Databricks: van Spark- naar ai-deskundige

Onderzoeksafdeling zit in Amsterdam vanwege CWI

21 augustus 2023 14:00 | Teus Molenaar

Topic Datamanagement

In tien jaar tijd van een bedrijf dat met Apache Spark een platform bood voor verwerking en analyse van grote hoeveelheden data naar een onderneming die de spil is in toepassing van artificiële intelligentie (ai). Onder meer dankzij de onderzoeksafdeling van Databricks in Amsterdam. Dat stelt Matei Zaharia, chief technology officer (cto) en medeoprichter van het Amerikaanse bedrijf.

Tijdens de Data + AI Summit van Databricks eind juni in San Francisco spreekt Computable met Matei Zaharia. Naast cto is hij universitair hoofddocent computerwetenschappen aan de Stanford Universiteit.

We gaan terug naar 2015, het jaar dat de eerste Spark Summit plaats vond in Amsterdam. Hoezo Amsterdam? Blijkt dat een onderzoeksafdeling van het bedrijf, dat zijn hoofdkantoor heeft in San Francisco, in Amsterdam is gevestigd. 'Dat heeft te maken met de aanwezigheid van het Centrum voor Wiskunde en Informatica – CWI, legt Zaharia uit. 'Er was een sterke database-onderzoeksgroep. Die zat achter de oprichting van Vectorwise in 2004, een krachtige, analytische database. In 2010 is het overgenomen door het Amerikaanse softwarehuis Action Corporation. Maar die data-kennis is nog steeds aanwezig in Amsterdam.'

Destijds begon de onderzoeksafdeling met tien medewerkers; inmiddels zijn er meer dan honderdvijftig mensen die louter onderzoek doen en productbeheer. 'We hebben wel meer collega's op ons kantoor in Amsterdam, maar deze medewerkers doen alleen research.'

Apache Spark

In 2009 was Zaharia student aan de Universiteit van Californië, Berkeley. Hij stond aan de wieg van Apache Spark, een opensource-framework voor de verwerking en analyse van grote hoeveelheden data. Daarbij gaat het om zulke grote hoeveelheden dat gangbare databases er niet goed mee overweg kunnen.

Matei Zaharia, cto Databricks.

In 2013 richt hij Databricks op. Samen met medestudenten Ali Ghodsi (nu ceo van Databricks), Reynold Xin (manager software engineering), Patrick Wendell (hoofd engineering), Andy Konwinski (adviseur) en Berkeley-professor Ion Stoica (executive chairman). 'Veel bedrijven deinsden ervoor terug om een open-source-toepassing te gebruiken. Ze wilden niet iets van een universiteit. Daarom hebben we een gehamaste versie ontwikkeld met dienstverlening erom heen. Dat gaf meer vertrouwen.'

Van meet af aan biedt Databricks een platform via cloud computing. 'Wij werkten tijdens onze studie met grote hoeveelheden data, maar we beschikten op de universiteit niet over computers die daarmee overweg konden. Daarvoor gebruikten we de cloud. Daar hadden we dus al veel ervaring mee opgedaan', verklaart Zaharia.

Natuurlijke taal

Wat waren de onderwerpen destijds in 2013 tijdens het evenement in Amsterdam? 'Het werken met grote datasets – eventueel in de cloud – via Spark was nieuw in die tijd. Mensen wilden vooral weten hoe je dat doet. Maar vooral welke andere bedrijven er al mee aan de slag waren gegaan. Ze wilden ervaringen met anderen delen. Omdat wij een commerciële versie van Spark hadden ontwikkeld, genaamd Databricks Runtime, speelden we een soort bemiddelaarsrol. Wij konden gebruikers aan elkaar voorstellen.'

Nu tien jaar later, zo vertelt Zaharia, is het aanbod veel completer. Waarbij Databricks heel veel aandacht heeft besteed – en dat nog steeds doet – om de technologie voor iedereen toegankelijk te maken. 'Machines zijn in het geheel niet toegerust om de inhoud van data te begrijpen en weten dus ook niet wat je ermee kunt doen. LLM's, large language models (grote taalmodellen, TM), veranderen dat. De meest toegepaste programmeertaal tegenwoordig is Engels; of elke andere natuurlijke taal. Je kunt tegen een computer praten. Sommigen krijgen daar sciencefictionachtige beelden bij en zien Star Trek voor zich. Ik weet niet of dat gaat gebeuren, maar wel dat data voor iedereen binnen een organisatie beschikbaar komen zonder dat daar een legertje data scientists aan te pas moet komen.'

Meer ruimte

Lakehouse IQ bijvoorbeeld traint datamodellen op specifieke processen en taalbegrippen van een onderneming, zodat je een ai-toepassing krijgt die naadloos past bij een bepaalde organisatie. 'Er zijn worden die hetzelfde zijn, maar in verschillende branches een andere inhoud hebben. Ons product leert dat, zodat er geen misverstanden ontstaan. In de kunstwereld is een curator iemand die een tentoonstelling inricht. In het bedrijfsleven is een curator iemand die een faillissement afhandelt, om een voorbeeld te noemen. Wij zorgen ervoor dat er modellen worden gemaakt die de context begrijpen en door iedereen benaderbaar zijn.'

