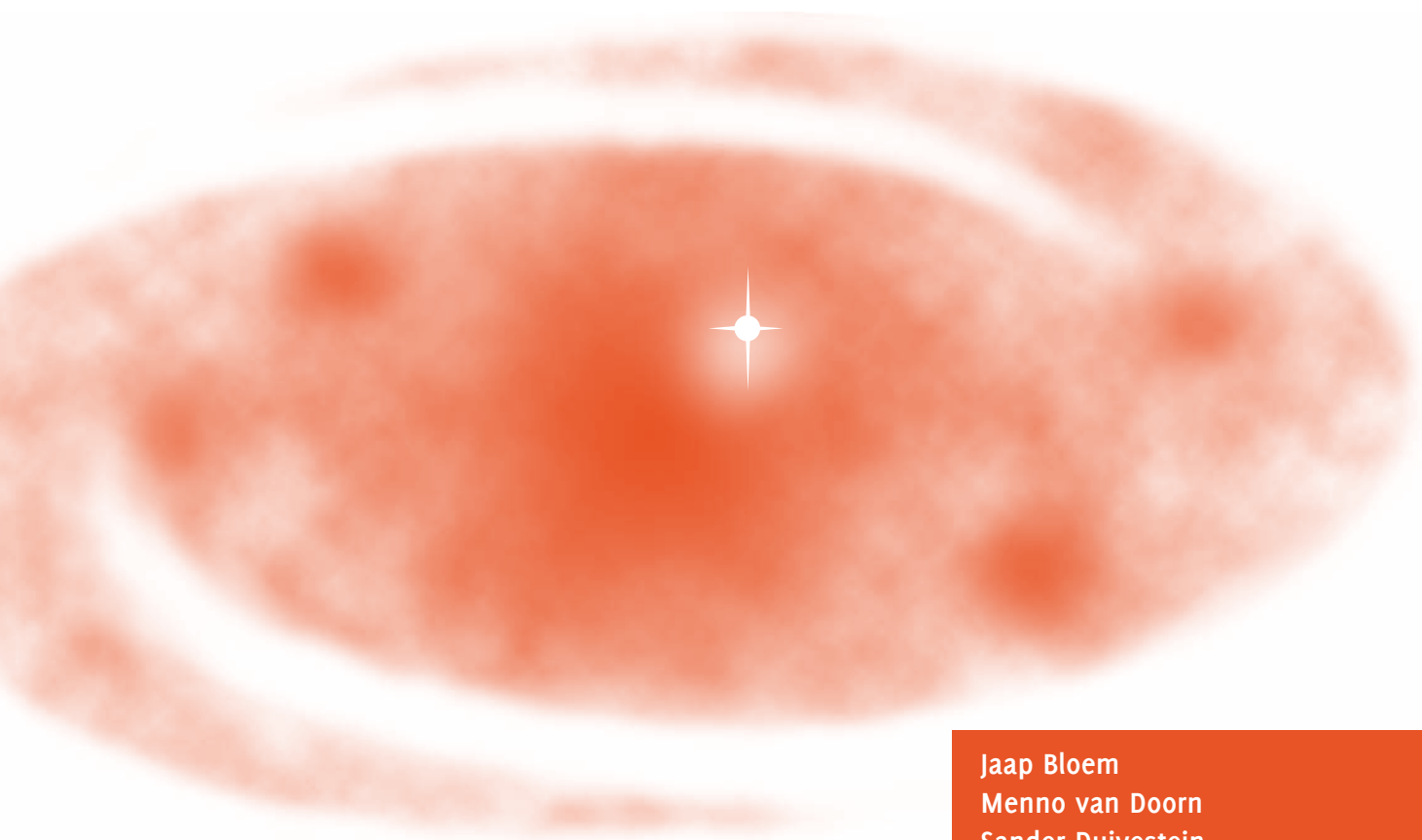


VINT-onderzoeksnotitie ① van 4
VINT-onderzoeksnotitie ② van 4
VINT-onderzoeksnotitie ③ van 4
VINT-onderzoeksnotitie ④ van 4

Helderheid creëren met Big Data



Jaap Bloem
Menno van Doorn
Sander Duivestein
Thomas van Manen
Erik van Ommeren



SOGETI

VINT | Vision • Inspiration • Navigation • Trends

vint.sogeti.com

vint@sogeti.nl

Inhoud

De Big Data-onderzoeksnotities van VINT	3
1 Digitale data als nieuwe industriële revolutie	4
2 Totaal datamanagement in elke organisatie	7
3 Doe mee aan onze Big Data-discussie op www.sogeti.com/vint/bigdata/questions	9
4 Waarom eigenlijk 'big'?	11
5 Het belang van Big Data	13
6 Big Data is Extreme Information Management	16
7 Big Data in organisaties anno 2012	18
8 Met Big Data van Big Science naar Big Business	22
9 Big Data als nieuw Data Science-tijdperk	25
10 Kennisachterstand inlopen is essentieel	26
11 Big Data in klinkende munt	28
Samenvatting en Social Business Analytics	29
Literatuur en illustraties	31



2e druk augustus 2012

© 2012 Sogeti VerkenningInstituut Nieuwe Technologie VINT

Boekproductie LINE UP boek en media bv, Groningen



VINT | Vision • Inspiration • Navigation • Trends

De Big Data-onderzoeksnotities van VINT

Sinds 2005, toen het begrip Big Data pas werd gelanceerd – opmerkelijk genoeg vanuit O'Reilly Media, dat een jaar eerder met Web 2.0 was gekomen – is Big Data een steeds actueler onderwerp geworden. Qua technologieontwikkeling en businessadoptie is het Big Data-veld sterk in beweging, en dat is een understatement.

In deze startnotitie over Big Data van in totaal vier onderzoeksnotities geven we antwoord op de vraag wat Big Data eigenlijk is, waarin het verschilt van bestaande dataudiging, hoe de transformatieve potentie van Big Data wordt ingeschat en hoe het anno 2012 zit met adoptie en plannen daarvoor.

VINT beoogt daarin helderheid te scheppen door ervaringen en visies in perspectief te presenteren: onafhankelijk en aangekleed met voorbeelden. Maar lang niet alle antwoorden zullen kunnen worden gegeven, sterker: er zullen meer vragen bij u opkomen. Bijvoorbeeld over de roadmap die u voor Big Data wilt hanteren. Over management en governance. Of over hoe u uw organisatie misschien moet herinrichten. Over de privacy-issues die Big Data-analyse oproept, zoals ten aanzien van social analytics. En over wat nieuwe algoritmes en systemen ons wellicht nog zullen brengen.

De nieuwe datafocus is een zoektocht met veel vragen aan het begin en zeker ook gedurende de reis. Daarom wisselen we graag met u van gedachten: online op www.sogeti.com/vint/bigdata/questions en natuurlijk in persoonlijke gesprekken. Door actief deel te nemen aan de discussie helpt u uzelf en ons om de gedachten ten aanzien van Big Data aan te scherpen. Om door voortschrijdend inzicht te komen tot heldere en verantwoorde beslissingen. Samen bepalen we zo de concrete invulling van de komende drie onderzoeksnotities na deze kick-off over Big Data.

Ter inspiratie treft u in deze notitie zeven vragen aan waarover we graag uw mening vernemen. In de pdf van dit document kunt u op de betreffende buttons klikken. Vervolgens wordt u direct naar de discussie in kwestie geleid.



Join the
conversation

1 Digitale data als nieuwe industriële revolutie

In 2012, zo'n veertig jaar na de start van het informatietijdperk, zijn alle ogen gericht op de basis daarvan: op digitale data. Dat lijkt misschien weinig nieuws, maar de toevloed van verschillende datatypen plus de snelheid waarmee die trend tot in lengte van dagen zal doorzetten, is opzienbarend. Data, data en nog eens data: we bevin-den ons in een uitdijend data-universum, vol met onontdekte verbanden. Dat is niet abstract en algemeen, maar specifiek en heel concreet, want elk nieuw inzicht kan het begin zijn van een goudmijn. Zo simpel en fundamenteel is deze data-explosie dat Joe Hellerstein van Berkeley University spreekt van een nieuwe industriële revolutie. Een revolutie op basis van digitale data, die de motor zijn van compleet nieuwe bedrijfs-economische en maatschappelijke mogelijkheden.

Begin mei 2012, op de Cloud Computing Conference van Goldman Sachs, presenteerde Shaun Connolly van Hortonworks data als 'The New Competitive Advantage'. Connolly geeft daar zeven redenen voor, twee businessgeoriënteerde, drie technische en twee financiële:

Businessredenen

1. Nieuwe innovatieve businessmodellen worden mogelijk.
2. Nieuwe inzichten die concurrentievoordeel opleveren.

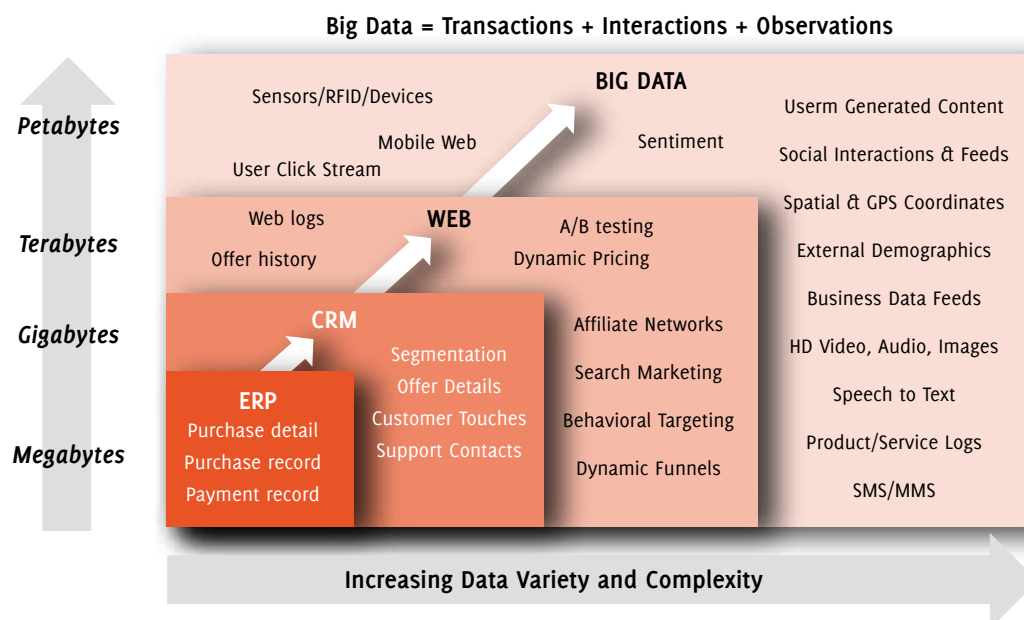
Technische redenen

3. De winning en opslag van data blijven allebei exponentieel groeien.
4. Data in verschillende vormen treffen we overal aan.
5. Traditionele oplossingen voldoen niet aan nieuwe complexiteitseisen.

