

Big data, Open data en nieuwe kansen voor bedrijvigheid

Door: Hans Wormer

De begrippen Big Data en Open Data hebben de laatste tijd niet over gebrek aan belangstelling te klagen. Big Data zit volgens onderzoeksbureau Gartner nu bovenin de 'hype cycle' en elke week is er wel ergens een seminar of bijeenkomst. Vanuit Brussel klinkt al tijden de opdracht aan nationale overheden om hun data open te maken en Nederland doet dit o.a. via de website Open Data Nederland. Gemeenten beginnen ook steeds meer datasets vrij te geven.

In principe staan de begrippen Big Data en Open Data los van elkaar maar beiden worden steeds vaker met elkaar in verband gebracht. Big Data kan Open Data worden en Open Data kan ook Big Data zijn. Hoe zit dat nu met die relatie, waar komen de begrippen bij elkaar, wat kunnen we ermee en levert dat ook nog wat op?

Open Data

Bij Open Data gaat het de afgelopen jaren met name over wat je zou kunnen noemen *gestructureerde overheidsdata*: het publiekelijk beschikbaar stellen van de administratieve gegevens van overheden. Deze gegevens werden voorafgaand aan de opslag keurig in een datamodel geordend. Vooraf werd duidelijk bepaald over welke subjecten (personen) en objecten (dingen) bepaalde attributen (gegevens) geregistreerd zouden worden. Hoe die subjecten en objecten onderling gerelateerd waren (1:1, 1:n, n:m) en wat de kenmerken waren lag vast. En de definities van de attributen werden keurig vastgelegd in een gegevenswoordenboek, al dan niet met de zegen van de Vereniging van Nederlandse Gemeenten of het ministerie van BZK. Bij gemeenten gaat het bijvoorbeeld om de gemeentelijke basisadministratie personen, sociale zakenadministratie, belastingadministratie, leerlingenadministratie, sportaccommodatieverhuur, maar ook het kadastrale systeem, ruimtelijke beheers gegevens en kaartmateriaal.

Big Data

Big Data is een trend in de IT die juist vooral betrekking heeft op ongestructureerde gegevens waar we tot op heden nog niet zo veel mee konden doen als met de gestructureerde gegevens. Natuurlijk gaat Big Data over veel data maar het zijn ook *data van een andere aard*. De komende jaren zit de grote datagroei niet bij de

administraties van organisaties maar in data die door individuen (sociale netwerken, e-commerce) en apparaten (medische scanners, DNA-sequencers, sensoren, bewakingscamera's, geluidsmeters) worden voortgebracht. Om uit dit soort data waarde te halen is er een heel stelsel aan nieuwe kennis, vaardigheden, tools en technieken in opkomst. En laten we ons niet vergissen. Het is een golf die de komende jaren over ons spoelt.

Tot niet zo lang geleden was 80% van alle gegevens gestructureerd en zat doorgaans in een database. Nu is het al andersom en is meer dan 80% van de gegevens ongestructureerd en dit percentage neemt alleen maar toe. Daarnaast verdubbelt elke twee jaar de hoeveelheid data. Big Data betekent dus ook nieuwe gegevens van een andere aard waarmee we om moeten zien te gaan. We zullen moeten leren om er meer waarde uit te halen dan tot nu toe. De kosten van opslag gaan onvermijdelijk omhoog als er in 2020 16 tot 43 keer zoveel gegevens worden opgeslagen: de vooruitgang in de opslagtechniek zal niet voldoende zijn om deze toestroom helemaal te kunnen opvangen als het om opslagkosten gaat. We weten dus dat het meer gaat kosten maar gaat het ook meer opleveren?

Waarde uit al die data halen wordt dan ook een noodzaak. We hebben toegevoegde waarde nodig uit de groeiende berg aan gegevens en naar mijn overtuiging kan dat ook. Het wordt een nieuwe Goldrush. Wie weet de klompjes goud te halen uit deze grote modderberg aan gegevens?

