Machine Learning Pipeline Automation

Amar Kisoensingh

5 april 2021

Voorwoord

Samenvatting

Summary

Afkortingen

PaaS Platform as a Service 15, 16

Begrippenlijst

cloud computing platform first 16

datamining 15

machine learning is het ontwikkelen van algoritmes en technieken waarmee computers kunnen leren. 13, 15

machine learning pipeline is een pijplijn waarin gedefiniëerd staat hoe data stroomt. first 15, 16

vendor lock-in is niet de mogelijkheid hebben om naar een andere dienst/provider te gaan die dezelfde service biedt. 15, 16

Lijst van figuren

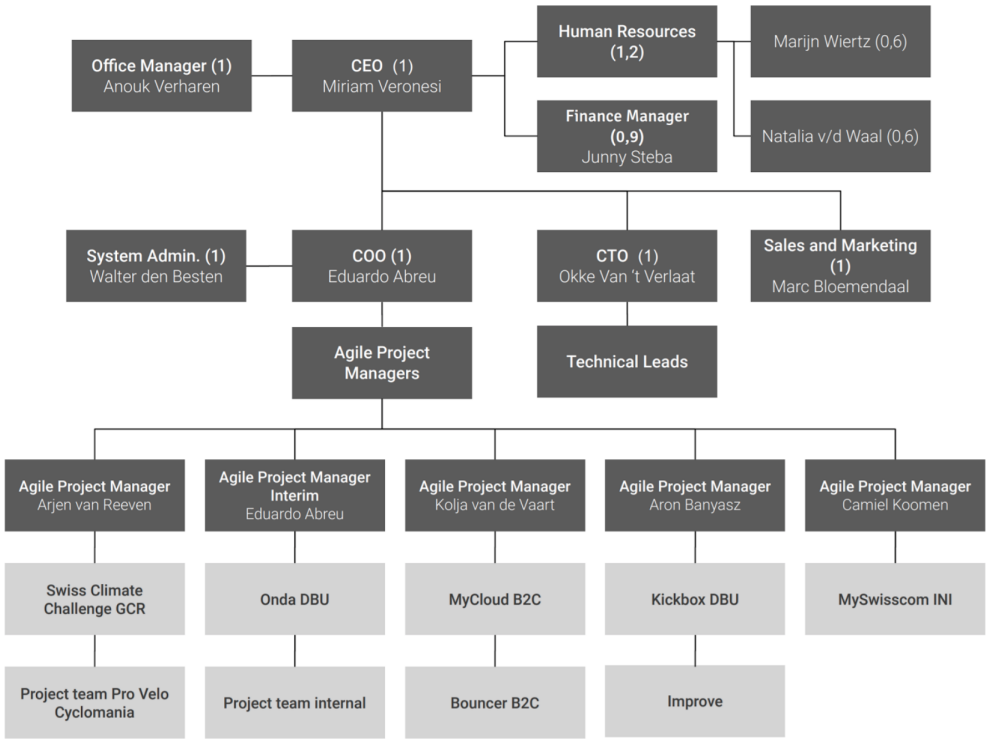
Lijst van tabellen

Inhoudsopgave

1 Inleiding

NGTI is een software ontwikkel bedrijf dat gevestigd is in Rotterdam. Met het starten van nieuwe projecten begint NGTI met de probleem stelling, mockups en wireframes en prototyping. Vervolgens wordt een applicatie voor mobiel en/of webgebruik ontwikkelt en wordt support geleverd voor bijvoorbeeld updates of het oplossen van bugs [1]. Naast het maken van een applicatie op maat biedt NGTI ook andere diensten, zoals een app framework of white label apps [2].

Kijkend naar hoe het bedrijf opereert is het een vrij 'platte' structuur. Formeel bestaat er wel een hiërarchie (Figuur 1.1), maar deze is in de praktijk niet zo gauw terug te vinden.



Figuur 1.1: Organogram van NGTI op 29-03-2021 [3].