Financiële redenen

6. De kosten van datasystemen blijven stijgen als percentage van het IT-budget.
7. Nieuwe standaardhardware en open-sourcesoftware brengen kostenvoordelen met zich mee.

Vanwege deze combinatie schuiven volgens Connolly daarom nu de traditionele datawereld van de businesstransacties en die van de interacties en observaties ineen. Onder de formule *Big Data = Transactions + Interactions + Observations* is het doel: meer business, een hogere productiviteit en nieuwe commerciële kansen.



Source: Contents of above graphic created in partnership with Teradata, Inc.

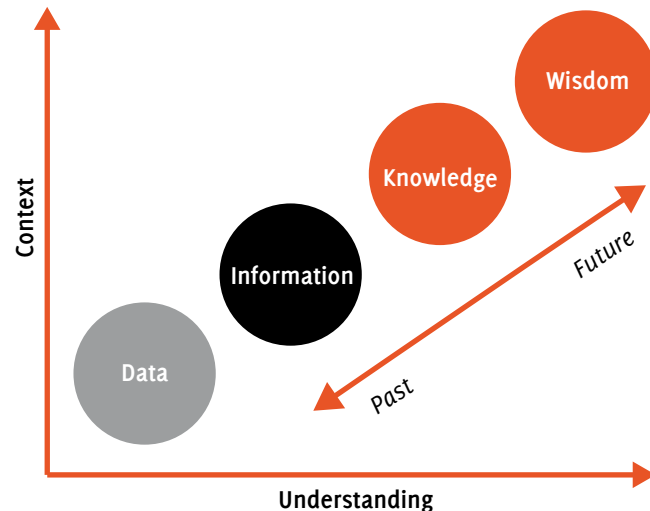
Digitale data als basis

We leven momenteel in ten minste drie tijdperken met digitale data als basis: het informatietijdperk, het sociale tijdperk en het Big Data-tijdperk. Zo staat het in de *List of Periods* van Wikipedia, die onze hele historie bestrijkt. De explosieve datagroei komt werkelijk overal vandaan. Van businesstransacties, van mobiele devices, van sensoren, van sociale en traditionele media, van hd-video, van cloud computing, van aandelenmarkten, van webclicks en ga zo maar door. Al die data worden gegenereerd in de interactie tussen mensen, machines, applicaties en combinaties daarvan. Wie er zich weinig bij kan voorstellen, moet maar eens een kijkje nemen in een publiek toegankelijke uithoek van ons nieuwe data-universum: het Linked Open Data-domein op <http://lod-cloud.net>. De visualisatie van dat datanetwerk en de onderdelen daarvan maakt meteen duidelijk wat er wereldwijd aan de hand is, in alle sectoren van de economie, de maatschappij en de wetenschap, en wederom ook weer in combinatie.

Alles is informatie

Organisaties bestaan dankzij informatie en in de wetenschap is er tegenwoordig een stroming die zegt dat eigenlijk alles informatie is. Aan die informatie liggen data ten grondslag en hoe meer relevante gegevens we hebben, des te beter zijn we in staat om de meest uiteenlopende zaken te begrijpen en beter te kunnen anticiperen op wat komen gaat. Dat is nodig om de juiste beslissingen te kunnen nemen, zeker in deze tijden van hyperconcurrentie en van crisis. De ongekende data-intensiteit in het Big Data-tijdperk, dat we uitgerekend nu zijn binnengestapt, is een zegen, zeggen de voorstanders. Immers, de analyse van complete datasets is per definitie de enige echte manier om volledig te kunnen begrijpen en voorspellen. Daar is geen speld tussen te

krijgen, en dankzij moderne en betaalbare IT – hardware, software, netwerken, algoritmes en applicaties – kan de analyse van complete datasets nu ook daadwerkelijk van de grond komen.



Join the conversation

Vraag 1
Winnen feiten het definitief van intuïtie?

www.sogeti.com/vint/r1q1

Big Data-case: verlies van klanten

Tot voor kort waren we gedwongen om steekproeven te nemen en daar analyses op los te laten. Maar hoe sample je een netwerk of een verzameling subnetwerken? Als een telecomprovider inzicht wil in de omstandigheden waaronder een subnetwerk van vrienden en kennissen plotseling naar een concurrent verhuist (*churn*), dan hebben we waarschijnlijk te maken met meer dan 10 miljoen bestaande en recente abonnees, met informatie over hun gebruik, hun uitgaven aan diensten en wie hun vrienden zijn, dus wie er hoe vaak werd gebeld of ge-sms't. We hebben dan te maken met kantelpunten: een deel van het subnetwerk loopt weg en vervolgens gaat de rest ook. Op zich heel voorstelbaar: als collega's of vrienden zijn overgestapt en bij de concurrent nu beter en goedkoper af zijn, dan is er een sociale en economische prikkel om zelf ook te vertrekken. Een provider wil dit natuurlijk voorkomen en moet daarbij alle data in beschouwing nemen. Wordt er namelijk een aselechte steekproef genomen van een miljoen klanten bijvoorbeeld, dan zijn de vriendenkringen waar het om gaat niet meer intact en daarmee ontvalt de basis om te kunnen voorspellen. Sampling is hier dus niet de methode. Om goed zicht te krijgen op de kantelpunten moeten we alle data in hun samenhang bekijken. Vervolgens kunnen we op basis van zich ontwikkelende patronen in een vroeg stadium anticiperen op 'churn' met gerichte retentieacties en -programma's.

Fraudedetectie

Een ander voorbeeld waarbij we de complete dataset nodig hebben, is fraudedetectie. Het signaal is dan zo klein dat er niet met steekproeven kan worden gewerkt tot het signaal is geïdentificeerd. Daarom moeten ook hier alle data worden geanalyseerd. Het wordt met recht een overduidelijk geval van Big Data als er bij fraude gekeken wordt naar ‘collusie’, naar illegale samenwerking die erop gericht is om anderen zoveel mogelijk te hinderen en kapot te maken, zoals in de casinowereld. Churn en fraude-detectie zijn een greep uit de toepassingsmogelijkheden van Big Data Analytics (zie ook paragraaf 7).

Big Data Success Stories

Onder de titel *Big Data Success Stories* publiceerde IBM In oktober 2011 een illustratieve reader met twaalf verschillende cases, om te laten zien wat Big Data nou eigenlijk betekent. Die vraag beantwoorden wij hier ook. In de volgende paragraaf al om precies te zijn en in paragraaf 7: ‘Big Data in organisaties anno 2012’. Voor dit moment houden we het erop dat Big Data-analyse verder gaat dan wat er met traditionele relationele databases mogelijk is en dat de trend is om steeds meer nieuwe datatypen in beschouwing te nemen. Met alle empirische data die voor het oprapen liggen, lijkt het erop dat we in de toekomst steeds vaker alleen maar op een slimme manier naar de feiten hoeven te kijken, waardoor uiteindelijk zelfs veel theorie- en modelvorming als tussenstappen kunnen worden overgeslagen. Deze Big Data-beloofte werd al verwoord in 2008, in het artikel ‘The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.’

2 Totaal datamanagement in elke organisatie

Big Data, de enorme datagroei die in alle opzichten hand over hand toeneemt, vraagt om *totaal datamanagement* in elke organisatie. Dat zegt onder meer The 451 Group.

Overall komen steeds meer data vandaan: van traditionele transactionele data tot sensoren en RFID-tags, en niet te vergeten sociale media, internet, clouds en mobiele devices. Gestructureerd, semigestructureerd of ongestructureerd, het maakt niet meer uit, want de IT-mogelijkheden voor dataverwerving en -verwerking, en de betaalbaarheid daarvan, groeien vrolijk mee.

Datagroei overstijgt de Wet van Moore

Hoewel de datatoevloed tegenwoordig de Wet van Moore overtreft – elke 18 maanden het dubbele aantal transistors op een chip tegen steeds lagere kosten – zijn we toch in staat om daar zinvol mee om te gaan. Dat is mogelijk door geavanceerde hard-

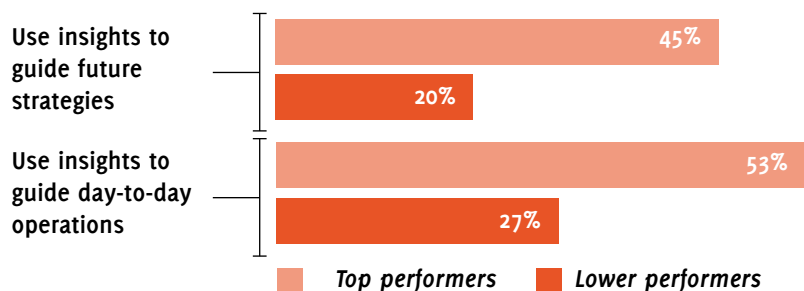
ware, software, netwerken en datatechnieken. We zijn kortom in staat om met iedereen in de organisatie het hele dataveld te bespelen. Wie dat goed doet, zegt Gartner onder meer in de presentatie *Information Management Goes 'Extreme': The Biggest Challenges for 21st Century CIOs*, kan 20 procent beter presteren dan de concurrentie die het nalaat:

'Through 2015, organizations integrating high-value, diverse, new information types and sources into a coherent information management infrastructure will outperform their industry peers financially by more than 20%.'

