Van lokale opslag naar clouddiensten

In de derde aflevering van zijn big data-serie gaat Klaas Jan Mollema in op verschillende vormen van fysieke dataopslag. Niet alleen technische aspecten bepalen de keuze voor een opslagvorm, maar ook bedrijfskundige en ethische aspecten.

Klaas Jan Mollema MSc.

In de eerdere afleveringen verkenden we de terminologie rond big data en gingen we in op datakwaliteit. Elk bedrijf wenst immers veilige, toegankelijke en betrouwbare data om zijn beslissingen op te nemen. In deze aflevering benoemen we de voor- en nadelen van databeheer in de cloud en bekijken we enkele clouddiensten.

De cloud

Na de introductie van internet in de jaren negentig volgde al snel cloudcomputing. Reeds in 1997 werd de term gebruikt in de literatuur en in 2000 werden de eerste clouddiensten aangeboden door grote partijen als Amazon en Microsoft. Cloudcomputing is het via een netwerk op aanvraag beschikbaar stellen van hardware, software en gegevens. De term 'cloud' komt voort uit de schematische weergave in de vorm van een 'wolk met computers' waarbij de eindgebruiker niet meer weet op welke computer de software draait die hij gebruikt of waar de gegevens zijn opgeslagen. Clouddiensten zijn abonnementsdiensten die we in drie verschillende vormen kennen:

> Software as a Service (SaaS) biedt de gebruiker eindapplicaties aan;
> Platform as a Service (PaaS) biedt een platform aan waar gebruikers toepassingen op een geïntegreerde manier kunnen gebruiken (onder andere toegangsbeheer en portaalfuncties);
> Infrastructure as a Service (IaaS) biedt infrastructuur aan, zoals opslagcapaciteit, servers en netwerken.

SaaS-diensten kent de gebruiker van Google Docs, Microsoft Office 365 en weblogomgevingen als WordPress. De applicatie draait volledig op de server en gratis of voor een klein bedrag per maand maak je gebruik van de dienst. PaaS-diensten worden vooral toegepast voor het verzorgen van 'single sign on' binnen organisaties, zodat met één inlogactie alle toepassingen toegankelijk zijn. IaaS-diensten bieden in eerste instantie vooral de hardware en de infrastructuur aan. Maar het is ook mogelijk om een compleet geïnstalleerde configuratie te huren, zoals een Cloudera-omgeving voor Hadoop-databeheer.

Dataopslag

Het opslaan en verwerken van data in de cloud is een Infrastructure as a Service-oplossing. Verschillende aanbieders bieden servercapaciteit aan in de cloud. Via een webomgeving neem je een abonnement op een server die dan geheel of gedeeltelijk tot je beschikking komt voor de opslag van gegevens. Door meerdere servers in te zetten vergroot je niet alleen de opslagcapaciteit, maar kun je ook voor de back-up van gegevens zorgen (zie aflevering 1 van deze big data-serie). Ook vergroot de inzet van meerdere servers de rekencapaciteit die je kunt inzetten voor de analyse en verwerking van gegevens.
Doordat je de capaciteit alleen inkoopt wanneer je die nodig hebt, is een abonnement op een clouddienst erg geschikt voor situaties waar schaalbaarheid van opslagruimte en rekencapaciteit per moment erg verschillen.
Andere voordelen van cloud-opslag zijn de goede fysieke beveiliging van datacentra, de degelijke koeling van de servers en het feit dat je het technisch beheer uitbesteedt aan de aanbieder van de dienst. Tevens sluit werken in de cloud naadloos aan op mobiele vormen van werken.

Lokale dataopslag

Maar dataopslag in de cloud brengt ook nadelen met zich mee. Doordat gegevens 'ergens' in de cloud worden opgeslagen is de fysieke locatie van de gegevens niet bekend. Juridisch, ethisch en privacy-technisch een lastig probleem wanneer het om gevoelige informatie gaat. Fysieke opslag in bijvoorbeeld Amerika zorgt ervoor dat de gegevens vallen onder de Patriot Act en dus inzichtelijk zijn voor de Amerikaanse veiligheidsdiensten. Soms bieden aanbie-

ders fysieke opslag aan op een specifiek werelddeel om dit probleem te voorkomen. Maar ondanks dat Google servercapaciteit in elk land opzet, kunnen ze niet te allen tijde garanderen dat de gegevens die in dat land worden geproduceerd ook echt binnen de landsgrenzen worden opgeslagen.*

Om er zeker van te zijn dat data in het land van je organisatie wordt opgeslagen, zal je een eigen serverruimte of datacentrum moeten inrichten. Dat vraagt om een goede infrastructuur, stroomvoorziening, koeling, back-up en beveiliging plus de kennis om servers te beheren. Maar dan heb je je data wel veilig zelf tussen je muren.

Big data-omgevingen

Er is een divers aanbod van big data-omgevingen, uiteenlopend van 'Software as a Service-' tot 'Infrastructure as a Service'-oplossingen. We lopen de belangrijkste even door.

> Microsoft Power BI online: Power BI is een business intelligence (BI)-toepassing die je lokaal kunt installeren als BI-tool of in de cloud kunt gebruiken als SaaS-omgeving. Power BI lijkt op een mix tussen Excel en Access en biedt krachtige visualisatiemogelijkheden.

