Machine Learning Service

bit.ly/azuremltdc

avanade

©2018 Avanade Inc. All Rights Reserved.

Contatos

Site/Blog/Email:
http://www.thaissasanches.com.br
http://meetup.com/pt-BR/DevelopersBR/
https://meetup.com/pt-BR/ai-brasil/

Redes Sociais:
Linkedin: /in/thaissa-bueno-sanches
Github: thayssa1186
Twitter: thayssa1186

Vagas Avanade:
https://careers.avanade.com/jobsenusurl/SearchJobs/?3_56_3=19753

Quem sou eu?

- Formada em tecnologia em rede de computadores pela UNIVEM/Marilia.
- Especialista em desenvolvimento .NET e Java.
  Pós Graduanda Machine Learning e Deep Learning na IGTI.
- Consultora de TI e Arquiteta
- Faço parte da coordenação do evento TDC – Trilha de IA
- Organizadora do AIFest 2018
- Uma das coordenadoras do Developers BR e IA Brasil além de varias comunidades de tecnologia que participo.

Obrigada

Thaissa Bueno Sanches
Consultant at Avanade

avanade