De spelregels blijven hetzelfde, maar de tactiek verandert. Net als in het verleden willen we informatie opwerken uit ruwe data en daar intelligente nieuwe inzichten uit winnen die betere en snellere businessbeslissingen mogelijk maken. Big Data is als het ware een oproep aan organisaties om hun Business Intelligence-inspanningen op een radicaal hoger plan te brengen. Op basis van de juiste technologie, de juiste processen, de juiste rollen en de juiste kennis en kunde, Data Science geheten. Een en ander door de hele organisatie heen en volcontinu.

Big Data is een nieuwe fase

Daarmee is Big of Total Data een nieuwe fase in de trend die in 2010 werd gekwantificeerd door MIT Sloan Management Review en het IBM Institute for Business Value in de studie *Analytics: The New Path to Value*. Bijna de helft van de best presterende organisaties, zo bleek toen, gebruikten digitale data voor hun langetermijnstrategie tegen slechts een vijfde van de underperformers. Voor de dagelijkse operatie was dat zelfs ruim de helft van de topperformers tegen ruim een kwart van de lager presterende organisaties. Het advies is dus om de analyse van digitale data in haar volle omvang met voorrang aandacht te geven.



Natuurlijk willen organisaties zo'n advies niet in de wind slaan, te meer omdat het logisch voortbouwt op bestaande Business Intelligence en de doelstelling van economisch gewin. Maar er moet nogal het een en ander aan vereisten worden geregeld en



Vraag 2
Hoe koppelt u realtime Big Data aan de operationele sturing van uw bedrijf?

www.sogeti.com/vint/r1q2

ingericht. Behalve op de potentie en beloften van Big Data gaan we ook daar in deze onderzoeksnotitie op in. De ambitie van alle Big Data-notities is om met u over deze belangrijke materie van gedachten te wisselen en om samen de mogelijkheden voor u te exploreren.

3 Doe mee aan onze Big Data-discussie op www.sogeti.com/vint/bigdata/questions

De Big Data-vraagstukken waarover we naar aanleiding van de onderzoeksnotitie die voor u ligt, graag ideeën en ervaringen uitwisselen, zijn om te beginnen drieërlei:

- A. Uw Big Data-profiel: hoe ziet dat eruit?
- B. Tien Big Data-managementuitdagingen: wat zijn uw issues?
- C. Vijf vereisten voor uw Big Data-project: bent u er klaar voor?

Nota bene

De interactie over deze en aanverwante zaken vindt plaats op onze website, maar zeker ook tête-à-tête wat ons betreft. Wekelijks zullen wij nieuwe onderzoeksinzichten met u delen via blogposts, e-mail- en twitter-alerts. Het begeleidende videomateriaal met toonaangevende deskundigen is bedoeld als inspiratie om de hele Big Data-thematiek vanuit verschillende invalshoeken verder te doordenken en bediscussiëren.

A. Uw Big Data-profiel: hoe ziet dat eruit?

Bij Big Data gaat het om ongewoon grote, vaak gedistribueerde verzamelingen van semigestructureerde of ongestructureerde data. Die zijn bovendien vaak incompleet en niet makkelijk toegankelijk. Ongewoon groot betekent het volgende, en dat is afgemeten aan de uiterste grenzen van de huidige standaard-IT en relationele databases: petabytes aan data of meer, miljoenen mensen of meer, miljarden records of meer, of een complexe combinatie hiervan. Met minder data en een grotere complexiteit kunt u dus ook een serieuze Big Data-uitdaging hebben, zeker wanneer uw tools, kennis en kunde niet up-to-date zijn. Bovendien bent u in het laatste geval ook niet goed voorbereid op toekomstige dataontwikkelingen. Semigestructureerd of ongestructureerd betekent dat de verbanden tussen data-elementen niet helder zijn en er waarschijnlijkheden moeten worden bepaald.

B. Tien Big Data-managementuitdagingen: wat zijn uw issues?

1. Hoe gaat u om met de toenemende hoeveelheden semigestructureerd en ongestructureerde data? Naar schatting 80 procent van de data die het meest waardevol zijn voor organisaties, valt buiten de traditionele relationele datawarehousing en datamining waar Business Intelligence tot op heden voornamelijk op was gericht.
2. Die nieuwe waardevolle data zijn afkomstig van een range aan verschillende datatypen en databronnen. Weet u welke voor u van belang zijn en heeft u een plan om ze strategisch in te zetten?
3. Heeft u zicht op de complexiteit van uw data, afzonderlijk en in combinatie? En weet u wat u allemaal in welke volgorde wilt weten? Nu en in de toekomst?
4. Nieuwe inzichten verkregen uit de combinatie van gestructureerde en ongestructureerde data zijn soms beperkt houdbaar. Weet u wat de gewenste snelheid van de verwerking en analyse van verschillende data en datacombinaties is? Welke vraagstukken die u wilt oplossen, vragen om een realtime aanpak? Denk er goed om dat voor realtime beslissingen ook realtime processen nodig zijn.
5. Heeft u nagedacht over de kosten van uw nieuwe datamanagement? Hoe zijn die opgebouwd: naar datadomeinen, technologie en expertise bijvoorbeeld?
6. De opslag van alle data die u wilt analyseren en bewaren stelt hoogstwaarschijnlijk nieuwe eisen aan uw IT-infrastructuur. Bent u daar planmatig mee bezig en kijkt u ook naar performance?
7. Hoe zit het met de beveiliging van alle data?
8. De opslag en beveiliging van Big Data zijn samen van groot belang inzake uw datagovernance, risicomanagement en compliance. Betreft u de juiste afdelingen en functionarissen bij uw Big Data-activiteiten?
9. Nieuwe businessinzichten opwerken vanuit grote datahoeveelheden vereist een totaalaanpak in de organisatie. Daarvoor zijn nieuwe kennis en kunde nodig. Zijn die aanwezig en hoe worden ze geborgd en verder ontwikkeld?
10. Weet u wat uw Big of Total Data-inspanningen betekenen voor uw energieverbruik?

C. Vijf vereisten voor uw Big Data-project: bent u er klaar voor?

Vanuit de managementuitdagingen hierboven vatten we hier vijf fundamentele voorwaarden voor u samen die nodig zijn om aan een concreet Big Data-project te kunnen beginnen:

1. Uw organisatie beschikt over de juiste mindset en cultuur. Er is door de organisatie heen geen twijfel over het nut van een Big of Total Data-aanpak, u weet waarmee u wilt beginnen en wat de doelstellingen voor de toekomst zijn.
2. Er is voldoende managementsupport en het is duidelijk wie de executive sponsors zijn.



Vraag 3

Wat is de beste aanpak om voor Big Data een gewillig oor bij het management te vinden?

www.sogeti.com/vint/r1q3

3. De benodigde expertise en ervaring inzake Data Science en Big Data-frameworks en -tools zijn aanwezig en geborgd.
4. Er is voldoende budget gealloceerd voor de benodigde trainingen om expertise en ervaring, mindset en cultuur te vormen en te laten beklijven.
5. Er zijn voldoende resources en budget voor de ontwikkeling van Big Data-toepassingen en u heeft daarbij de juiste partners en leveranciers geselecteerd.

4 Waarom eigenlijk 'big'?

We noemen iets 'big' – Big Mac of Big Mama bijvoorbeeld – om er de aandacht op te vestigen. Maar als we er geen plaatje bij hebben, dan roept 'big' meteen ook fundamentele vragen op. Precies dat is het geval bij Big Data, maar ook bijvoorbeeld bij het aanverwante Big Science. Hoe groot is Big Data bijvoorbeeld eigenlijk en ten opzichte van wat?



'Big' is niet zo'n handige term

De analisten van Forrester en Gartner zijn het daarom roerend met elkaar eens: bij nader inzien is 'big' misschien niet zo'n handige naam voor de datavloed die hand over hand toeneemt. Beide bureaus, en anderen met hen, houden het liever op 'extreme' in plaats van 'big'. Vanuit de statistiek heeft die term ook een langere historie.

In het dagelijks leven slaat 'big' op heel concrete oversized verschijnselen. Maar onvoorstelbare hoeveelheden digitale data onttrekken zich nu eenmaal zich aan het oog. Bovendien is er meer aan de hand dan hoeveelheid alleen.

Big Data en Web 2.0

Het is geen toeval, dat O'Reilly Media de term Big Data een jaar na Web 2.0 introduceerde, want inderdaad hebben veel waardevolle Big Data-situaties betrekking op

consumentengedrag. Web 2.0 was aanleiding om de interactie op internet opnieuw te doordenken en een paar flinke stappen verder te brengen. Op dezelfde manier vraagt de kwalificatie Big Data aandacht voor de businesskansen van enerzijds de toevloed aan gegevens en anderzijds de nieuwe technologieën, technieken en methoden die daarop zijn gericht.

Een simpel antwoord

Zoals gezegd overstijgt de toename van data momenteel de Wet van Moore. Verschillende typen data in combinatie met de benodigde analysesnelheid zijn nu de grootste uitdaging. Samen met het beperkte aantal mensen dat goed met Big Data overweg kan. In 2020 zal er 35 zettabyte aan digitale data zijn. Dat is een stapel dvd's die van de aarde tot halverwege Mars reikt. Facebook heeft 70 petabyte en 2700 multiprocessornodes; de zoekmachine Bing 150 petabyte en 40.000 nodes. Maar wat betekent Big Data precies voor organisaties? Big Data kunnen we vanuit de vraagstukken benaderen, maar natuurlijk ook vanuit de oplossingen. Het simpelste antwoord is van Forrester Research en luidt als volgt:

'Big Data: Techniques and Technologies that Make Handling Data at Extreme Scale Economical.'