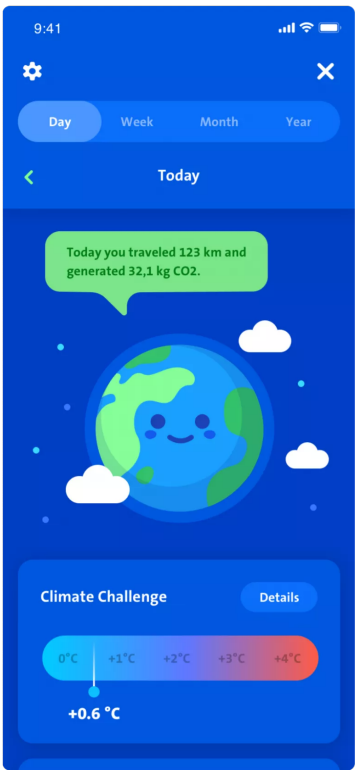
Swisscom heeft verschillende dochterbedrijven [4] waarvan NGTI er een van is. Sinds maart 2021 is het bekend gemaakt dat Swisscom van plan is om een afdeling, Swisscom DevOps Center, te fuseren met NGTI. Omdat de fusie onzekerheid met zich meebrengt voor de structuur en manier hoe NGTI werkt, zal de situatie vóór de fusie aangehouden worden gedurende het afstuderen.

1.1 Projecten

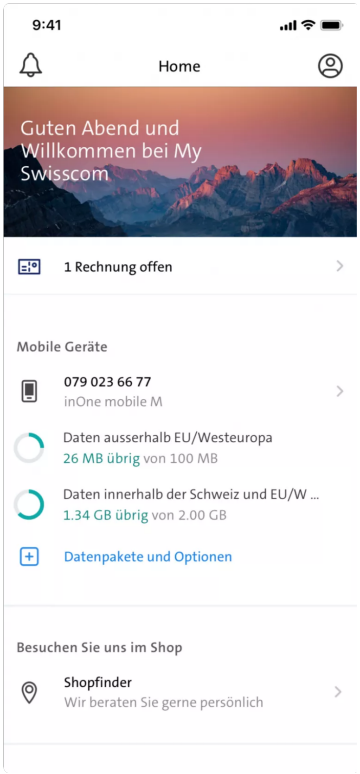
NGTI heeft een vrij breed portfolio met apps voor verschillende doeleinden. Een van deze apps is de Climate Challenge App [5]. Me deze app kunnen gebruikers hun C02-voetafdruk en impact in kaart brengen. Er wordt bijgehouden hoeveel kilometer de gebruiker reist en

met welke vervoersmiddel. De app is onderdeel van twee bestaande nieuws apps, Blick en Bluewin [6]. Het doel is om de gebruiker aan te sporen om groener te reizen. Een screenshot van de app is te zien in Figuur 1.2

Een andere oplossing is de My Swisscom App [7]. Dit is een native app voor Android en iOS waarbij Swisscom-klanten hun contract kunnen bestellen, wijzigen of beëindigen. In de app kunnen klanten ook de dataverbruik zien en instellingen voor abonnementen wijzigen. Een screenshot van de app is te zien in Figuur 1.3.



Figuur 1.2: Screenshot van de Swiss Climate Challenge app [5].



Figuur 1.3: Screenshot van de My Swisscom App [7]

1.2 Tools die worden gebruikt

Om productief te zijn gebruikt NGTI een aantal tools en programma's om producten te maken en te communiceren met zowel collega's als klanten. De meest gebruikte en belangrijkste zijn Slack, Google Workspace en Zoom.

1.2.1 Slack

Interne communicatie gaat via Slack. Het programma faciliteren collega's om elkaar met een lage instap te benaderen en berichten die voor het hele bedrijf relevant zijn te versturen. Ook zijn er 'channels' beschikbaar over specifieke onderwerpen, zoals: *#dev*, *#ios* en *#test-automation*.

