Wie wil er een data lake?

Kuddegedrag

Tekst **Daan van Beek**

Overheidsorganisaties zoeken naar oplossingen om de groeiende digitale informatiehuishouding duurzaam op orde te brengen én te houden. De zelf in elkaar geknutselde Access-database, het gekoesterde Excelparadijs of misschien zelfs BI-tools zoals Cognos lijken dan opeens hopeloos ouderwets. Hebben we geen modernere tools nodig? Het datameer (*data lake*) is de nieuwste hype in de wereld van data-analytics.

Of het nu gaat om mobiele telefonie, cloud computing, social media of nieuwe datacenters, Nederland loopt graag voorop. Dat is niet vreemd, want de adoptie van nieuwe technologie kan positief bijdragen aan het succes van een organisatie, zo concludeerden onderzoekers van Rabobank Research vorig jaar nog. Niemand wil achterblijven. Ook de overheid niet. Kuddegedrag is een bekend fenomeen. Doe mij dan ook maar zo'n data lake in Apache Hadoop, datalab of hippe AI-engine.

Het grootste verschil tussen een traditioneel datawarehouse en een modern data lake is dat een datawarehouse gevuld is met gestructureerde data. Het data lake daarentegen bevat ruwe, ongestructureerde data die in het oorspronkelijke formaat is opgeslagen. Het behouden van deze detailinformatie kan handig zijn. Maar net als in een echt meer kun je in een data lake verdrinken. Bovendien mag je volgens de nieuwste privacywetgeving (AVG) niet zomaar allerlei vertrouwelijke persoonsinformatie opslaan.

Daan van Beek
www.passionned.nl/over-pg/

Eindbaas Passionned Group & auteur van De intelligente organisatie en Datacratisch werken.

Mensen zijn niet alleen kuddedieren, maar ook verstokte verzamelaars. Gestructureerde, semi-gestructureerde en ongestructureerde data: alles gooien ze in het data lake. Je begint met een helder meer en eindigt met een zompig moeras. Veel mensen denken ten onrechte dat meer (data) altijd beter is. Het inrichten van een data lake is echter nooit een doel op zich. Dataopslag kost ondanks de forse prijsdalingen in de storagewereld nog steeds veel geld. Bovendien kun je een data lake heel makkelijk vullen, maar de juiste data in de juiste vorm naar boven halen, is een ander verhaal. Voor de data-analyse heb je ook dure *data scientists* nodig en die zijn schaars.

Een data lake vullen is heel makkelijk, maar de juiste data in de juiste vorm naar boven halen, is een ander verhaal

Kortom: er zijn misschien wel meer redenen te bedenken waarom je juist niét zou moeten investeren in een data lake. Leveranciers hoor je daar bijna nooit over. Onafhankelijke consultants wel. Misschien is een 'gewone' big database wel afdoende …

Blijf dus kritisch. Voor je het weet zit je opgescheept met een stuwmeer aan data waarvoor niemand in de organisatie zich verantwoordelijk voelt of enig gevoel bij heeft. Het meer ligt er rimpelloos bij. De adoptie stagneert en de investering rendeert niet. Data moet eerst door de mens gaan voordat er een vonk opgloeit. Dat is een noodzakelijke voorwaarde voordat mensen bewegen richting een verbeterdoel. Tot slot: neem elke beslissing met je hoofd, hart én ziel. ●

Zie ook www.od-online.nl

Reactie op column Daan van Beek

Een dure les!

Tekst **Geert-Jan van Bussel**

Daan van Beek legt de vinger op de zere plek: het kuddegedrag van mensen en instanties om zo snel mogelijk achter een nieuwe hype aan te lopen. We konden dat al eerder zien bij de blockchain, die na de hype ook weer met beide voeten op de grond is gezet. Het 'data lake' is zo'n nieuw verschijnsel. Verzamel alle 'ruwe' data die je hebt en krijgt, gooi ze bij elkaar, laat er analysesoftware op los, en voilà: alle vragen kunnen worden beantwoord.

