

DATASCIENCE VERGT HARD EN NOEST WERKEN

De CIO als datafabrieksmanager

Datascience is hot. Lees de vacatures en blogs er maar op na. Ieder bedrijf 'doet' aan big data, is grote data-lakes aan het bouwen, en wil spannende dingen met machine-learning of deep learning doen. Elke universiteit levert groepen data-scientists af die vol verwachting de bedrijven instromen, bedenkend hoe zij met hun algoritmes de digital transformatie mogelijk gaan maken.

Maar helaas. Datascience bestaat voor meer dan 90 procent uit hard en noest werken. Niet aan algoritmes en mooie machine-learningmodellen, maar aan data. Aan methodisch werken. En aan automation. Aan het opzetten van een fabriek voor je dataverwerking.

Data-wrangling

De data die je binnenkrijgt is zelden zo mooi als de Kaggle-sets die schoon zijn en opgesteld vanuit een duidelijke probleemstelling. Dus de eerste stappen in datascience bestaan uit scripts schrijven om de data binnen te krijgen, te bedenken hoe deze op te schonen, wat te doen met outliers en anomalies, hoe om te gaan met missende waarden, en of je in het kader van bijvoorbeeld de GPDR al vanaf in-

gestion je data moet anonimiseren. JADS (Jheronimus Academy of Data Science in Den Bosch) noemt dit met een herkenbare term 'data-wrangling'. Belangrijk is om deze stap direct goed neer te zetten: denk na over je data-pipeline.

Datastorage

Parallel moet je al nadenken over het vak dat vroeger de database-engineer of de warehouse-expert uitvoerde. Kies je een relationele database, NoSQL of unstructured database om je *wrangled* en *analyzed* data op te slaan? Wat gebruik je om terabytes aan images op te slaan en snel te kunnen processen? Moet je direct schaalbaar opzetten met meerdere nodes, of denk je daar pas later aan? Hoe bescherm je de data tegen verlies of ongewenst gebruik? Moet je alleen de resultaten van je machine-learningmodel opslaan of ook de tussenstappen en inputs? Nadenken over datastorage is ook een onderdeel van je data-pipeline, en iets om vanaf de start mee te nemen in je besluiten.

Analytics

Met de opgeschoonde dataset die slim opgeslagen staat kun je vervolgens aan de slag met analytics voor je eerste inzichten. Wil je slim zijn, dan moet je weten wat er precies gebeurt, welke data echt zinvol is en welke patronen er te onderkennen

zijn. Data is er genoeg, maar tot nog toe kunnen alleen mensen het echt interpreteren. Bijvoorbeeld correlaties, hiervoor heb je domeinkennis nodig. Want wat voor jou als statisticus een aha-moment is omdat je een sterke correlatie vindt vanuit de data, is voor de domeinexpert een 'duh'-opendeurmoment: logisch dat je dat wordt als het regent.

Hetzelfde geldt voor de algoritmes en modellen die een datascientist toepast. Er bestaan genoeg libraries en toepasbare modellen, maar als je niet begrijpt wat deze doen met je data, welke variabelen wel en niet van belang zijn, wanneer er sprake is van overfitting, dan kom je niet ver. Dan heb je een cool model met een lousy uitkomst.

CIO als datafabrieksmanager

Aan de CIO of CDO de taak om dit te sturen. Sturen op een goede data-pipeline. Sturen op begrip van de data alsook op domeinkennis. Sturen op doorvragen en nadenken. Datascience is lastig en is zeker geen wondermiddel. Het systeem doorgronden vergt manjaren verdieping, en de businesscase moet je altijd blijven zoeken vanuit boerenverstand en domeinkennis.

Marianne Faro is director van Itility.