Nieuwe versus klassieke systemen

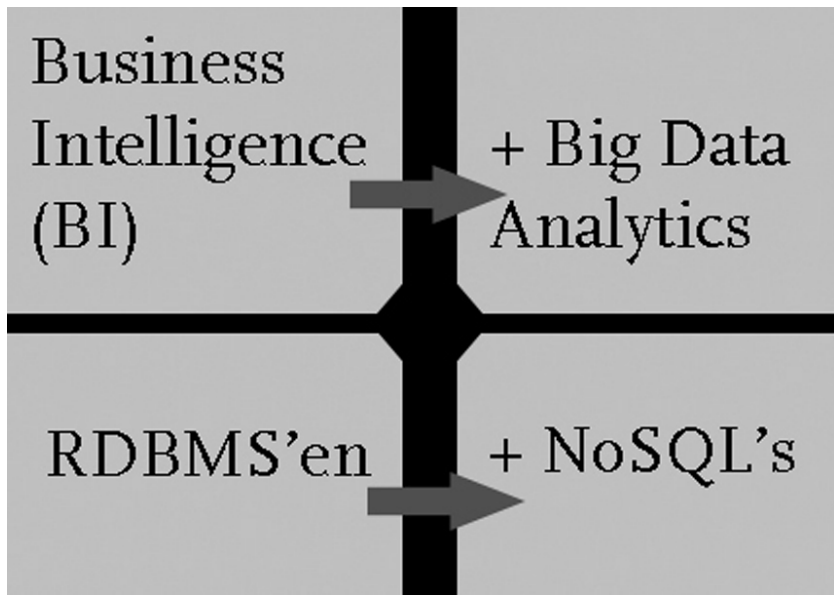
Na ruim 30 jaar dominantie krijgen de relationele databasemanagement systemen (RDBMS'en) concurrentie of beter gezegd aanvulling. De DB2's, Oracle's, SQL-servers van deze wereld blijven belangrijk en vormen nog steeds een miljardenmarkt. Maar er komt een hele nieuwe kolom naast van nieuwe databases. De zogenaamde NoSQL's zoals CouchDB, MongoDB, en MarkLogic. Deels zijn deze databases al een eind op weg een breed enterprise aanbod te bieden, deels moet deze stap nog gezet worden voordat organisaties alle zekerheden en extra dienstverlening kunnen krijgen die bij een enterprise suite horen. Maar dat is op zich geen onbekend verschijnsel, zo ging het indertijd ook met de relationele databasemanagement systemen.

Big Data betekent niet dat de klassieke Business Intelligence (BI) tools en statistische pakketten op de schroothoop kunnen, ze blijven belangrijk. Ik heb het dan over Microsoft SQL Server, IBM Cognos,

SAP Business Objects, SAS, SPSS etc. Zij krijgen gezelschap van nieuwe bedrijven die zich puur richten op Big Data-technieken als social networking mining, tekst mining, video analyse, sentimentanalyse en deep analytics. Alle klassieke BI leveranciers spelen hierop in door binnen hun productaanbod deze technieken op te nemen. Met Big Data Analytics kun je ook rapporteren over wat je ziet op basis van beelden, geluiden, teksten, muisklikken, sensoren in een brug of dijk etc.

Open Data is veelal nog data die uit de al eerder genoemde gestructureerde individuele systemen komt. Het gros van alle Open Data is op dit moment data van de overheid, vanuit het bedrijfsleven of vanuit instellingen is nog maar op zeer beperkte schaal Open Data beschikbaar.

