MLops voor Data Engineers

Opzet, programma en planning

# Opzet

In één dag willen we de basis van MLOps meegeven. Dat is ambitieus: één dag is kort. Deelnemers hebben zich daarom voorbereid op het gebied van containers, om te snappen hoe dit werkt, en wat voor flexibiliteit dit geeft.

In de training zullen we niet verdergaan met containers per sé, maar richten we ons op het doen van MLOps met behulp van Azure Machine Learning.

## Bedoeling

Bij Achmea worden containers gehost in OpenShift, en artifacts in Artifactory. In het oorspronkelijke overleg met Leon hebben we hier het volgende over gezegd:

*Belangrijk: niet één programmeertaal, maar programmeer\*vaardigheden\*:*

*- SQL, Python, R? Keuze is aan jou uiteindelijk*

*Geldt ook voor OpenShift:*

*- Deelnemers moeten een container kunnen maken*

*- Het concept van Artifactory moet duidelijk zijn*

*- Deployment laten doen*

*- Docker images*

*Op een gegeven moment: ML model in container stoppen, "hoe gaat de data nu stromen?"*

Uit de omschrijving van de cursusgids:

*In twee dagdelen (= één dag) ervaren deelnemers hoe Machine Learning geoperationaliseerd kan worden. We kijken naar concepten (immutable containers, scheiding van data en logica, aanwezigheid van alle dependencies) en maken zelf eenvoudig een container aan. Deelnemers ervaren hiermee hoe eenvoudig het is om alle dependencies “in te pakken” en vervolgens op elk platform te kunnen uitvoeren. Ook bespreken we de implicaties voor bewaarplicht van algoritmen en afhankelijkheden van software die ontwikkeld is door collega’s. De toepassing van containers en hoe dit binnen Achmea werkt wordt hiermee helder. Dit geeft Data Engineers krachtige  
gereedschappen in handen om Machine Learning-modellen te kunnen operationaliseren.*

MLOps in Azure ML is dus niet de gewenste richting. Moeten we dan ook niet gaan doen. In plaats daarvan container maken + artifact repository + Kubernetes (AKS).

## Oorspronkelijke planning

Oorspronkelijk hadden we dit als volgt gepland:

|  |  |
| --- | --- |
| Tijdstip | Activiteit |
| 9:00 | Start |
| 9:30 | uitleg *containerization* + demo |
| 10:00 | uitleg Docker + demo |
| 10:15 | Pauze |
| 10:30 | zelf doen: Docker bouwen |
| 11:00 | Packages round-up + demo |
| 11:30 | Gesprek over waarom je dit zou willen in je organisatie |
| 12:00/12:15 | Lunch |
| Middag | MLOps workshop vanuit Microsoft |

Bij de nadere invulling bleek dat dit te ambitieus was: de MLOps workshop is behoorlijk geavanceerd en technisch.

# Ideeën

Het idee is nu om een standaard model te pakken, aangeleverd door een Data Scientist. Door dat netjes in te pakken in een container kunnen we het eenvoudig hosten op een Container Instance, en wellicht zelfs op Kubernetes.

## Uitdagingen

De uitdaging hier is dat deelnemers niet op hun eigen laptop nu al kunnen werken met Docker, Kubernetes en/of OpenShift. Daarom moeten we een manier hebben dit zoveel mogelijk op Azure en/of Azure DevOps te doen. Vanuit de eerdere ervaring met Azure DevOps ben ik er echter niet gerust op dat mensen dit goed (werkend) hebben staan. Daarom liever wat minder DevOps.

Een optie hier is om KubeFlow te gebruiken. Hier zijn duidelijke instructies voor hoe dit aan te pakken op Azure.

Een andere optie is:

* Twee Azure DevOps-organizations aanmaken
* In elke DevOps-organization 5 deelnemers stoppen
* In de DevOps-organisaties een pipeline maken die een bestaand ML model omzet naar een container
  + Bestaand script pakken
  + Opsplitsen in training en inference
  + Container maken m.b.v. bestaande pipeline (invullen van de juiste script-namen)
  + Deployment naar ACR, ACI

Deze sluit meer aan bij wat cursisten nu al kunnen, namelijk:

* Beperkt python (invullen bij de puntjes)
* Beperkt DevOps (idem)
* (Zeer) beperkt containers

# Programma

Het programma ziet er als volgt uit:

|  |  |  |
| --- | --- | --- |
| Tijd | Onderwerp | Todo |
| 9:00 | Welkom  Rondje vragen  Recap containers | Plan: hoe recap containers? |
| 9:30 | Checkpoint: is het idee van de container duidelijk?  Werkvorm: leg elkaar uit in een groepje | Groepjes maken  Aanwijzen wie er wat uitlegt |
| 10:00 | Koffie |  |
| 10:15 | Lab: docker maken  Advanced: Python-scripts kopiëren + uitvoeren | VM maken met Docker aan boord  **Stappenplan opstellen**  Advanced bron: <https://towardsdatascience.com/deploy-machine-learning-pipeline-on-cloud-using-docker-container-bec64458dc01> |
| 11:00 | Recap Dockerfile  Demo in DevOps  Registratie ACR + uitleg | DevOps Demo klaarzetten Registratie ACR uitleggen |
| 12:00 | Lunch |  |
| 13:00 | DevOps + ACR + ACI lab | **Stappenplan opstellen**  **DevOps blueprint opzetten** |
| 14:00 | Koffie |  |
| 14:15 | Uitleg Kubernetes + demo |  |
| 15:00 | Lab: deploy naar K8s |  |
| 15:30 | MLOps: het grote plaatje | Slides / plaatje selecteren  **Demo in Azure ML klaarzetten** |
| 16:00 | Afsluiting / uitloop |  |

Voorbereiding voor aanvang:

* ~~Inrichten omgevingen (zie bijgevoegde Excel-file)~~
* Groepjes indelen (zie bijgevoegde Excel-file voor deelnemers)
* Bronnen voor “het grote plaatje” waaraan het verhaal opgehangen gaat worden kiezen
* Demo van Kubernetes voorbereiden
* Voorkennis check inplannen

Groepjes uitleg:

1: Ari + Mike + Ronald

2: Hans + Adrie + Peter + Jette

3: Dirk + Frank + Eelco + Ron

Als Mike er niet is: Ron -> 1