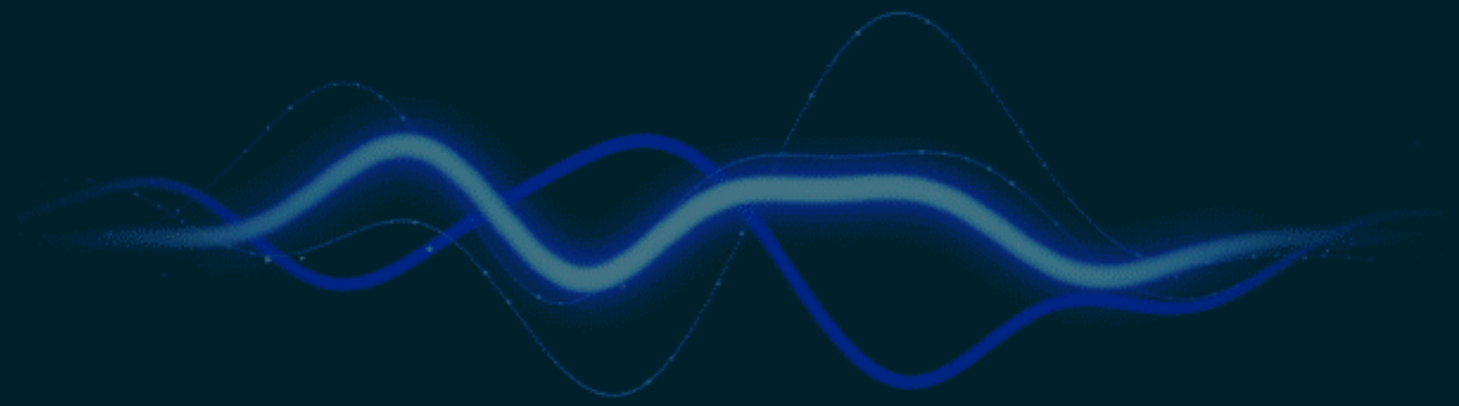
# Implantando modelos de Deep Learning em Azure Container Instance




avanade

# Implantando modelos de Deep Learning em Azure Container Instance

Thaissa Bueno Sanches - Consultant at Avanade

# Azure Machine Learning service

# WHAT IS AZURE MACHINE LEARNING SERVICE?

Set of Azure Cloud Services

Python SDK

That enables you to:

- ✓ Prepare Data
- ✓ Build Models
- ✓ Train Models
- ✓ Manage Models
- ✓ Track Experiments
- ✓ Deploy Models

avanade

# Azure Machine Learning service feature

**Workspace** – Top level container for model management and experimentation. Also, creates and manages storage, container registry, key vault and app insights

**Experiments** – Grouping of training runs for a given script

**Pipelines** – Stitch together multiple stages to create machine learning workflows

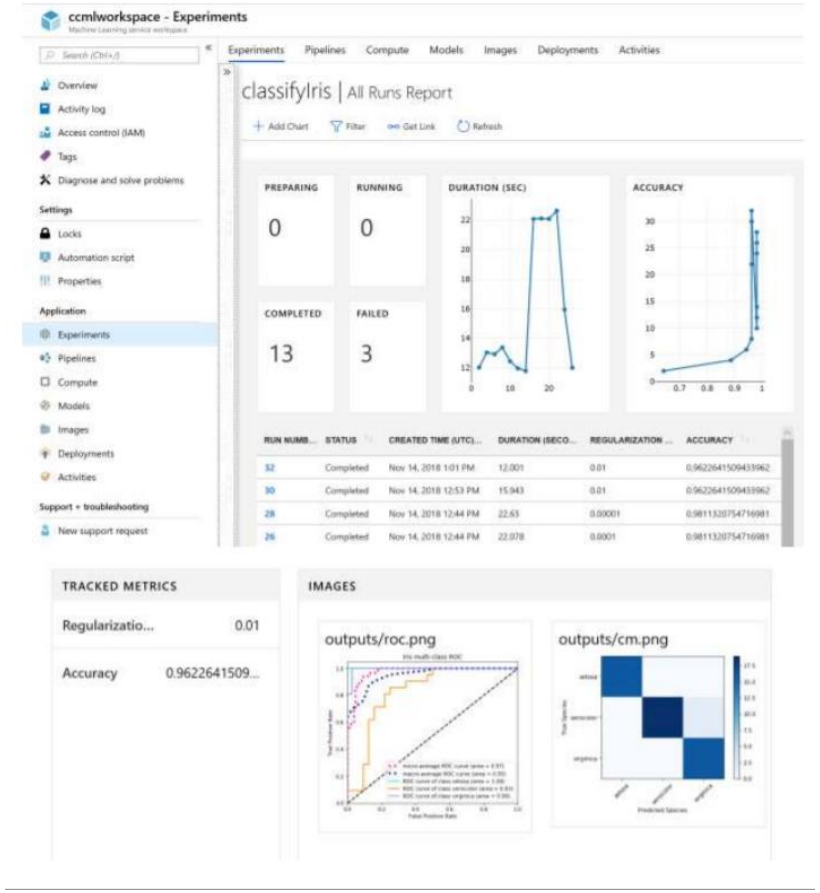
**Compute** – Compute resources for training or image deployment