Schematisch kan de ontwikkeling die we gaan doormaken als volgt worden weergegeven:



Gebruik van data: Koppelen versus combineren

Gestructureerde gegevens moet je koppelen en een koppeling moet je (laten) beheren. Iedereen moet over een fiets of iedereen moet over een rijwiel spreken. Of je moet expliciet vastleggen dat een fiets een rijwiel is. In de Big Data wereld is dit niet altijd relevant. Met nieuwe semantische technieken maakt het niet meer uit of iemand over een

fiets of over een rijwiel praat. Als het ene systeem over een fiets praat en het andere over een rijwiel dan hoef je alleen maar te begrijpen dat dit zo is zodat de data uit beide systemen te combineren valt. Ingewikkeld wordt het als je landentabellen, salarisschalen of belastingpercentages veranderen. En al helemaal als je facturen van het systeem voor verhuur van sportaccommodaties er qua structuur heel anders uit zien dan het financieel systeem vereist. Waar in de oude wereld Master Data Management een noodzaak was om hier mee overweg te kunnen kan er met Big Data tools ook anders gewerkt worden waarbij de bronnen geen link of afstemming met elkaar hebben maar wel gecombineerd gebruikt kunnen worden.

Een inmiddels klassieke oplossing om een integratie of koppeling van toepassingen mogelijk te maken en data uit te wisselen en bij elkaar te brengen gaat via het gebruik maken van middleware. Maar middleware is duur want als er “links of rechts” iets verandert dan heb je onderhoud nodig van je middleware. Zonder correcte koppelingen functioneren de toepassingen echter niet meer en werkt vervolgens het werkproces niet efficiënt meer of zelfs helemaal niet. Een leerplichtambtenaar staat dan voor de verkeerde huisdeur omdat de koppeling met burgerzaken niet meer werkt en de verhuizing dus niet is doorgegeven (of verwerkt). De financiële administratie raakt overbelast omdat de 2500 facturen van de sportaccommodatieverhuur opnieuw met de hand moeten worden ingebracht in het financieel systeem. Daarnaast is de manier om uit systemen die zo werken gecombineerde data te krijgen doorgaans opgelost via een datawarehouse, de bron voor Business Intelligence.

Big Data is een wereld waarin het niet zozeer gaat om een koppeling van data maar om vooral ook informatie te halen uit de combinatie van heel diverse soorten gegevens. Een tractor in de precisielandbouw die de beelden van een videocamera gemonteerd op de voorkant combineert met data uit een GPS-verbinding op het dak, eerder verkregen data uit satellieten, de gegevens van de nitraatsensoren in de ploeg en de data uit de motor.

Maar Big Data gaat ook over het combineren van administratieve gegevens met sociale media. Door gegevens van een gemeentelijke afdeling Onderzoek & Statistiek (historie) te combineren met wat bewoners zelf over hun wijk melden op sociale netwerken (actueel) kun je monitoren hoe het met de wijk gaat. Zo heb je een ‘early warning system’ waarmee veel sneller noodzakelijke acties mogelijk zijn. Dit kan bijvoorbeeld problemen voorkomen of hier kan al in een vroeg stadium op worden geanticipeerd. Door dit soort data open te

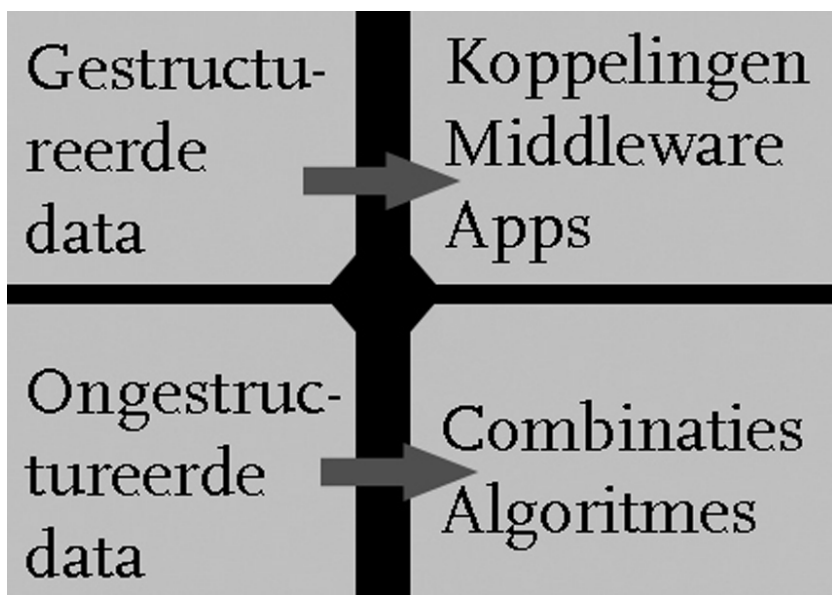
maken creëren we Open Data met een grote toegevoegde waarde waar vervolgens via allerlei apps weer gebruik van gemaakt kan worden.