Net als The 451 Group en Gartner maakt ook Forrester dus geen onderscheid tussen Big of Little Data. Vergeleken met vroeger zijn er veel nieuwe en andere data bijgekomen, en dat gaat maar door, maar data zijn data. Ze gaan hand in hand en alleen met de goed doordachte integratie van het hele spectrum van verschillende ordegrootten komen we echt verder. We hebben te maken met één dataspectrum, één continuüm, en dat moeten organisaties stap voor stap strategisch exploreren.

Eén groot datacontinuüm

Ruim dertig jaar eerder gold dit precies zo voor de groei van wetenschappelijke activiteit: klein en groot. In zijn boek *Reflections on Big Science* (1967) schreef atoomwetenschapper Alvin Weinberg toen:

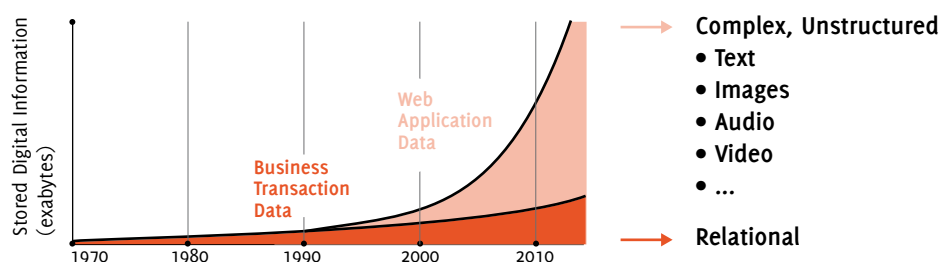
'The scientific enterprise, both Little Science and Big Science has grown explosively and has become very much more complicated.'

Zo is het geconstateerd ten aanzien van science en precies zo geldt het ook voor data. Lees maar mee met wat Chirag Metha zegt. Als Technology, Design & Innovation Strategist was Metha bij SAP verbonden aan het Office of the CEO:

'Today, technology — commodity hardware and sophisticated software to leverage this hardware — changes the way people think about small and large data. It's a data continuum. [...] Big Data is an amalgamation of a few trends — data growth of a magnitude or two, external data more valuable than internal data, and shift in

*computing business models. [...] **Big Data is about redefining what data actually means to you.** [...] This is not about technology. This is about a completely new way of doing business where data finally gets the driver's seat.'*

Big Data wil dus helemaal niet zeggen dat we bijvoorbeeld Little of Small Data, of Medium, Large enzovoort maar moeten vergeten, integendeel. Waar het om gaat, is dat we tegenwoordig alle data in alle hoedanigheden kunnen en moeten bekijken. Technologisch is dat mogelijk en businesswise is het wenselijk, zo niet noodzakelijk.



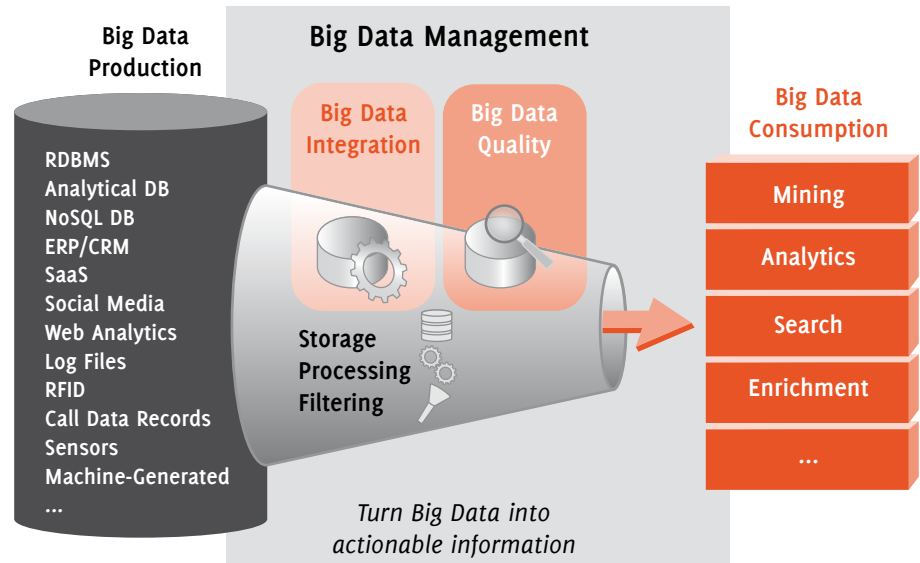
Met name is dat het geval omdat 80 procent van alle nieuwe data niet-relatueel of ongestructureerd is en in combinatie met transactiedata voor organisaties de meest waardevolle informatie bevat. Volgens sommigen hoeven lang niet alle op het eerste gezicht ongestructureerde data dat te blijven en kunnen zulke data met relatief weinig moeite vaak toch in een structuur worden ondergebracht.

5 Het belang van Big Data

Waarom we al die data willen hebben en bekijken is duidelijk. Onder meer sociale media, web analytics, logfiles en sensoren leveren waardevolle informatie op, de kosten van IT-oplossingen blijven dalen en de computerverwerkingskracht neemt toe. Informatieoverdaad heeft daarmee grotendeels afgedaan: in principe zijn organisaties in staat om de informatievloed te managen en te gelde te maken. Wie excelleert in de verwerving, de verwerking en het management van waardevolle data, zegt Gartner, zal financieel gezien 20 procent beter kunnen presteren dan de concurrentie.

In organisaties is het aandeel van ongestructureerde data, zoals documenten, e-mail en afbeeldingen, momenteel 60 tot 80 procent. Van alle data-analyse in organisaties bevat nu 5 tot 15 procent een sociale component om de gestructureerde data te verrijken. Dat moet omhoog, niet in de laatste plaats vanwege alle externe data die nog kunnen worden meegenomen.

Ook het Internet of Things wordt een steeds rijkere bron van data. Op dit moment, zegt Cisco-CTO Padmasree Warrior, zijn er 13 miljard devices verbonden met internet



en in 2020 zullen dat er 50 miljard zijn. IDC verwacht dat er tegen die tijd ruim 1 biljoen sensoren met het internet zullen zijn verbonden. Alle bijbehorende datastromen kunnen interessante inzichten opleveren ten behoeve van betere businessbeslissingen.



Vraag 4
Wat is de belangrijkste nieuwe spelregel voor organisaties ten aanzien van Big Data?

www.sogeti.com/vint/r1q4

We staan aan het begin van Big Data

Banken behoren tot de top van organisaties die zich met Big Data bezighouden, maar in het rapport met de veelzeggende titel *Big Data: Harnessing a Game-changing Asset* van de Economist Intelligence Unit geeft Paul Scholten, COO Retail & Private Banking bij ABN AMRO, volmondig toe dat de bank nog in het explorerende stadium is als het gaat om de uitnutting van met name ongestructureerde sociale data:

'We are used to structured, financial data. [...] We are not so good at the unstructured stuff. [...] The company is just beginning to understand the uses of social media, and what might be possible in terms of improving customer service.'

Mark Thiele zegt het heel pakkend door Big Data anno 2012 te vergelijken met de start van het world wide web. Thiele is Executive vp Data Center Technology bij Switch, de operator van het SuperNAP-datacenter in Las Vegas, het grootste en meest krachtige ter wereld:

'Big Data today, is what the web was in 1993. We knew the web was something and that it might get Big, but few of us really understood what "Big" meant. Today, we aren't even scratching the surface of the Big Bata opportunity.'

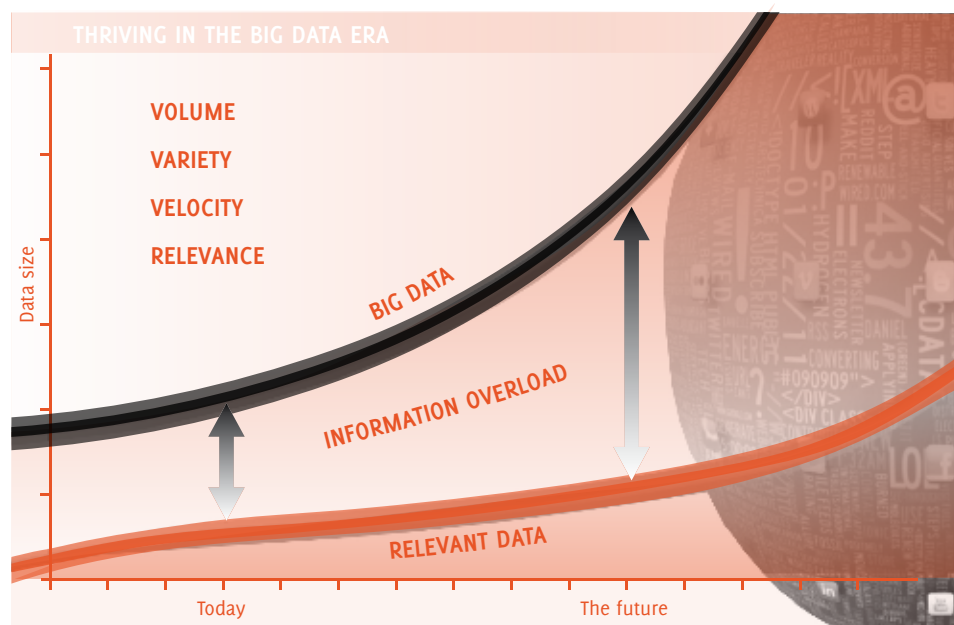
Geen afzonderlijk fenomeen

Als er één ding duidelijk is, dan is het wel dat Big Data geen afzonderlijk fenomeen is. Bovendien benadrukt 'big' met name het kwantitatieve aspect. Gelukkig roept dit onmiddellijk de nodige vragen op, waardoor we genoodzaakt zijn om verder na te denken over Big Data.

In maart 2012 publiceerde Credit Suisse Equity Research het rapport *The Apps Revolution Manifesto, Volume 1: The Technologies*. De auteurs daarvan zien met name de convergentie van Service-Oriented Architecture, Cloud, Fast Data, Big Data, Social en Mobile als bepalend voor de waarde die nieuwe enterpriseapplicaties zullen bieden. Deze ontwikkeling schat Credit Suisse Equity Research net zo transformatief in als client/server en webapplicaties in het verleden waren.

