Data Scientist: Het beroep van de 21ste eeuw

Wil van der Aalst (20-12-2013)

Negentig procent van alle wereldwijd beschikbare data is in de afgelopen twee jaar geproduceerd. Ook al zien we al 50 jaar een exponentiële groei van gedigitaliseerde data, pas nu wordt duidelijk wat de invloed is van deze overvloed aan data. Nieuwe diensten en analyses rond Big data zullen onze samenleving veranderen. Alleen organisaties die slim gebruik maken van de stortvloed aan gegevens over het gebruik van producten en diensten zullen overleven. Helaas ligt de nadruk van Big data initiatieven vaak op het genereren en opslaan van enorme hoeveelheden data in plaats van de analyse ervan. Het gaat om het slim gebruiken van data en hiervoor is een nieuwe beroepsgroep nodig: de *data scientist*. Zoals de informatica is voortgekomen uit de wiskunde, ontstaat de nieuwe data science discipline uit een combinatie van bestaande disciplines waaronder informatica, wiskunde, elektrotechniek, sociologie en bedrijfskunde. Er is nu al een tekort aan data scientists en de verwachting is dat dit tekort in de komende jaren alleen maar groter zal worden.

Big Data als de nieuwe olie

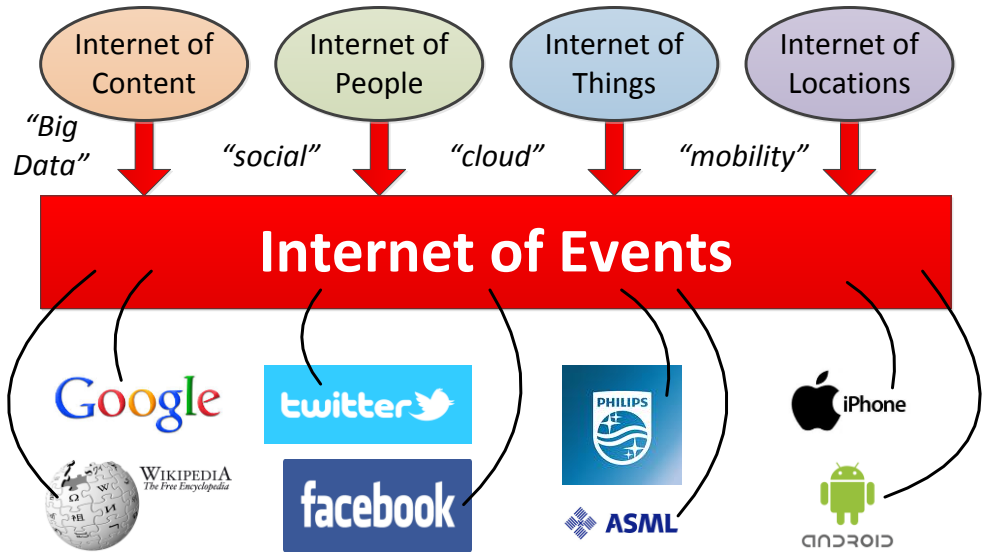
In 2006 poneerde Clive Humby de stelling "Data is de nieuwe olie". In de afgelopen jaren is duidelijk geworden dat het belang van data inderdaad vergelijkbaar is. Gegevens zijn immers een primaire grondstof geworden voor het functioneren van bedrijven en onze samenleving. Voordat olie gebruikt kan worden als brandstof in een auto zijn er diverse stappen nodig: exploratie, winning, transport, raffinage, opslag, en distributie. Soortgelijke stappen komen we tegen bij data: het vinden van relevante data (exploratie), het extraheren van deze ruwe data (winning), transport van data, het bewerken van data door middel van filtering en aggregatie (raffinage), opslag van data, en de distributie van data. Er zijn echter ook belangrijke verschillen tussen data en olie. Het kopiëren en transporteren van data is relatief eenvoudig. Olie kan niet gekopieerd worden (anders zou de prijs niet zo snel stijgen) en transport is kostbaar en tijdrovend. Data is specifiek en in tegenstelling tot olie minder uitwisselbaar. Twee records in een tabel met klantgegevens zijn niet uitwisselbaar, dit in tegenstelling tot olie. Als we 40 liter Euro 95 tanken bij een tankstation, maakt het niet uit waar de brandstof vandaan komt. De benzine is immers niet voorbestemd voor een bepaalde klant op een bepaalde dag. Data die niet specifiek heeft geen betekenis. Het getal 47 heeft alleen betekenis als we weten wat het uitdrukt, bijvoorbeeld de leeftijd van een bepaalde persoon op een bepaald tijdstip.