1.2.2 Google Workspace

Met Google Workspace kunnen bestanden en documenten gemaakt, opgeslagen en gedeeld worden. Omdat dit via een browser kan, hoeven werknemers geen software te installeren. NGTI gebruikt het ook om collaboratief en parallel te werken aan hetzelfde document.

1.2.3 Zoom

Voorheen werd Zoom alleen gebruikt om te videobellen met collega's en geïnterviewden. In de tijd van het pandemie is Zoom echter een belangrijke speler geworden om effectief samen te werken. Meetings zoals introducties van nieuwe collega's of demo's van producten worden online gehouden.

1.3 Aanleiding opdracht

NGTI heeft voorzien dat ze haar applicaties 'slimmer' moet maken door machine learning (ML) in te zetten. Niet alleen zorgt dit voor een betere gebruikerservaring, maar geeft NGTI ook een voorsprong op haar concurrenten.

De opdracht bestaat uit drie onderdelen:

1. Onderzoek naar hoe het opzetten van een pipeline en het maken van de acties in elke stap geautomatiseerd kunnen worden
2. Onderzoek naar hoe het trainen van een model versimpeld kan worden voor de developer
3. Onderzoek naar het maken van een platform-agnostische oplossing

Hierbij zal een proof-of-concept (PoC) gemaakt worden om aan te tonen of het haalbaar is. Een diepere duik in het probleem en het definiëren van de onderdelen is te vinden in hoofdstuk 2.

1.4 Leeswijzer

TODO: Schrijf leeswijzer voor elk hoofdstuk

2 Probleemanalyse

Zoals beschreven in paragraaf 1.3 is NGTI genoodzaakt om machine learning in te zetten om haar applicaties 'slimmer' te maken. Hier zijn een aantal redenen voor, onder andere:

1. gebruikerservaring verbeteren
2. voorsprong hebben op concurrenten

Het 'slimmer' maken van applicaties kan op verschillende manieren, maar met machine learning kan een platform gebouwd worden waarmee elke richting op gegaan kan worden. Om machine learning te implementeren in haar applicaties loopt NGTI tegen een aantal obstakels op, namelijk: expertise vereist in het machine learning pipeline domein, tijd om een pipeline op te zetten en vendor lock-in.

2.1 Expertise Machine Learning

Machine learning is geen triviaal onderwerp. Om een model te trainen is kennis nodig van verschillende domeinen: datamining, software engineering en statistieken. In een multidisciplinair team is het voor één teamlid niet nodig om alle domeinen te beheersen.

Doordat er voorkennis nodig is om een model te trainen en een pipeline goed op te zetten, is het vaak te hoogdrempelig voor developers. De expertise is daar bovenop niet in een korte tijd te vergaren.

2.2 Opzetten pipeline

Bovenop de complexiteit van machine learning zelf bestaan er verschillende manieren om een model te trainen. Een machine learning pipeline opzetten is daar een van. Een pipeline is een workflow dat bestaat uit een aantal stappen die doorgelopen kan worden om een model te trainen. In elke stap worden acties uitgevoerd, zoals het verwijderen van onbruikbare data of de prestatie van modellen vergelijken en een rapport met uitslagen genereren. Het opzetten van zo een pipeline én de actie(s) in de stappen definiëren kost tijd en kennis. Daarnaast zijn de stappen en acties vaak hetzelfde voor verschillende pipelines. Het automatiseren en hergebruiken van stappen en acties tussen pipelines zou tot tijdwinst leiden.

2.3 Vendor lock-in

Er bestaan een aantal diensten, zogenoemde Platform as a Service (PaaS), waarbij je een pipeline kan opzetten en acties kan definiëren. Een van de problemen met een PaaS is vendor lock-in. Dit betekent dat, als er eenmaal een pipeline is opgezet, de overdraagbaarheid van de pipeline naar een andere PaaS vrijwel onmogelijk is. Ook zijn de opties en mogelijkheden om uit te breiden in de toekomst gelimiteerd.