**Models** – Encapsulation of trained models

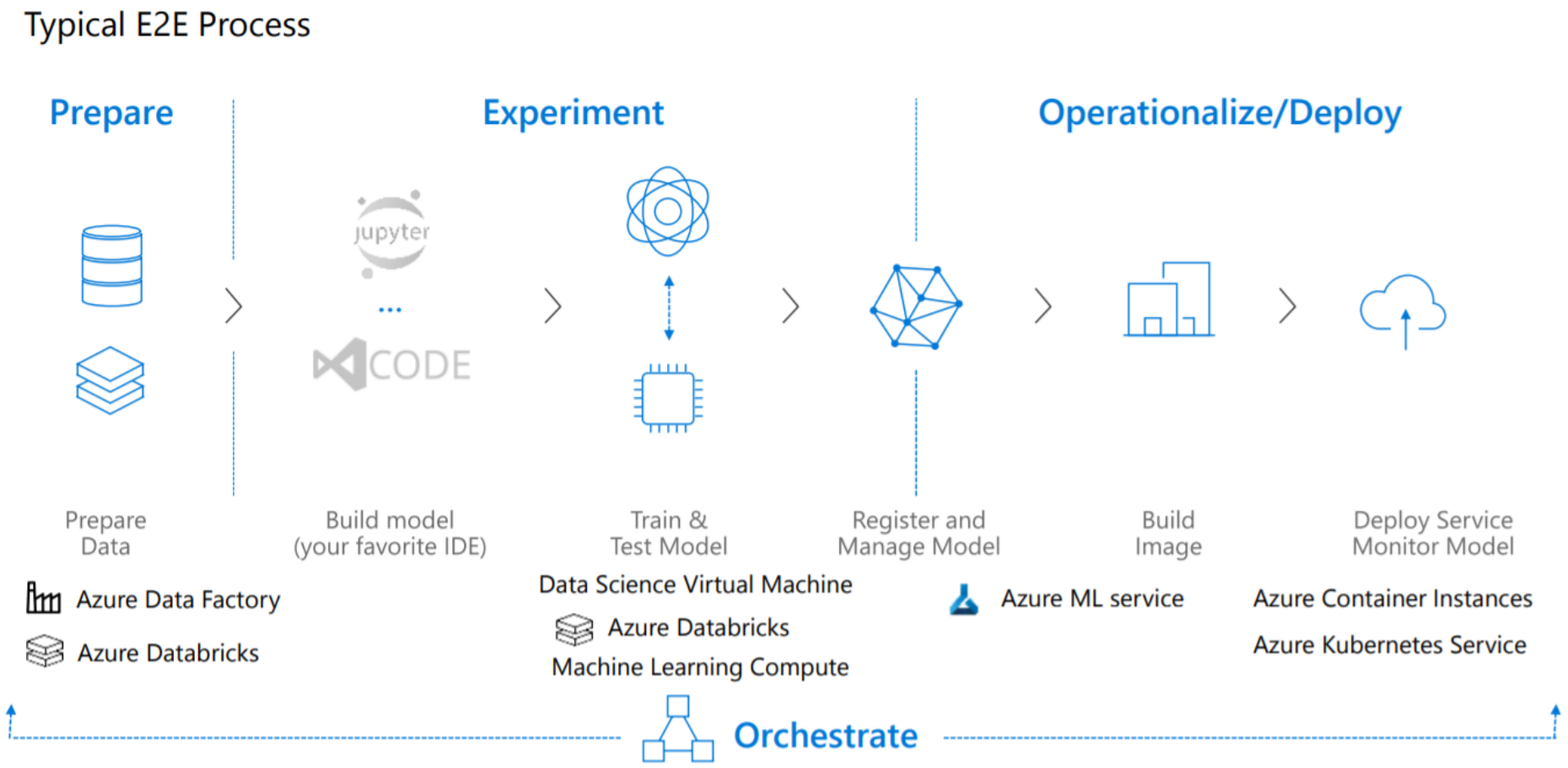
**Images** – Docker images containing trained models to be **deployed** to a compute target