Volume, Variety, Velocity

Al in 2001 maakte Doug Laney duidelijk – toen bij META Group en tegenwoordig bij Gartner – dat er bij de gestage datagroei drie factoren op elkaar kunnen inwerken: de datahoeveelheid (*Volume*), gestructureerde, semigestructureerde en ongestructureerde datatypen (*Variety*) en de gewenste analysesnelheid (*Velocity*). Tegenwoordig voegen we daar vaak *Complexity*, *Value* en *Relevance* aan toe. De laatste twee, omdat we graag precies willen weten wat we met alle data willen en kunnen, om er niet voor niets tijd, geld en moeite in te steken.

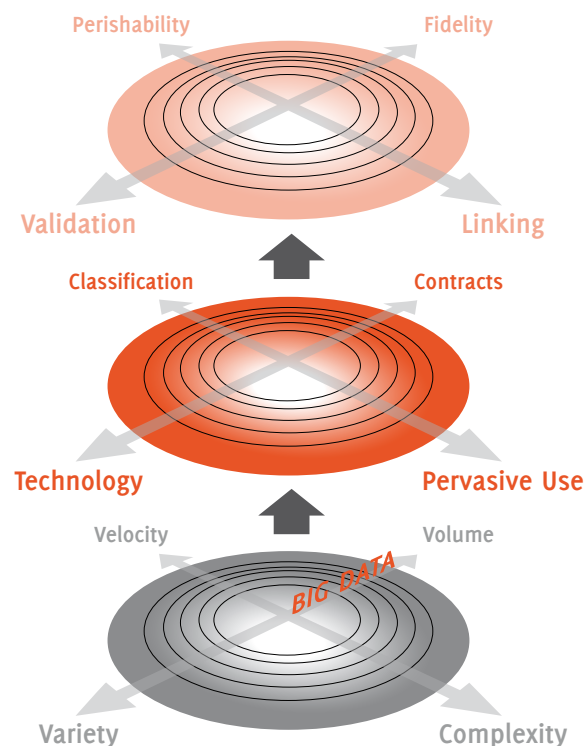


Big Data als de Next Frontier

Vandaaruit, zo voorspelt het McKinsey Global Institute in zijn rapport *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, kan de juiste uitnutting van Big Data honderden miljarden dollars profijt opleveren voor de verschillende sectoren van de Amerikaanse economie. McKinsey onderstreept het grote sectorale verschil (zie paragraaf 11) van het gemak waarmee Big Data kan worden verworven, afgezet tegen de waarde die Big Data-benutting naar verwachting kan hebben, en benadrukt de noodzaak om de kennisachterstand in organisaties inzake de omgang met (Big) Data weg te werken (zie paragraaf 10).

6 Big Data is Extreme Information Management

Het basismodel van Volume, Variety en Velocity heeft Gartner nu uitgewerkt tot de drie interacterende lagen met telkens vier dimensies hierna. De resulterende twaalf factoren grijpen ineen en moeten allemaal doelbewust worden geadresseerd in het informatiemanagement van de 21ste eeuw: afzonderlijk en als geheel.



Simpelweg staat er van beneden naar boven het volgende. Vanuit met name de variëteit en de complexiteit van een toenemende hoeveelheid data – vaak ook realtime – is het op basis van de juiste technologische toepassingen in combinatie met de intensieve inzet van alle data goed mogelijk om gevalideerde uitspraken te doen en verbanden te leggen die businessbeslissingen kwalitatief op een hoger plan brengen.

Als we Big Data als vertrekpunt nemen, dan hoort dat naar de letter thuis aan de volumekant. Variëteit en snelheid zijn op dat niveau de andere dimensies van Doug Laney. Een extra toevoeging is de complexiteit van niet alleen de data maar ook van de ‘use cases’: de manier waarop alle data met elkaar in verband worden gebracht vanuit relevante en waardevolle vraagstellingen. Een concrete typologie op basis van de ‘formule’ *Big Data = Transactions + Interactions + Observations* gaven we al in paragraaf 1.

Op het middelste niveau gaat het om toegang en controle. Om te beginnen zijn er altijd afspraken (*contracts*) over hoe welke (*classification*) informatie moet worden vastgelegd en hoe die kan worden gebruikt. Sociale media en cloud computing bieden prachtige kansen maar er is wel nieuwe technologie (*technology*) nodig om te zorgen dat er altijd en overal gebruik van kan worden gemaakt (*pervasive use*).

De bovenste laag gaat over betrouwbaarheid van informatie (*validation, fidelity*). Ze moet niet alleen relevant en accuraat zijn bij de verwerving (*perishability*), maar ook in de use case: in het gebruik. Van belang is tevens de vraag of er in combinatie met andere informatie (*linking*) verrijking plaatsvindt.

In zijn totaliteit moeten organisaties in Big Data-verband antwoord geven op de bekende zes standaardvragen: wat, wanneer, waarom, waar, wie en hoe? De eerste vier betreffen de inrichting van uw Enterprise Information Architecture en de laatste twee die van uw Enterprise Information Management.

Wat?	Wat zijn de juiste data en informatie?
Wanneer?	Wat zijn de juiste lifecyclefasen daarvan?
Waarom?	Wat zijn de juiste kenmerken?
Waar?	Wat zijn de juiste interfaces voor interactie?
Wie?	Wat zijn de juiste rollen in de organisatie?
Hoe?	Wat zijn de juiste informatieactiviteiten?

Dit is kort en goed de concretisering die bij de standaardvragen hoort. Ze dienen als richtsnoer voor de verdere inrichting van Big Data-, Total Data- of Extreme Information Management-processen.



Vraag 5
In hoeverre is Big Data
een oplossing op zoek
naar een probleem?

www.sogeti.com/vint/r1q5

EIM en Big Data Governance

Concrete handvatten voor Extreme Information Management geeft het Big Data Governance Maturity Framework van IBM. De bijbehorende checklist bevat ruim 90 aandachtspunten op 11 deelgebieden. Dit verhelderende materiaal treft u hier aan:

ibmdatamag.com/2012/04/big-data-governance-a-framework-to-assess-maturity

7 Big Data in organisaties anno 2012

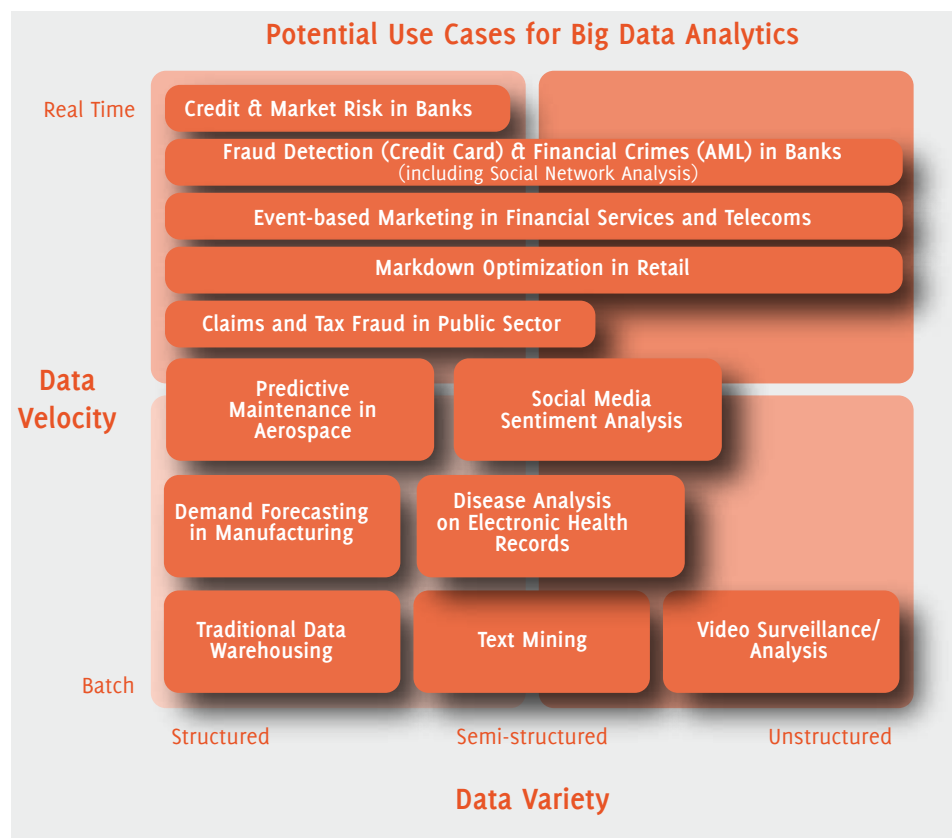
Langs de assen snelheid (Velocity) en datatypen (Variety) – dus bewust abstraherend van datahoeveelheid (Volume) – hebben SAS en IDC de op de volgende pagina afgebeelde voor de hand liggende potentie van Big Data Analytics voor organisaties anno 2012 opgesteld.

Data Science als sport

Het gewenste intensieve samenspel tussen medewerkers op het gebied van Big Data en het huidige gebrek aan expertise en ervaring in organisaties geven ruimte aan de Web 2.0-aanpak die crowdsourcing heet. Een voorbeeld van zo'n online-initiatief in Big Data-dienstverlening is het Australische Kaggle.com, dat een sport maakt van Big Data-uitdagingen: *'We're making data science a sport.'* In hun online arena, zoals Kaggle het noemt, kunnen data-cracks deelnemen aan wedstrijden. Organisaties bieden hun data en vraagstukken aan, die vervolgens door de aan Kaggle verbonden experts vakkundig tot op het bot worden geanalyseerd. De beste oplossing wint en sleept de uitgelopen prijs in de wacht. Roem, prijzengeld en spelplezier is waar de datagladia-toren om strijden:

'Kaggle is an arena where you can match your data science skills against a global cadre of experts in statistics, mathematics, and machine learning. Whether you're a world-class algorithm wizard competing for prize money or a novice looking to learn from the best, here's your chance to jump in and geek out, for fame, fortune, or fun.'