Alhans, daar werkt Databricks aan. De eerste publieke testfase is aanstaande. Tijdens het data+ai evenement in San Francisco heeft de onderneming tal van aankondigingen gedaan. Zoals de Marketplace, de overname van MosaicML, Lakehouse Apps, Clean Room, Deltalake 3.0. Er leek geen einde aan te komen.

'Eind 2013 hadden we zo'n twintig medewerkers, nu zitten we wereldwijd op meer dan vijftduizend. Onze omzet is navenant gestegen. Dat betekent eenvoudigweg dat we meer ruimte hebben om nieuwe diensten en producten te ontwikkelen. Dat doen we trouwens niet alleen, want dat lukt geen enkel bedrijf tegenwoordig. Je moet samenwerken met andere. Zo werken we nauw samen met Microsoft dat Delta Lake, ons open-source-product, gebruikt als basis voor zijn Azure Delta Lake Storage.'

Data democratization

Databricks is er alles aan gelegen om het werken met (grote hoeveelheden) data voor iedereen toegankelijk te maken. Met een mooi woord noemen ze dat data democratization. Waarbij Databricks verantwoordelijkheid draagt voor de juiste gegevensbeveiliging, privacybescherming en governancemaatregelen met als doel dat gegevens nauwkeurig, betrouwbaar en veilig worden gebruikt, zelfs wanneer de toegang wordt gedemocratiseerd. 'Toen ik studeerde en een website wilde bouwen, had ik vier boeken nodig. Dat is tegenwoordig niet meer nodig. Het is gemakkelijk gemaakt. Dat gaan we doen voor ai', besloot Zaharia.

Lijf

In San Francisco loopt Computable Ivo Everts tegen het lijf. Hij is tegenwoordig strategic solutions architect in het Amsterdamse kantoor. Ooit begonnen als de eerste solutions architect voor Databricks in Nederland. Hij begrijpt wel hoe Databricks aan de ontwikkelafdeling in Amsterdam is gekomen. 'Nog steeds is Peter Boncz adviseur, zeg hij. Boncz is professor aan het CWI, gespecialiseerd in grote datasystemen.

"Shell gebruikt Datalake om te voorspellen hoeveel en welke snacks er nodig zijn op benzinestations"

Everts is nauw betrokken bij hoe ABN Amro en Shell werken met Lakehouse. 'Het gaat altijd om een partnerschap. We werken samen met de klanten aan de bouw van de modellen. Bij ABN Amro gebruiken ze ons platform om razendsnel transacties door te lichten op eventueel frauduleus handelen. En er zijn ook andere use-cases.' Shell is evenwel de grootste klant. 'Daar wordt ons platform gebruikt om processen minder vervuilende stoffen te laten uitstoten. Wij zitten met Lakehouse in het hart van hun data-architectuur. Zo hebben we digitale twins gemaakt van de raffinaderijen. Om preventief onderhoud te kunnen plegen, onder andere.'

Als een bedrijf eenmaal met Lakehouse gaat werken, zo is Everts' ervaring, dan zijn er altijd wel meer toepassingen te verzinnen. 'Zo gebruikt Shell Datalake om te voorspellen hoeveel en welke snacks er nodig zijn op benzinestations.'

Betere kwaliteit

Er zijn meer klanten in Nederland en België. 'Zoals Ahold Delhaize, maar ook een verdeelsingsbedrijf in Nederland. Dit gebruikt ons platform om goede van slechte zaden te onderscheiden', zegt Everts. Het delen van data, op een veilige, gereguleerde manier, wordt steeds belangrijker, vindt hij. 'Om smart cities te bereiken, zal je met veel partijen data moeten delen. Wij faciliteren dat.'

Waar Databricks over vijf jaar staat, is volgens Everts moeilijk te voorspellen. De ontwikkelingen gaan razendsnel. Duidelijk is wel dat llm's de zaken gaan opschudden. 'Ze gaan SQL-queries schrijven, code genereren. Er komen miljoenen modellen. Menselijke interventie blijft nodig, maar over de hele linie krijg je een betere kwaliteit.'

Wat vond u van dit artikel?

Dit artikel delen:

Lees verder

Datamanagement | Carrière | Business Analytics | SAN | Databases | Apache | Opensource | Diensten | Kunstmatige intelligentie | CWI | Databricks

Databricks pompt 1,6 miljard in 'datahuis aan meer'

31 AUGUSTUS 2021 22:41

Uw reactie

LET OP: U bent niet ingelogd. U kunt als gast reageren maar dan wordt uw reactie pas zichtbaar na goedkeuring door de redactie. Om uw reactie direct gepubliceerd te krijgen moet u eerst rechtsboven inloggen of u registreren

Naam

E-mailadres

Reactie

Ik ga akkoord met de voorwaarden.

Jaarbeurs b.v. gaat zorgvuldig en veilig om met uw persoonsgegevens. Meer informatie over hoe we omgaan met je data lees je in het privacybeleid

Preview