**Deployments** – Deployment of an image to web services, IoT Edge and FPGAs

**Data Stores** – Abstraction layer over Azure storage



# Custom AI



# Azure ML Concept – Model Management

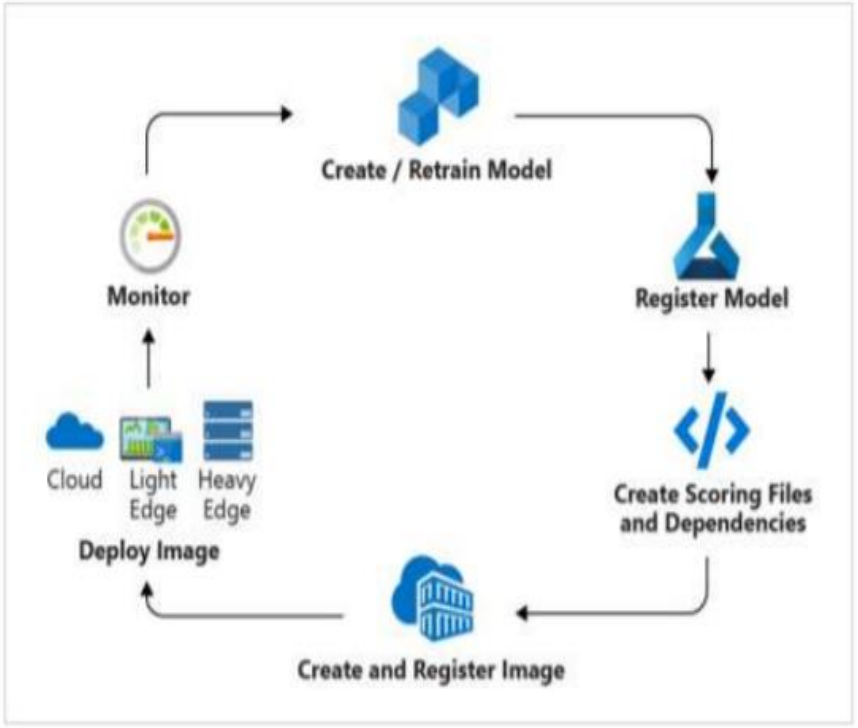
**Model Management in Azure ML usually involves these four steps**

**Step 1:** Register Model using the Model Registry

**Step 2:** Register Image using the Image Registry (the Azure Container Registry)

**Step 3:** Deploy the Image to cloud or to edge devices

**Step 4:** Monitor models—you can monitor input, output, and other relevant data from your model.



# Azure ML Concept – Computer Target

Compute Targets are the compute resources used to run training scripts or host your model when deployed as a web service.

They can be created and managed using the Azure Machine Learning SDK or CLI.

You can attach to existing resources.

You can start with local runs on your machine, and then scale up and out to other environments.

## Currently supported compute targets

| Compute Target | Training | Deployment |
|---|---|---|
| Local Computer | ✓ | |
| A Linux VM in Azure (such as the Data Science Virtual Machine) | ✓ | |
| Azure ML Compute | ✓ | |
| Azure Databricks | ✓ | |
| Azure Data Lake Analytics | ✓ | |
| Apache Spark for HDInsight | ✓ | |
| Azure Container Instance | | ✓ |
| Azure Kubernetes Service | | ✓ |
| Azure IoT Edge | | ✓ |
| Field-programmable gate array (FPGA) | | ✓ |

# Azure ML Concept – Image Registry

## Image contains

1. A model.
2. A scoring script used to pass input to the model and return the output of the model.
3. Dependencies needed by the model or scoring script/application.

## Two types of images

1. **FPGA image**: Used when deploying to a field-programmable gate array in the Azure cloud.
2. **Docker image**: Used when deploying to compute targets such as Azure Container Instances and Azure Kubernetes Service.