Geavanceerde analyses zonder data zijn hetzelfde als sportwagens zonder benzine. Er worden op dit moment echter ongelooflijke hoeveelheden data geproduceerd. Volgens sommigen wordt op dit moment in 10 minuten meer data gegenereerd dan in de periode van de prehistorie tot 2003 (5 exabytes). Dit is een direct gevolg van de Wet van Moore die stelt dat het aantal transistors in een geïntegreerde schakeling door de technologische vooruitgang elke 2 jaar verdubbelt. De beschikbare rekenkracht en opslagcapaciteit neemt exponentieel toe. Als transportmiddelen sinds 1970 een soortgelijke ontwikkeling hadden doorgemaakt, konden we nu in 24 milliseconden naar New York

vliegen en een rondje rond de wereld rijden op slechts 38 milliliter benzine. Deze cijfers illustreren de spectaculaire ontwikkelingen op IT gebied en het belang van data science als vakgebied.

Internet of Events (kader)

Steeds meer organisaties, mensen, en machines zijn continu verbonden met het internet en genereren events die gebruikt kunnen worden voor uiteenlopende vragen. Denk bijvoorbeeld aan de röntgenapparatuur van Philips die aan het internet hangt. Door middel van process mining worden de event logs van deze apparaten geanalyseerd om te ontdekken hoe ze echt in ziekenhuizen gebruikt worden, wanneer ze stuk gaan, en waarom ze stuk gaan. Dit kan gebruikt worden om bijvoorbeeld de röntgenbuis te vervangen net voordat deze stuk gaat. We spreken ook wel over het "Internet of Events" als containerbegrip voor het "Internet of Content" (klassieke informatiebronnen zoals webpagina's), het "Internet of People" (sociale media zoals Twitter en Facebook), het "Internet of Things" (apparaten die aan het internet hangen of RFID tags hebben), en het "Internet of Locations" (bijvoorbeeld data met locatiebepaling gegenereerd door smartphones). Denk bijvoorbeeld aan de nieuwe iPhone 5S die meer dan 14 sensoren heeft om onder meer beweging, richting, licht, locatie, geluid, en zelfs vingerafdrukken te bepalen. Steeds opnieuw worden er weer innovatieve manieren gevonden om massaal gegeneerde data nuttig te gebruiken. Op dit moment wordt het mobiele telefoonverkeer al gebruikt om files in kaart te brengen. In de toekomst zullen ruitenwissers wellicht doorgeven dat ze gebruikt worden om zo een betere weersvoorspelling mogelijk te maken. Dit zijn slechts enkele voorbeelden die laten zien dat het groeiende "Internet of Events" steeds weer nieuwe diensten en producten mogelijk zal maken.



De waarde van data

Volgens de website www.twalue.com is de waarde van de auteur zijn Twitter-account $334.37 waard (@wvdaalst). Door de marktwaarde van bedrijven als Twitter, Facebook en Google te delen door het aantal gebruikers is eenvoudig vast te stellen dat het surfgedrag van een gemiddelde tiener een waarde van honderden euro's vertegenwoordigd. Twitter, Facebook en Google leven van data en moeten een kostbare infrastructuur in de lucht houden zonder betalende eindgebruikers te

hebben. Deze investeringen worden betaald voor derden die willen betalen voor data en aandacht (denk aan reclame en diensten). In het algemeen geldt: "Als je niet betaalt, ben je zelf het product".