Bij Open Data krijg je de data 'as is' en moet je doorgaans zelf maar uitzoeken wat er mee zou kunnen. Daarnaast is lang niet altijd een garantie ten aanzien van de kwaliteit of integriteit van de bron. De eigenaar kan de data aanpassen, verplaatsen of zelfs verwijderen. De onvoorspelbaarheid die daar soms mee gepaard gaat maakt dat er bij het gebruik maken van Open Data soms erg creatief te werk gegaan moet worden.

Waarde haal je uit de totale set aan data door brute (reken)kracht, door middel van slimme of soms ingewikkelde algoritmes en soms is het gewoon een eenvoudige correlatie waar niemand aan had gedacht. Met een slim algoritme kun je soms al na een paar muisklikken inschatten of een koper op een website van een webshop een man of een vrouw is. Met nog slimmere algoritmes en een combinatie van data zijn we in staat te zien of iemand jong of oud is, arm of rijk en kunnen we inschatten of iemand echt kooplustig is en hier het aanbod zelfs op aanpassen. Als we maar veel data hebben of real-time beschikbaar (zoals met sensors) en goede algoritmes krijgen we inzichten die tot op heden niet mogelijk waren. We kunnen aan het kleurpatroon van de babycamera "zien" of het goed gaat met het kind. We kunnen aan de uitspraak en woordgebruik van een psychiatrische patiënt afleiden of het richting een crisissituatie gaat. We kunnen aan een medische scan zien of er sprake is van een voorstadium van een carcinoom. Daarbij helpt de machine de menselijke kennis en intuïtie.

Algoritmes die zijn ontwikkeld in een specifieke sector of domein kunnen ook op heel andere terreinen ingezet worden. Zo worden algoritmes om sterren, zwarte gaten en planeten te ontdekken in het heelal ingezet om kankercellen te ontdekken in en brei aan cellen in pathologische beelden.

Er gaat met Big Data een wereld open om tot nieuwe combinaties van gegevens te komen. En dit brengt ons al dicht in de buurt van een heel ander verschil tussen de klassieke datawereld en de Big Data wereld, die raakt aan Open Data.



Creatief en intuïtief versus mathematisch en analytisch

Big Data vraagt naast mathematische vaardigheden (de algoritmes) ook over veel creatieve denkkraft en een goede intuïtie. Juist iets wat veel mensen die zich met Open Data bezig houden in huis hebben. Immers, bij veel Open Data staat bij het beschikbaar stellen niet vast met welk doel het gebruikt zal worden en wat het moet opleveren. Dit is zo belangrijk omdat de zoektocht naar goudklompjes in de modderberg lang niet altijd gaat via een rationeel rechtlijnig proces en vooraf vastgesteld proces.

Als een manager vraagt: “hoeveel moet ik erin steken en wat komt er dan uit?” dan is het antwoord niet altijd eenduidig. Het gaat om het inschatten van wat er allemaal mogelijk is en wat je niet weet moet je leren. Behoedzaam stapjes maken maar er wel voor gaan. Op tijd weten te stoppen, maar soms ook doorzetten. Met beperkt kosten maken beginnen en met een doel voor ogen. Brainstormen. Proberen. Doorgaan na een mislukking. Verrassende wendingen nemen en ongebaande paden inslaan. Dat is meer een aanpak die hierbij past.