Het willen aanleggen van een data lake is een van ieder gezond verstand gespeende reactie van organisaties omdat ze blijkbaar hun informatiemanagement niet op orde hebben. We hebben hele goede informatiemanagementtechnologie beschikbaar om (automatisch) data op te nemen in onze beschikbare opslagsystemen, om het te kunnen classificeren, metadateren, contextualiseren, beschikbaar stellen, bewaren en (op zijn tijd) te vernietigen. Data verliest snel waarde, blijft niet actueel en het is de vraag wat de zin is om het te bewaren en te analyseren om iets te weten te komen waar je organisatorisch eigenlijk niets mee kunt.

Kwaliteit van data

Om nog maar niet te spreken over de kwaliteit van de meeste 'ruwe' data die een data lake in geloodst wordt. Is die data wel betrouwbaar, volledig, integer, authentiek? We weten het eigenlijk niet. Wat wel bekend is, is dat ongeveer 30% van de data die organisaties in hun systemen opnemen niet aan die criteria van kwaliteit kan voldoen. We weten niet waar die data vandaan komt, wie die data heeft gegenereerd en hoe die data is bewerkt. De context van ontstaan en verwerking is onbekend. We hebben geen zicht op de samenstelling van de data, of ze gemanipuleerd is, of we ze moeten bewaren of vernietigen, of er privacy-implicaties zijn, en ga zo maar door!

Ik weet dat de enorme vloed aan data die op organisaties afkomt informatiemanagement moeilijk maakt. Maar data dan maar ongestructureerd, in verschillende oorspronkelijke formaten en volledig gespeend van context in een stuwmeer van data storten en er maar op vertrouwen dat algoritmes, kunstmatige intelligentie en *machine learning* het probleem van het vinden van de juiste informatie op het juiste moment oplossen, geeft blijk van enorme naïviteit. De ervaring leert dat dat niet gebeurt.

De beperkingen van algoritmes

Algoritmes zijn uitermate subjectief. Ze kunnen veel, maar niet objectief zoeken en vinden. De antwoorden (of voorspellingen) zijn (bewust of onbewust) gepredestineerd door de ontwikkelaar van het algoritme. Zo is in de Verenigde Staten gebleken dat gebruikte algoritmes in de rechterlijke macht de bestaande discriminatie versterkten. Datagedreven rechtspraak blijkt veel subjectiever te zijn dan de traditionele rechtspraak, omdat gedrag en emotie volledig worden uitgesloten. In het stationskwartier van Eindhoven zien we bij de toepassing van 'slimme' technologie eenzelfde verschijnsel optreden, waarbij individuen door technologie als onderdeel van een groep worden geclassificeerd en op basis daarvan worden behandeld. Zonder de zekerheid te hebben dat dat individu daadwerkelijk tot de gekozen groep behoort! Het is ook bekend dat algoritmes zijn ontworpen om het vooraf gewenste antwoord te 'vinden', dat vervolgens als 'waarheid' wordt aangenomen.

Misplaatst vertrouwen

Een data lake is een uiting van ongebreideld vertrouwen in 'slimme' technologie. De technologie echter is een 'black box'. We weten niet wat daarin gebeurt. We kennen het algoritme niet. We weten niet of alle data in het 'lake' wel betrokken wordt bij de analyse. We weten niet wat het doel is en we weten ook niet wie er eigenlijk verantwoordelijk voor is. Misschien willen we dat ook wel niet weten. Maar blijkbaar is dat 'niet weten' voldoende om een miljoeneninvestering te doen in technologie en implementatie.

Net als bij de blockchain, waarbij gebleken is dat uiteindelijk van alle gestarte projecten vanaf 2015 wereldwijd meer dan 80% 'mislukt' is. Maar er is wel veel geleerd, zo wordt verzekerd.
Dat is wel een dure les!!

Dr. G.J. van Bussel

Directeur van Van Bussel Document Services en docent-onderzoeker aan de Hogeschool van Amsterdam.

Zie ook www.od-online.nl