## Image Registry

Keeps track of images created from models.

Metadata tags can be attached to images. Metadata tags are stored by the image registry and can be used in image searches

# Azure ML Concept – Deployment

**Deployment is an instantiation of an image. Two options:**

## Web service

A deployed web service can run on Azure Container Instances, Azure Kubernetes Service, or field-programmable gate arrays (FPGA).

Can receive scoring requests via an exposed a load-balanced, HTTP endpoint.

Can be monitored by collecting Application Insight telemetry and/or model telemetry.

Azure can automatically scale deployments.

## IoT Module

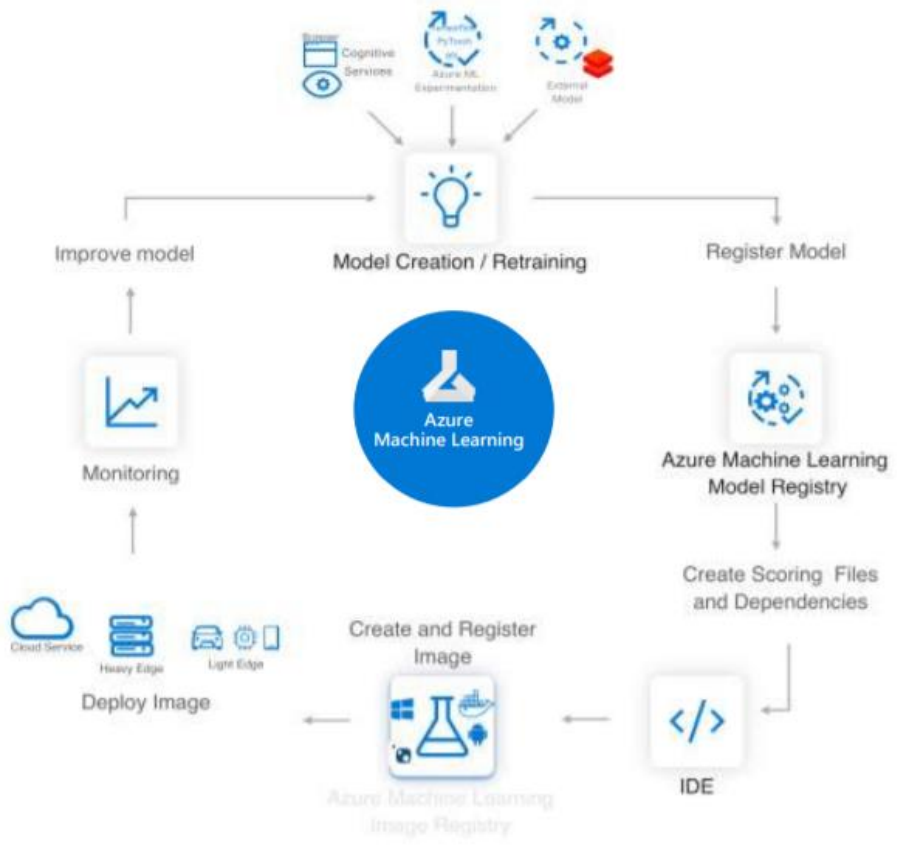
A deployed IoT Module is a Docker container that includes the model, associated script and additional dependencies.

Is deployed using **Azure IoT Edge** on edge devices.

Can be monitored by collecting Application Insight telemetry and/or model telemetry.

Azure IoT Edge will ensure that your module is running and monitor the device that is hosting it.

# Azure ML service
# Lets you easily implement this AI/ML Lifecycle




## Workflow Steps

Develop machine learning training scripts in Python.

Create and configure a compute target.

Submit the scripts to the configured compute target to run in that environment. During training, the compute target stores run records to a datastore. There the records are saved to an experiment.

Query the experiment for logged metrics from the current and past runs. If the metrics do not indicate a desired outcome, loop back to step 1 and iterate on your scripts.