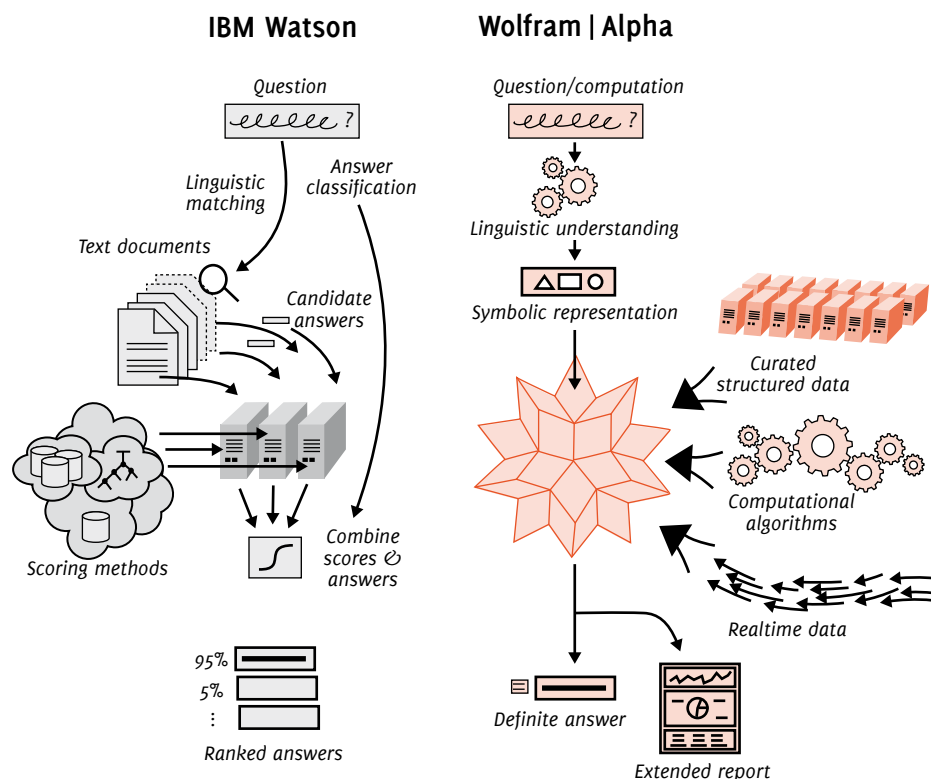
Ontwikkelingen als Kaggle zijn erg interessant, omdat de potentie van innovaties en van innovatief ondernemerschap op basis van Big Data hoog wordt ingeschat. State-of-the-artcomputersystemen als Watson van IBM en Wolfram|Alpha (zie pagina 20) spelen hier ook een grote rol. Deze en andere intelligente computers worden ingezet bij steeds meer Big Data-uitdagingen: van banken tot aan het Smart Grid en de gezondheidszorg.



Het Social Business Analytics-voorbeeld van 'churn', de afkalving van het klantenbestand, dat bijvoorbeeld sterk speelt in de telecom, is aan het begin van deze notitie behandeld, in paragraaf 1.

Het Smart Grid

Op het snijvlak van Big Data en het zogeheten Smart Grid lopen momenteel wereldwijd een groot aantal pilotprojecten. Grid monitoring is een van de aandachtsgebieden, zoals in het Tennessee Valley Authority-project. Daar gaat het nu om 9 miljoen huishoudens en ruim 4 miljard metingen per dag, die samen 500 terabyte aan data opleveren. Typische toepassingen zijn storingen ontdekken en het energieverbruik in de gaten houden. Slimme meters zijn er voor elektriciteit, gas en water. In 2013 zullen er naar verwachting 270 miljoen operationeel zijn. Gaan we een stap verder, naar intelligente huizen, dan zullen die op basis van 100 sensoren per huishouden straks elk 4 tot 20 petabyte per jaar aan data genereren. De behoefte aan Big Data-applicaties in de nutsvoorziening neemt dus toe en de zich ontwikkelende deregulering werkt dit in de hand.



Gezondheidszorg

Gezondheidszorg is een breed veld dat ons allemaal direct aangaat. Wat betreft het klinische gebruik van Big Data, voor zorgbehandeling dus, is het om te beginnen een groot voordeel om informatie door de tijd heen op uiteenlopende manieren te kunnen volgen. Daarnaast kan er een begin worden gemaakt met patroonherkenning, met name de detectie van gebeurtenissen die niet vaak voorkomen of die niet waarneembaar zijn als er gekeken wordt naar kleine populaties. Een mooi voorbeeld is hoe Google met Big Data-analyse van zoektermen in staat is om realtime te volgen hoe een griep epidemie om zich heen grijpt. Nog veel indrukwekkender is hoe het wetenschappelijke Global Viral Forecasting-project Big Data gebruikt om wereldwijde pandemieën als hiv en H1N1 te voorkomen. Daar moeten we agressief proactief in zijn, want het uitblijven van resultaten heeft geleerd dat we niet kunnen blijven afwachten terwijl potentiële rampen zich aan het ontwikkelen zijn.

Onze genenkaart voorbij

Een fundamentele Big Data-ontwikkeling op gezondheidsgebied is de ambitie van het Broad Institute, een initiatief van MIT en Harvard, om het Human Genome Project, dat werd afgerond in 2003, uit te breiden. Over een periode van 13 jaar slaagden wetenschappers er uiteindelijk in om alle 20.000 tot 25.000 genen in kaart te brengen plus de 3 miljard basisparen van het menselijke DNA. Wat het megaproject vooral

uitwees, is dat genen maar een klein deel uitmaken van ons genoom en dat er nog veel fundamentele elementen moeten worden opgezocht en onderzocht.

Daarop legt het Broad Institute zich nu sinds 2003 toe, en dan met name op de vraag hoe cellen precies informatie verwerken, wat niet alleen leidt tot een beter begrip van het genoom, maar ook grote therapeutische waarde heeft. Samen met anderen onderzoekt het Broad Institute momenteel onder meer de celmutaties die kanker veroorzaken, de moleculaire structuur van virussen, bacteriën et cetera die verantwoordelijk zijn voor infectieziekten en de mogelijkheden daarvan voor medicijnontwikkeling.

Genoombiologie en het onderzoek naar celcircuits behoren tot de meest belangrijke Big Data-uitdagingen van dit moment. Eind 2011 had het Broad Institute 8 petabyte aan data. Het instituut bouwt doorlopend aan tientallen specialistische softwaretools om de data op de gewenste manier te kunnen analyseren. Alle software en data kunnen door iedereen worden gedownload.

Social analytics

Warenhuizen gebruiken social analytics om op basis van terabytes aan zoekopdrachten, blogposts en tweets het online-aanbod snel te kunnen aanpassen aan de wensen van klanten. Binnen een paar dagen in plaats van zes weken, zoals vroeger normaal was. Moderne social-analytics-tools zijn geoptimaliseerd voor gebruik door business-professionals en kunnen met allerlei databronnen overweg: publiek toegankelijke bronnen, eigen data en die van partners.

De datastroomrevolutie

Software voor de analyse van datastromen wordt gebruikt om realtime afwijkingen en nieuwe patronen in de data te kunnen ontdekken. Organisaties kunnen op die manier meteen nieuwe inzichten opdoen en snel de beslissingen nemen die op basis van de meest recente gebeurtenissen nodig zijn. Denk aan tweets die gemonitord worden of blogposts, videobeelden, elektrocardiogrammen, gps-data, sensoren van uiteenlopende aard en financiële markten. Moderne datastroomsoftware maakt het mogelijk om realtime complexe verbanden te monitoren in situaties die de mogelijkheden van relationele databases en traditionele analysemethoden te boven gaan. Van patiëntenzorg tot een betere service aan klanten, datastroomsoftware biedt opzienbarend nieuwe mogelijkheden.

Medische complicaties voorkomen

In ziekenhuizen worden doorlopend onder meer de ademhaling, de hartslag, de bloeddruk en de temperatuur van patiënten in de gaten gehouden. Om de vaak subtiele waarschuwingssignalen van complicaties beter te kunnen detecteren moeten datastroomsystemen worden ingezet. Die zijn in staat om ruim voordat de symptomen zich voordoen de eerste indicatoren van complicaties op te pikken. Voorheen werden 1000 metingen per seconde geaggregeerd tot patiëntrapportages per half uur

of om het uur, maar dat is veel te grof. Datastroomsysteem zijn in dit geval van vitaal belang om proactief te kunnen ingrijpen.

Een optimale service

Een ander voorbeeld is de dienstverlening aan klanten. Internet en sociale media hebben klanten empowered en kieskeuriger gemaakt. We vertrouwen elkaars oordeel gemiddeld drie maal meer dan advertenties van organisaties. Het is dus van vitaal belang om goed te luisteren naar wat klanten en anderen online te melden hebben en wat ze onderling uitwisselen. De verbetering van dienstverlening vereist tegenwoordig nauwlettende aandacht voor commentaren op websites, in e-mails, in tekstberichten en op sociale media. Als medewerkers dat handmatig doen, is dat veel te traag en treden er te veel inconsistenties op in de rapportage en de opvolging. Met geavanceerde datastroomsoftware voor contentanalyse zijn organisaties tegenwoordig in staat om dat soort ongestructureerde data automatisch te analyseren en te categoriseren naar bepaalde termen en zinsneden die voorkomen. Het autoverhuurbedrijf Hertz verdubbelde op deze manier de productiviteit van zijn klantenservice.