2.4 Doelstelling

Om het probleem op te lossen is onderzoek en experimentatie nodig op verschillende vlakken. De gewenste oplossing is het ontwikkelen van een systeem waarbij developers met weinig tot geen kennis een model kunnen trainen. Het systeem moet de infrastructurele taken voor zich nemen, zoals het opzetten van een pipeline, de stappen en acties automatiseren. Daarnaast moet het systeem ook platform agnostisch zijn.

2.5 Bestaande oplossingen om pipelines op te zetten

Er bestaan een aantal oplossingen voor vrijwel alle problemen die NGTI ondervindt. Elke oplossing is een PaaS van een derde partij waarbij vendor lock-in inherent is. Dit maakt ze ongeschikt maar betekent echter niet dat ze nutteloos zijn. Er kan namelijk gekeken worden hoe een pipeline wordt opgezet, welke acties de stappen verricht en daar vervolgens (gedeeltelijk) het systeem op baseren. Daarnaast wordt bij alle oplossingen een expertise van ML op een bepaalde niveau verwacht.

De oplossingen kunnen gecategoriseerd worden in twee groepen:

1. Machine learning pipeline specifieke services
2. Cloud computing platformen

2.5.1 Machine learning pipeline specifieke service

Bedrijven zoals Algorithmia [8] en Valohai [9] bieden alleen diensten om pipelines op te zetten. Ze zorgen voor het databehoud dat door de gebruiker wordt geüpload en het trainen van het model. Verder kan er toezicht gehouden worden op de kosten, beschikbaarheid en prestatie van het model.

Valohai heeft documentatie een aantal blog posts die bij het ontwerpen van het systeem relevant zouden kunnen zijn.

2.5.2 Cloud computing platformen

De drie grote cloud computing platformen Amazon, Azure en Google hebben meer te bieden dan alleen een pipeline opzetten, zoals het hosten van een website, database of virtuele server. De cloud computing platformen hebben hetzelfde probleem als de machine learning pipeline specifieke services; vendor lock-in is onvermijdelijk. Wat wel een mogelijkheid zou kunnen zijn is dat de andere services van de cloud computing platformen gebruikt kunnen worden als onderdeel van het systeem.

Het systeem zou bijvoorbeeld een server kunnen aanmaken, een model trainen, het resultaat downloaden en vervolgens de server verwijderen.

2.6 Frameworks om cloud computing platformen te beheren

Gedurende de vooronderzoek zijn frameworks dat cloud computing platformen beheert en frameworks waarmee een pipeline uitgerold kan worden naar boven gekomen. Deze

frameworks kunnen deel uitmaken van de oplossing. Een framework dat cloud computing platformen kan beheren is Terraform [10]. Met Terraform is het mogelijk om een plan te schrijven waarin bijvoorbeeld staat welke type server nodig is. Bij het uitvoeren van het plan spreekt Terraform een cloud computing platform naar keuze aan om de server op te starten. Terraform kan vervolgens controleren of de server draait en de server afsluiten wanneer het niet meer nodig is.

Kubeflow [11] is een ander framework waarmee een pipeline kan worden opgezet. Het verschil met Terraform is dat Terraform flexibeler is met wat er aangemaakt kan worden op een cloud computing platform. Kubeflow kan alleen een pipeline uitrollen.

2.7 Hoofd- en deelvragen

Uitgaand van de drie obstakels kan de hoofdvraag als volgt worden geformuleerd:

In welke mate kan een machine learning pipeline worden geautomatiseerd onafhankelijk van de onderliggende cloud computing platform?

De hoofdvraag kan worden onderbouwd met vier deelvragen. Om te beginnen is het verstandig om te weten welke stappen er in een machine learning pipeline zit:

Uit welke stappen bestaat een machine learning pipeline?