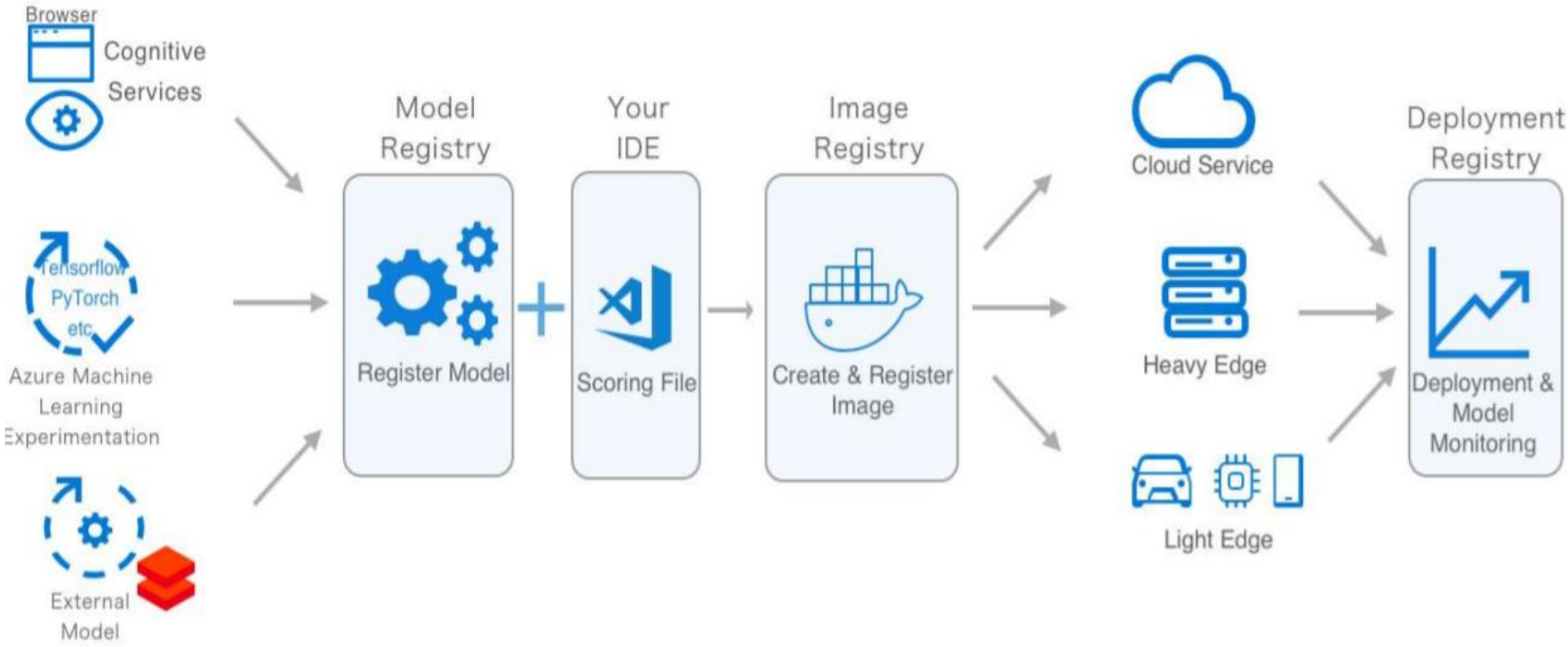
Once a satisfactory run is found, register the persisted model in the model registry.

Develop a scoring script.

Create an Image and register it in the image registry.

Deploy the image as a web service in Azure.

# Azure ML: How to deploy models at scale



# Productive Services

| Customer journey | Data Prep | Build and Train | Manage and Deploy |
| --- | --- | --- | --- |
| Python ML developer |  **Azure ML service** (Pandas, NumPy etc. on AML Compute) |  **Azure ML service** (OSS frameworks, Hyperdrive, Pipelines, Automated ML, Model Registry) |  **Azure ML service** (containerize, deploy, inference and monitor) |
| Apache Spark / Big Data |  **Azure Databricks** (Apache Spark Dataframes, Datasets, Delta, etc.) |  **Azure Databricks + Azure ML service** (Spark MLib and OSS frameworks + Automated ML, Model Registry) |  **Azure ML service** (containerize, deploy, inference and monitor) |