Vraag 6
Hoeveel privacy bent u bereid op te offeren om een optimale service te krijgen?

www.sogeti.com/vint/r1q6

Visionaire fase

De voorbeelden ten aanzien van Big Data zijn nu nog tamelijk rudimentair. Wellicht is dit een indicatie van de fase waarin de ontwikkelingen rond Big Data zich bevinden. Organisaties baseren hun onderscheidende waarde nu nog niet op hun capaciteit om met Big Data om te gaan. De echte 'helden' van dit tijdperk hebben we nog niet kunnen identificeren, waardoor de disruptieve potentie slechts door de voorbeelden heen schemert. We bevinden ons in het visionaire stadium, waarin volop wordt geëxperimenteerd. Gedurende het Big Data-onderzoek en bij de publicatie van de verschillende onderzoeksnotities zal VINT daarom met name aandacht besteden aan cases op verschillende gebieden, vanuit verschillende invalshoeken en sectoren.

8 Met Big Data van Big Science naar Big Business

In de wereld van de Grote Wetenschap, Big Science, ontwikkelt Big Data zich het hardst. Over 10 jaar zullen 2800 radiotelescopieën in het Square Kilometer Area-project (SKA), het grootste Big Science-project ooit, dagelijks 1 miljard gigabyte aan data genereren. Dat is net zoveel als het hele internet op een doordeweekse dag in 2012. Al in 2008 kondigde Chris Anderson in *Wired* de *Petabyte Age* af en riep Joseph Hellerstein, van UC Berkeley, de *Industrial Revolution of Data* uit. Ter vergelijking: per uur verwerkt Google anno 2012 in totaal 5 petabyte oftewel 5000 terabyte per uur.

Big Data, Big Science en Big Bang

De begrippen Big Data, Big Science en Big Bang hebben alle drie betrekking op een compleet andere situatie dan we gewend waren. Big Bang hebben we te danken aan de Britse astrofysicus Fred Hoyle in een radio-uitzending uit 1949. Atoomwetenschapper Alvin Weinberg populariseerde Big Science in 1961 in het tijdschrift *Science*. En eigenlijk nog maar kort geleden, in 2005, kwam Roger Magoulas van O'Reilly Media op de proppen met de term Big Data. Voor organisaties welteverstaan: van op het individu gerichte *Next Best Offer Analytics* tot aan productieomgevingen en sensordata.

Big Business en Big Bucks

Het is dus een goede gewoonte om iets 'big' te noemen als we daar echt de aandacht op willen vestigen. Denk ook aan *Big Brother* (1949) van George Orwell, en niet te vergeten aan meer profane zaken als Big Business – grote (Amerikaanse) ondernemingen vanaf midden negentiende eeuw – en Big Bucks, die beide met Big Science en Big Data direct verband houden. Wat Big Data betreft stappen we momenteel van megabytes, gigabytes en terabytes door naar het duizelingwekkende tijdperk van de petabytes, de exabytes en de zettabytes. Dat gaat nu heel erg snel.

Overall gonst het van de kansen die voor het grijpen liggen om op Big Data te kapitaliseren. Het McKinsey Global Institute noemde Big Data in 2011 *'the next frontier for innovation, competition, and productivity'* en de Economist Intelligence Unit sprak onomwonden van *'a game-changing asset'*. Het zijn citaten uit de titels van twee richtingwijzende rapporten over Big Data, een actueel thema dat sterk in ontwikkeling is en waarover het laatste woord nog lang niet is gezegd. McKinsey maakt dat heel expliciet:

'This research by no means represents the final word on big data; instead, we see it as a beginning. We fully anticipate that this is a story that will continue to evolve as technologies and techniques using big data develop and data, their uses, and their economic benefits grow (alongside associated challenges and risks).'

Het Global Pulse-project

Als om de relativiserende woorden van McKinsey kracht bij te willen zetten, presenteerde secretaris-generaal Ban Ki Moon van de Verenigde Naties eind 2011 het zogeheten Global Pulse-project. Dat is erop gericht om samen met verschillende commerciële en academische partners via grote online datasets – *New Data* in Global Pulse-terminologie – de vinger aan de pols te houden van een aantal ontwikkelingen in de wereld met als doel om eerder en beter te kunnen ingrijpen. Er zijn vijf hoofdprojecten:



1. A Global Snapshot of Well-being through Mobile Phones
2. Real-Time E-Pricing of Bread
3. Tracking the Food Crisis via Online News
4. Unemployment through the Lens of Social Media
5. Twitter and Perceptions of Crisis-Related Stress

Data Science rules!

Ondanks zo'n richtingwijzend initiatief laat het Big Data-concept zich nog steeds het makkelijkst relateren aan wat we Big Science noemen. Daar zijn de Volume-, Variety- en Velocity-aspecten in combinatie met state-of-the-art hardware en -software het duidelijkst aanwezig, hoewel sommigen de Relevance en Value zullen bestrijden, zeker in crisistijd. Maar bovendien zijn de deeltjesversneller van het CERN en hypermoderne radiotelescopen wel even een paar maten groter dan waar we businesswise mee moeten dealen, en zijn ze datatechnisch van een heel andere orde. Hoe komen we dan met *Big Data* van *Big Science* naar *Big Business*? De kern van het antwoord daarop is *Data Science*, de kunst om bestaande data om te vormen tot nieuwe inzichten waarop een organisatie actie kan/wil ondernemen.

Zonder het tegenwoordig veelbesproken begrip Data Science te noemen benadrukt Chirag Metha, voormalig Technology, Design & Innovation Strategist voor het SAP Office of the CEO, vooral het belang van de tools en de samenwerking daaromheen, want Big Data is absoluut niet alleen voor experts. Het is belangrijk om zoveel mogelijk mensen bij de dataketen te betrekken, aldus Metha:

'Without self-service tools most people will likely be cut off from the data chain even if they have access to data they want to analyze. I cannot overemphasize how important the tools are in the Big Data value chain. They make it an inclusive system where more people can participate in data discovery, exploration, and analysis. Unusual insights rarely come from experts; they invariably come from people who were always fascinated by data but analyzing data was never part of their day-to-day job. Big Data is about enabling these people to participate – all information accessible to all people.'

9 Big Data als nieuw Data Science-tijdperk

Vanaf het begin was een belangrijk kenmerk van Big Science dat de geïsoleerd opererende wetenschapper tot het verleden behoorde. Maar het bleef geen onderscheidend kenmerk voor Big Science, want al gauw werd samenwerking over de hele linie de norm. Zonder goed gecoördineerde samenwerking is moderne wetenschap ondenkbaar. Het rapport *Big Science > Big Data > Big Collaboration: Cancer Research in a Virtual Frontier* uit oktober 2011 benadrukt dat vanuit Big Data-perspectief. Hier wordt Big Science op één lijn gesteld met Big Data en Big Collaboration. In het rapport zelf worden de drie 'Big's' uit de titel nog aangevuld met *Big Technology* oftewel *Big Compute*:

'Big Science generates dimensions of data points and high-resolution images to be deciphered and decoded. In cancer research, Big Data often require on-demand Big Compute across settings using a private cloud, a public cloud or mix of the two.'

Precies dat is ook wat er voor organisaties verandert als ze met Big Data aan de slag gaan. Zijn de bestaande technologieën en werkwijzen in een organisatie niet berekend op Big Data, dan is er een nieuwe aanpak nodig. Dat betekent: investeren in spullen, in mensen, in skills, in processen, in management en in governance. Voor onderzoeksbureau Gartner bijvoorbeeld is Big Data vooral letterlijk de Volume-component aan de basis van wat daar *Extreme Information Management* heet. Integraal onderdeel daarvan is *Data Science*, de 'wetenschap' die met Big Data, Fast Data, Total Data en Dynamic Data onvermijdelijk nu ook de organisatie binnenkomt. Chirag Metha geeft de volgende profielschets van een data scientist:

'The role of a data scientist is not to replace any existing BI people but to complement them. You could expect the data scientists to have the following skills:

- *Deep understanding of data and data sources to explore and discover the patterns at which data is being generated.*
- *Theoretical as well practical (tool) level understanding of advanced statistical algorithms and machine learning.*
- *Strategically connected with the business at all the levels to understand broader as well deeper business challenges and being able to translate them into designing experiments with data.*
- *Design and instrument the environment and applications to generate and gather new data and establish an enterprise-wide data strategy since one of the promises of Big Data is to leave no data behind and not to have any silos.'*



Vraag 7
Kunt u met Big Data
de toekomst beter
voorspellen?

www.sogeti.com/vint/r1q7

Big Data: een nieuwe microscoop

Ruim een eeuw geleden zette Frederick Taylor met zijn *Principles of Scientific Management* de verwetenschappelijking van organisaties op de agenda. Toen ging het om management: belangrijk, maar in essentie een kwestie van 'continuous improvement'. Met Big Data voorzien de enthousiastelingen een daadwerkelijk fundamentele omslag, zoals indertijd de microscoop. Dat is momenteel een geliefde analogie: we staan aan het begin van een nieuw tijdperk, te vergelijken met de start van de moderne wetenschap, zo'n 400 jaar geleden. Dankzij de digitale 'microscoop' die momenteel als het ware voor Big Data wordt ontwikkeld, kunnen we volgens MIT-hoogleraar Erik Brynjolfsson straks op allerlei terreinen veel wetenschappelijker en accurater analyseren en voorspellen. Eindelijk zijn we dankzij geavanceerde hardware en software in staat om razendsnel in en uit te zoomen. Teneinde structuren en verbanden te ontdekken ten behoeve van spectaculair betere inzichten, oplossingen en beslissingen: *Data Driven Decisions* en *Predictive Analysis*.