Vervolgens kunnen de verschillende cloud computing platformen in kaart worden gebracht:

Wat zijn de verschillen en overeenkomsten tussen cloud computing platformen waarmee een machine learning pipeline kan worden opgezet?

Verder kan het handig zijn om meerdere platformen aan te spreken. Dit zou kunnen met een bestaand framework. Hierbij is het, net als de vorige deelvraag, belangrijk om te weten welke frameworks er zijn:

Wat zijn de verschillen en overeenkomsten tussen frameworks waarmee cloud computing platformen beheerd kunnen worden?

Ten slotte wordt een PoC gemaakt om te laten zien of het probleem oplosbaar is. Hiervoor is een doordachte voorbereiden onmisbaar:

Hoe ziet de architecturale blauwdruk van een applicatie, waarmee een machine learning pipeline kan worden opgezet, die modulaire acties geautomatiseerd in stappen samenstelt, en die platform-onafhankelijk is, eruit?

3 Onderzoeksmethoden en scope

Om elke hoofd- en deelvraag te beantwoorden, maak ik bij elk gebruik van een onderzoeksmethode. Volgens Scribbr [12] zijn er twee onderzoeksmethoden: kwantitatief en kwalitatief. Bij een kwantitatief onderzoeksmethode wordt data verzameld waarmee grafieken of tabellen gemaakt kunnen worden. De focus bij een kwalitatief onderzoeksmethode ligt bij het verzamelen van verschillende interpertaties en opvattingen. Hierop kan optioneel een eigen interpertatie op gemaakt worden. [13].

Onder kwantitatief en kwalitatief vallen verschillende dataverzamelingsmethoden. Deze beschrijft simpelweg de manier hoe data wordt verzameld. Dit kan bijvoorbeeld met een enquête, literatuuronderzoek op websites en in boeken of een onderzoek over een lange periode [13].

Elke hoofd- en deelvraag is gekoppeld aan een onderzoeksmethoden. Vervolgens is beschreven welk(e) dataverzamelingsmethode(n) wordt gebruikt met een korte toelichting. Daarnaast wordt op een hoog niveau de scope bepaald en de requirements vanuit NGTI mocht die er zijn.

D1: Uit welke stappen bestaat een machine learning pipeline?	
Methode(s)	Kwalitatief
Dataverzamelings methode(n)	Literatuuronderzoek, fundamenteel onderzoek, toegepast onderzoek
Scope	Binnen de scope: • In kaart brengen uit welke stappen een pipeline bestaat • Acties die in een stap worden uitgevoerd • Of het mogelijk is om stappen te versimpelen / abstraheren voor developers • Of het mogelijk is om stappen en acties te automatiseren Buiten de scope: • Automatisering van stappen en acties • Een versimpeling van machine learning
Toelichting	Er wordt gekeken naar welke stappen er in een pipeline zitten. De theorie wordt vervolgens toegepast in een experiment. De nadruk ligt vooral of de mogelijkheid er is om stappen en acties te automatiseren en of machine learning versimpeld kan worden, niet dat er een uitwerking is.

Tabel 3.1: Deelvraag 1

D2: Wat zijn de verschillen en overeenkomsten tussen cloud computing platformen waarmee een machine learning pipeline kan worden opgezet?	
Methode	Kwalitatief
Dataverzamelings methode(n)	Literatuuronderzoek, vergelijkend onderzoek
Scope	Binnen de scope: • Inventarisatie met de ”long list short list” methode • Welke functionaliteit bieden de platformen op een machine learning pipeline op te zetten • Ervaring opdoen doormiddel van een pipeline te maken binnen twee cloud computing platformen • Basale inventarisatie en vergelijking van alternatieve manieren om met een cloud computing platform te communiceren Buiten de scope: • Prijs, performance en snelheid
Toelichting	Cloud computing platformen worden in kaart gebracht. Vervolgens wordt met criteria bepaald in een later stadium de lijst verkort tot twee kandidaten. Alternatieve manieren om met een cloud computing platform te communiceren is afgebakend tot first-party tools en frameworks dat een of meerdere platformen tegelijk kan aanspreken.