avanade

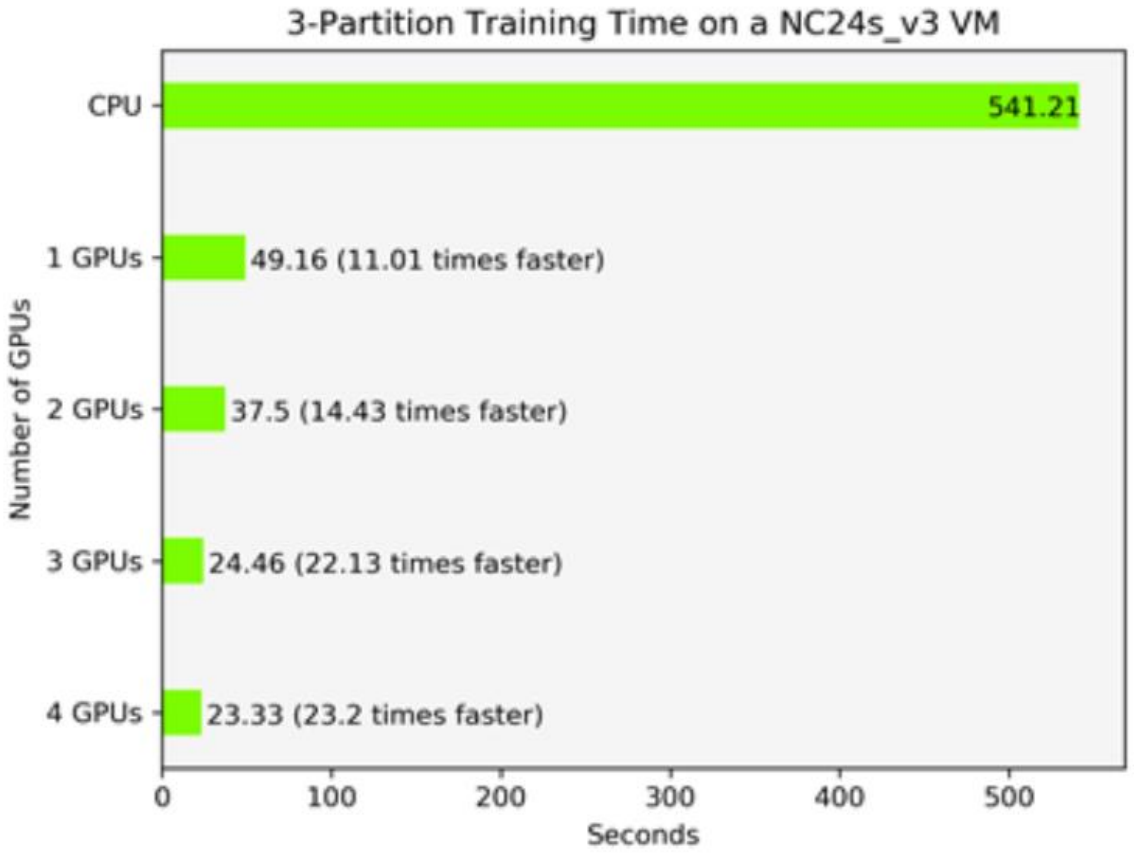
# Productive Services

| Comparable Table | | |
|---|---|---|
| | Azure Machine Learning Studio | Azure Machine Learning Services |
| Pros | • Rapid development (Drag and Drop)<br>• Works well with relatively simple datasets<br>• Pre-built ML algorithms<br>• Cheap | • Fast (VMs with GPUs)<br>• Different optimization methods, CI/CD pipeline<br>• Full control during training<br>• Manage computing resources (choose VM size)<br>• Use open source ML libraries |
| Cons | • Can be slow<br>• Limited optimization methods, operationalized architecture<br>• Less control during training<br>• Fixed computing resources | • More elaborate to build, require deeper knowledge of machine learning<br>• Deeper models need much more data with much more memory<br>• Higher costs for VM with GPU |

avanade

# CPU or GPU

# CPU VS GPU




avanade


# Deployment in Azure Container Instance

# ACI Deployment

```
from azureml.core.webservice import AciWebservice

aciconfig = AciWebservice.deploy_configuration(cpu_cores=1,
                                               memory_gb=1,
                                               tags = {"data": "textSentiment", "type": "classification"},
                                               description='Analise de sentimento')
```

```
%%time
from azureml.core.webservice import Webservice
from azureml.core.image import ContainerImage

# configure the image
image_config = ContainerImage.image_configuration(execution_script="score.py",
                                                   runtime="python",
                                                   conda_file="ccaanlsentimentdependence.yml")

service = Webservice.deploy_from_model(workspace=ws,
                                       name='ccaanlsentiment2-svc',
                                       deployment_config=aciconfig,
                                       models=[model],
                                       image_config=image_config)

service.wait_for_deployment(show_output=True)
```

avanade

avanade

# Demo

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

bit.ly/azuremltdc

avanade

# 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