‘Open Data mensen’ passen daarom uitstekend in de wereld van Big Data. Sommigen zal het (deels) aan de mathematische skills ontbreken. Maar een creatieve denker kun je de mathematische vaardigheden soms ook wel leren of laten samenwerken met iemand die dat beheerst. Een mathematisch persoon creatief en intuïtief leren

denken is minder eenvoudig. Maar zij kunnen de stap goed maken naar een toepassing waarbinnen de toegevoegde waarde van een algoritme wordt vastgelegd.

De kansen die Big Data en Open Data bieden

Big Data projecten worden nu nog bijna allemaal binnen bedrijven of instellingen uitgevoerd. Daarbij bevinden velen zich nog in de fase van verkennen en het inventariseren en verzamelen van allerlei beschikbare data, doorgaans de eigen data en soms kijkend naar wat er buiten de deur te vinden is. Een belangrijke aanjager is doorgaans het argument dat er commerciële waarde uit gehaald moet worden. Bij Open Data staat veelal de maatschappelijke relevantie centraal. Er zijn echter allerlei combinaties te maken tussen Big Data en Open Data. Door Big Data die nu nog gesloten is open te maken kunnen anderen hier ook van profiteren. Het delen van data door het beschikbaar te stellen, ook door bedrijven en instellingen, biedt enorme mogelijkheden. We kunnen in dit kader dan spreken van Open Big Data.

Naast de bestaande (open) datamarkten zijn er nieuwe datamarkten aan het ontstaan die veel kansen bieden voor nieuwe bedrijvigheid. De rode draad hierin is dat data uit de silo's gehaald wordt, we er bij moeten kunnen en weten wat er mogelijk is of mogelijk zou kunnen. Dat het delen van data en samenwerken tussen partijen die eigenlijk elkaars concurrenten zijn voor iedereen wat oplevert komt wel steeds meer in beeld.

Een recent voorbeeld is logistieke data die transporteurs met elkaar delen waardoor er veel efficiënter gepland kan worden. Met minder ritten tot gevolg wat niet alleen goedkoper is maar ook nog eens maatschappelijk relevant door minder vrachtverkeer op de weg. Dit soort ontwikkelingen zal invloed hebben op de wereld van Open Data en de mensen die hiermee actief zijn. Een mooi gegeven is dat de competenties van diegene die actief zijn met Open Data juist goed aansluiten bij wat de wereld van Big Data vraagt.

Kijken we naar de data *an sich* en naar de nieuwe tools en technieken die beschikbaar zijn dan zien we vooral dat het gaat om de toegevoegde waarde te halen uit *alle data*, zowel de gestructureerde data als de ongestructureerde data (beelden & scans, spraak, muisklikken, sensordata). Het is een enorme aanjager voor allerlei bedrijvigheid en nieuwe producten en diensten met kansen voor entrepreneurs die hier op inspringen. Het vinden van nieuwe combinaties en nieuwe arrangementen voor toegevoegde waarde staat

voorop. Ik wens u een creatieve, intuïtieve en ondernemende toekomst toe!



Hans Wormer (1961) is sinds eind 2011 intern programmamanager Almere DataCapital bij de Economic Development Board Almere.

Het meerjarige economische ontwikkelingsprogramma Almere DataCapital streeft naar een ecosysteem van bedrijven en instellingen die toegevoegde waarde diensten leveren op het terrein van Big Data.

Daarvoor was hij hoofd ICT van de gemeenten Almere, Ede en Lelystad. Hij werkte als management consultant bij onder andere Andersson, Elffers, Felix, CMG en Berenschot. Daar specialiseerde hij zich op het terrein van organisatie, plannings- en sturingsvraagstukken op het terrein van de overheids-ICT.

Dit artikel is geschreven met dank aan Oscar Wijsman.