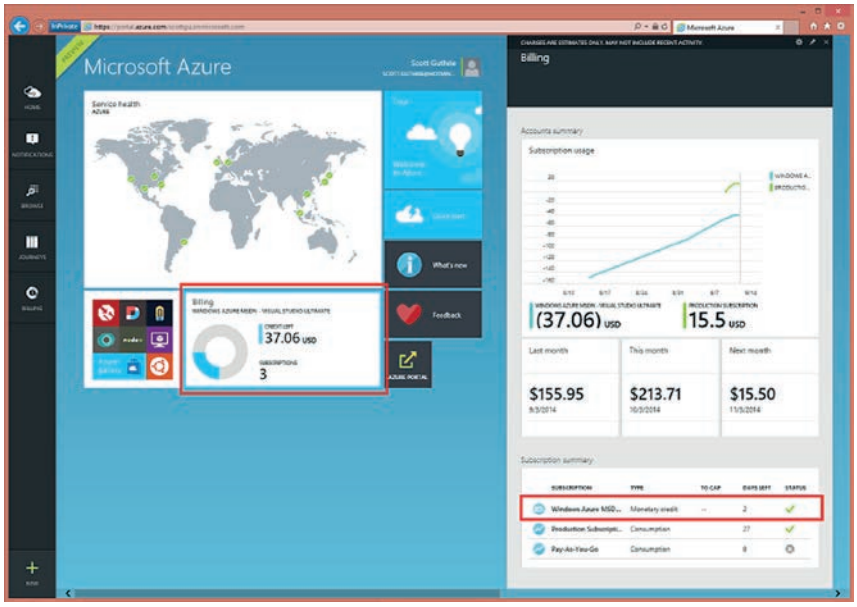
> Microsoft Azure: Azure biedt zowel Microsoft-based producten (SQL-server) als open source-omgevingen aan voor big data-opslag en -analyse (Hadoop). Koppeling met Power BI is mogelijk en er worden zoekmachinefuncties aangeboden.

> Google Cloud: Na het doneren van Hadoop en Map/Reduce aan de open source-community heeft Google de software doorontwikkeld naar 'Big Query': een big data opslagomgeving met een SQL-achtige interface. De infrastructuur die Google biedt is geschikt voor petabytes en je kunt zelf bepalen hoeveel processoren de data analyseren. Een andere dienst van Google Cloud is BigTable, een zeer schaalbare NoSQL-database. Doordat er geen relationeel model aan deze database ten grondslag ligt en zij kan worden opgeslagen over verschillende servers heen, is zij heel flexibel is en berekend op grote hoeveelheden data. Verder biedt Google Cloud een platform voor machine learning en mogelijkheden voor geografische representatie.

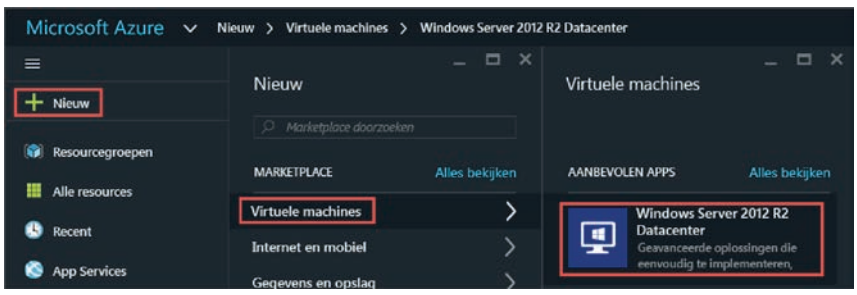
> Amazon Web Services: Amazon biedt een breed scala aan big data-diensten aan. Naast opslag- en analysetoepassingen bieden ze een Business Intelligence (quicksight)-service aan en een grafische processing unit voor het genereren (renderen) van grafische content.

> Oracle: Databaseleverancier Oracle biedt haar producten ook aan in de cloud.

Clouddiensten houden bij hoe lang de server online geweest is of hoeveel dataverkeer er gebruikt is (afbeelding 1). De eerste periode krijg je een gratis voucher om voor een bepaald bedrag een dienst te testen. Is dat bedrag opgesoupeerd, dan wordt het vervolggebruik gefactureerd. In een dashboardomgeving kun je nagaan hoeveel je gebruikt hebt en kun je eventueel beperkingen instellen zodat je server minder verkeer genereert of bijvoorbeeld alleen actief is wanneer je een bewerking doet.



Afbeelding 1. Dashboard van Microsoft Azure met datagebruik en kosten



Afbeelding 2. Aanmaken van een nieuwe server is snel en makkelijk

Snel aan de slag

Behalve dat je organisatie geen eigen systeembeheer meer hoeft te hebben om een omgeving in te richten, biedt een cloudomgeving nog een ander voordeel: je kunt heel snel schakelen. Doordat de aanbieders gebruik maken van standaard hardware kunnen ze ook standaard installaties aanbieden. Met één druk op de knop start (afbeelding 2) en stop je een server die vervolgens binnen enkele minuten geïnstalleerd en wel voor je klaarstaat. Ook servers bijschakelen wanneer je meer capaciteit nodig hebt gaat met hetzelfde gemak. Hierdoor ben je in de cloud altijd voorbereid op toekomstige ontwikkelingen. <

* Zie: www.nu.nl/internet/4361357/google-opent- groot-datacenter- in-eemshaven.html.

Klaas Jan Mollema MSc. (www.zijlmo.nl/mo) is specialisatiecoördinator Business Data Management aan de opleiding Informatica van Hogeschool Leiden.