Tabel 3.2: Deelvraag 2

D3: Wat zijn de verschillen en overeenkomsten tussen frameworks waarmee cloud computing platformen beheerd kunnen worden?	
Methode	Kwalitatief
Dataverzamelings methode(n)	Literatuuronderzoek, vergelijkend onderzoek
Scope	Binnen de scope: • Inventarisatie naar wat er aangemaakt, gewijzigd en verwijderd kan worden binnen een cloud computing platform • Hoe een machine learning pipeline op papier gemaakt zou worden met een framework • Experiment met het opzetten van een pipeline via het framework op een cloud computing platform
Toelichting	Er wordt gekeken naar welke frameworks er beschikbaar zijn en wat de verschillen/overeenkomsten zijn. Om de applicatie future proof te maken is het belangrijk om een framework te kiezen wat in de praktijk beproefd is en ondersteuning van een community heeft.

Tabel 3.3: Deelvraag 3

D4: Hoe ziet de architecturale blauwdruk van een applicatie, waarin een machine learning pipeline kan worden opgezet, die acties voorgeprogrammeerd zijn, en die platform-onafhankelijk is, eruit?	
Methode	Kwalitatief
Dataverzamelings methode(n)	Literatuuronderzoek
Scope	Binnen de scope: • Technische tekeningen Buiten de scope: -
Toelichting	De literatuuronderzoek slaat op of de technische tekeningen gemaakt zijn volgens een standaard zoals UML. Dit komt niet terug als theorie maar de bronnen worden wel vermeld.

Tabel 3.4: Deelvraag 4

H: In welke mate kan een machine learning pipeline worden geautomatiseerd onafhankelijk van de onderliggende cloud computing platform?	
Methode	Kwalitatief
Dataverzamelings methode(n)	Literatuuronderzoek
Scope	De scope wordt bepaald na de requirement analyse.
Toelichting	Onderzoek naar documentatie van gebruikte framework(s).

Tabel 3.5: Hoofdvraag

4 Stappen in een Machine Learning pipeline

5 Cloud computing platformen

6 Frameworks om platformen te beheren

7 Architecturale ontwerp van de oplossing

7.1 Literatuur

7.2 Design

7.3 Conclusie

8 Oplossing

8.1 Scope definiëren

8.2 Research techstack

8.3 Wireframe

8.4 Mockup

8.5 POC

8.6 Conclusie

9 Conclusie

10 Aanbeveling

11 Discussie

12 Reflectie

12.1 Deelvraag 1: Hoe werkt Machine Learning?

Bibliografie

[1] 22 mrt 2021. URL: https://www.ngti.nl/diensten/.

[2] 29 mrt 2021. URL: https://www.ngti.nl/oplossingen/.

[3] NGTI B.V. „Organogram".

[4] 29 mrt 2021. URL: https://www.swisscom.ch/en/about/beteiligungen-swisscom-uebersicht.html.

[5] 29 mrt 2021. URL: https://www.ngti.nl/cases/swiss-climate-challenge/.

[6] 29 mrt 2021. URL: https://www.swissclimatechallenge.ch.

[7] 29 mrt 2021. URL: https://www.ngti.nl/cases/my-swisscom-app/.

[8] Algorithmia Inc. 18 feb 2021. URL: https://algorithmia.com.

[9] Valohai. 22 feb 2021. URL: https://valohai.com.

[10] 5 apr 2021. URL: https://www.terraform.io.

[11] 5 apr 2021. URL: https://www.kubeflow.org.

[12] 25 mrt 2021. URL: https://www.scribbr.nl/scriptie-structuur/methodologie-in-je-scriptie/.

[13] 25 mrt 2021. URL: https://www.scribbr.nl/onderzoeksmethoden/kwalitatief-vs-kwantitatief-onderzoek/.

Bijlagen