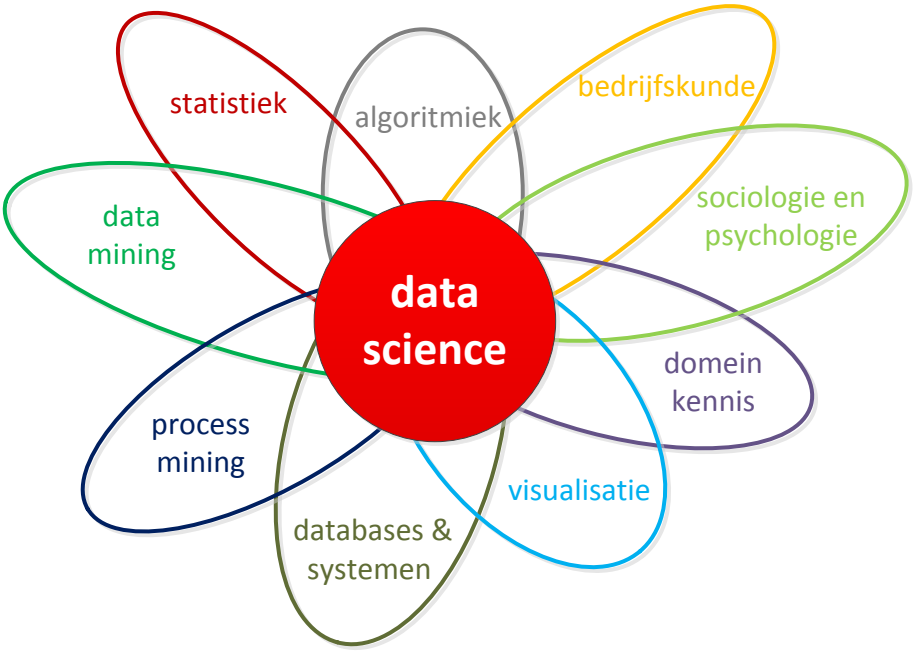
In toenemende mate zien bedrijven de waarde van data. Een recente studie van Bain & Company laat zien dat bedrijven die investeren in data science meer winstgevend zijn en sneller reageren. Organisaties van meer dan 50 medewerkers zullen in de toekomst niet kunnen overleven zonder slim gebruik te maken van de stortvloed aan gegevens.

Data Science Center Eindhoven (DSC/e) (kader)

Het Data Science Center Eindhoven (DSC/e) speelt in op het snel toenemende belang van (Big) data (www.tue.nl/dsce/). De expertises van twintig onderzoeksgroepen van de Technische Universiteit Eindhoven (TU/e) worden in dit onderzoeksinstituut gebundeld om met bedrijven samen te werken en ingenieurs aan de benodigde kennis te helpen. De oprichting van het DSC/e moet binnen enkele jaren leiden tot de eerste zelfstandige bachelor- en masteropleiding op het gebied van data science. De enorme belangstelling van het bedrijfsleven voor data science bleek tijdens de opening waarvoor in korte tijd de 700 beschikbare plaatsen vergeven waren. Het DSC/e werkt samen met bedrijven als Philips, Perceptive software, Adversitement, SynerScope, SAP en Fluxicon.

Een nieuw beroepsprofiel

Het aantal vacatures op het gebied van data science neemt op dit moment sterk toe. Sommigen noemen dit nieuwe beroep "The Sexiest Job of the 21st Centrury" (zie artikelen in bijvoorbeeld USA Today, Harvard Business Review, en Forbes). Dit roept de vraag op wat precies het profiel van een data scientist is. Net als in de jaren 70 en 80 toen de informatica als vakgebied ontstond als antwoord op de komst van computers, is nu door het beschikbaar komen van Big data een nieuw vakgebied aan het ontstaan. Een data scientist moet voldoende kennis van statistiek, data mining, process mining, visualisatie, databases, algoritmiek, en gedistribueerde systemen (denk aan Hadoop) hebben om data om te zetten in nieuwe inzichten, voorspellingen, en aanbevelingen. Het is echter niet voldoende om alleen technische vaardigheden te hebben. Sociologie, psychologie, bedrijfskunde en domeinkennis spelen een belangrijke rol bij het vertalen van ruwe data in proces- en productverbeteringen. Denk bijvoorbeeld aan de recente ophef over de werkwijze van de National Security Agency (NSA). Dit laat zien dat ethiek en privacy belangrijke elementen in de opleiding van de nieuwe beroepsgroep zijn. Data science zou zich moeten richten op maatschappelijke problemen. Zonder maatregelen wordt de gezondheidszorg onbetaalbaar, hiervoor zijn slimme analyses nodig om de juiste afwegingen te maken (meer efficiency zonder kwaliteitsverlies). Slimme elektriciteitsmeters kunnen op afstand uitgelezen worden, zonnepanelen kunnen overdag elektriciteit leveren en elektrische auto's kunnen energie tijdelijk opslaan. Deze ontwikkelingen zullen grote invloed hebben op onze energievoorziening. Voor de afstemming tussen vraag, aanbod en opslag zijn er nauwkeurige voorspellingen en regelsystemen nodig. Kortom, werk aan de winkel voor de data scientist.


statistiek
algoritmiek
bedrijfskunde
data
mining
sociologie en
psychologie
data
science
domein
kennis
process
mining
visualisatie
databases &
systemen