10 Kennisachterstand inlopen is essentieel

Als actueel businessstema, met torenhoge economische en maatschappelijke beloften, staat Big Data sterk in de belangstelling en is het bovendien volop in beweging. De komende tijd zal dat zo blijven en daarom is er behoefte aan een helder beeld. In dat verband, zo heeft het McKinsey Global Institute becijferd, zullen er ironischerwijze alleen al in de Verenigde Staten 140.000 tot 190.000 data-experts (data scientists) bij moeten komen en moet het aantal businessmensen dat met data overweg kan met 1,5 miljoen omhoog. Om verantwoord met Big Data te kunnen omgaan is om te beginnen een bepaald kennisniveau vereist dat momenteel in organisaties over de hele linie structureel ontbreekt. Volgens onderzoek van IBM uit 2011 willen organisaties wel heel graag, getuige de percentages hierna. Een paar jaar geleden gold nog het excuus dat de ontginning van Big Data alleen was weggelegd voor wetenschappelijke bollebozen en een select aantal organisaties. Voor elke andere partij was het gewoon te moeilijk en te duur. Dat is nu niet meer zo. Pioniers als Walmart, Tesco en Google hebben aangetoond dat data de bron kunnen zijn van bestendig concurrentievoordeel. Op dit moment heeft volgens IBM maar liefst 83 procent van de CIO's visionaire plannen om met nieuwe Business Intelligence & Analytics op basis van Big Data de concurrentiepositie van hun organisatie significant te verbeteren.



De Economist Intelligence Unit onderschrijft dit maar verdeelt tevens het huidige Big Data-gedrag in grote organisaties onder in het volgende volwassenheidskwartet:

- ◆ **Dataverspillers** (*data wasters*)

Van de dataverspillers geeft 30 procent geen prioriteit aan het verzamelen van data. De 70 procent uit deze categorie die wel data verzamelt, gebruikt ze volstrekt te weinig. Zulke organisaties presteren financieel onder de maat. *We treffen ze aan in elke economische sector.*

- ◆ **Dataverzamelaars** (*data collectors*)

Deze organisaties onderkennen het belang van data, maar beschikken niet over de middelen om er iets mee te doen, behalve de data opslaan. Ze hebben zichzelf compleet ondergedompeld in data. *We treffen ze vooral aan in de gezondheidszorg en de professional services.*

- ◆ **Datamanagers in spe** (*aspiring data managers*)

Dit is de grootste groep. Men is zich volledig bewust van het belang van Big Data voor de toekomst van de organisatie. Ze gebruiken data voor strategische besluitvorming en investeren daar hevig in. Maar het topniveau wat betreft prestaties hebben ze nog niet bereikt. *We treffen ze vooral aan in de communicatiehoek en in de retail.*

- ◆ **Strategische datamanagers** (*strategic data managers*)

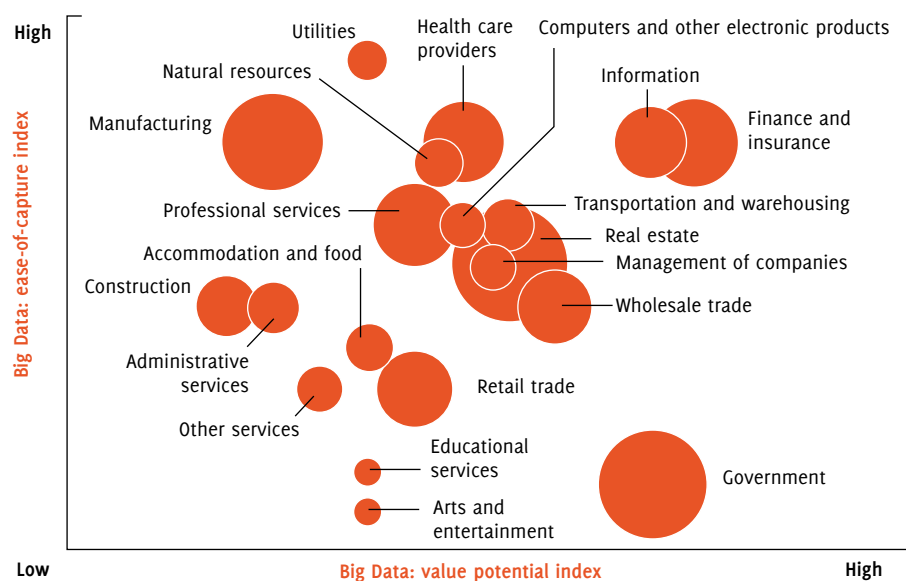
Dit is de meest geavanceerde groep van Big Data-gebruikers. Deze organisaties identificeren allereerst specifieke metriecken en data die geënt zijn op hun strategische doelstellingen. *We treffen ze vooral aan in de maakindustrie, de financiële dienstverlening en de technologiesector.*

Organisaties moeten dus niet alleen zomaar data verzamelen, maar ook de wens en competentie ontwikkelen om met zoveel mogelijk data aan de slag te gaan. Samen met de businessprofessionals moeten data scientists helpen om alle data te interpreteren en inzichten te genereren waar de organisatie daadwerkelijk iets aan heeft. Daarbij kan het gaan om gerichte vraagstukken of om explorerend dataonderzoek. De beweging is om een organisatie om te vormen van intuïtieve naar data-intensieve besluitvorming. Van de *heroïsche manager* die als het ware blind besluiten neemt in het besef dat er veel te weinig data zijn, naar de meer *scientific manager* die eerst op zoek gaat naar data en inzicht.

11 Big Data in klinkende munt

Waarom de Data Science-inhaalslag nodig is, kwantificeert McKinsey als volgt. Wereldwijd kunnen er volgens het bureau op basis van Big Data biljoenen (*trillions*) dollars en euro's aan waarde worden gegenereerd. Jaarlijks bijvoorbeeld 300 miljard dollar in de context van de Amerikaanse gezondheidszorg, 250 miljard euro voor de Europese overheid, ruim 100 miljard dollar voor Amerikaanse telecomproviders en tot 700 miljard voor hun klanten. Door goed te kapitaliseren op Big Data zou de Amerikaanse detailhandel ruim 60 procent meer nettomarge kunnen draaien en zou de maakindustrie uiteindelijk maar de helft van de huidige uitgaven te hoeven spenderen aan productontwikkeling en assemblage, terwijl het werkkapitaal met 7 procent zou afnemen.

Dit zijn voorbeelden uit het totaalplaatje voor sectoren van de Amerikaanse economie op de volgende pagina. Duidelijk blijkt het grote sectorale verschil van het gemak waarmee Big Data kan worden verworven, afgezet tegen de waarde die Big Data-benutting naar verwachting kan hebben. Het McKinsey Center for Business Technology publiceerde de grafiek aan het begin van 2012 in de reader *Perspectives on Digital Business* op basis van gegevens uit het rapport *Big Data: The Next Frontier for Innovation, Competition, and Productivity* van het McKinsey Global Institute uit mei 2011.



Om het gemak van verwerving (verticaal) te bepalen hebben de onderzoekers gekeken naar vier factoren: het aanwezige analytische talent, de IT-intensiteit, de data-driven mindset en de beschikbaarheid van data in een sector. De potentiële waarde (horizontaal) is een functie van de volgende vijf factoren: de hoeveelheid aanwezige data, de variatie in bedrijfseconomische performance, het contact met klanten en toeleveranciers, de transactie-intensiteit en de competitieve turbulentie in een sector. De grootte van de bolletjes in de grafiek geeft de relatieve bijdrage weer van een sector aan het bbp.

Met name waar het om veel mensen gaat, zoals in de nutsvoorziening en de gezondheidszorg, heeft Big Data potentie. Temeer vanwege het relatieve gemak waarmee Big Data kan worden gewonnen, zoals uit de grafiek hierboven blijkt. Wat dat betreft spant de nutsvoorziening de kroon. Qua combinatie van Big Data-verwervingsgemak, klantrelevantie, financieel gewin en bijdrage aan de economie staat de informatieverwerkende industrie, waaronder financiële dienstverleners, aan de top.

Samenvatting en Social Business Analytics

Big Data anno 2012 is vergelijkbaar met wat het web begin jaren negentig was. Er is een enorme versnelling gaande, alles wordt aan elkaar gelinkt en de bijbehorende visies worden gevormd. De verwachting van velen is dat de huidige datafocus de wereld op zijn kop zal zetten: economisch, maatschappelijk, qua innovatie en sociaal.

Organisaties staan momenteel voor de grote uitdaging om zich een voorstelling te maken van de concrete mogelijkheden van Big Data. Hoe zou Big Data een revolutie in uw bedrijfstak kunnen bewerkstelligen? Of wat zou er veranderen als u alles wat u zou willen weten ook inderdaad zou kunnen weten? Kunt u daarmee omgaan? Wilt u dat eigenlijk wel, en zo ja, hoe dan? En kunt u het zich veroorloven om de Big Data-ontwikkeling nog even af te wachten of er misschien maar helemaal niet aan mee te doen?

De kern van Big Data is dat we te maken hebben met één dataspectrum, één continuüm. Dat zullen organisaties stap voor stap strategisch gaan exploreren, want nieuwe mogelijkheden om beter beslissingen te kunnen nemen, willen we niet laten liggen. Om de urgentie voor uw organisatie te helpen bepalen hebben we in paragraaf 3 de volgende drie vraagstukken gepresenteerd en toegelicht:

- A. Uw Big Data-profiel: hoe ziet dat eruit?
- B. Tien Big Data-managementuitdagingen: wat zijn uw issues?
- C. Vijf vereisten voor uw Big Data-project: bent u er klaar voor?

De interactie over deze en aanverwante zaken vindt plaats op onze website, maar zeker ook tête-à-tête wat ons betreft. Wekelijks zullen wij nieuwe onderzoeksinzichten met u delen via blogposts, e-mail- en twitter-alerts. Het begeleidende videomateriaal met toonaangevende deskundigen is bedoeld als inspiratie om de hele Big Data-thematiek vanuit verschillende invalshoeken verder te doordenken en bediscussiëren.

Lang niet alle antwoorden zullen onmiddellijk kunnen worden gegeven, sterker: er zullen juist veel nieuwe vragen bij u opkomen. Het Big Data-thema is een zoektocht met veel vraagtekens aan het begin van en zeker ook gedurende de reis. Daarom wisselen we graag met u van gedachten: online op www.sogeti.com/vint/bigdata/questions en natuurlijk in persoonlijke gesprekken.

Door actief deel te nemen aan de discussie helpt u uzelf en ons om de gedachten ten aanzien van Big Data aan te scherpen. Om door voortschrijdend inzicht te komen tot heldere en verantwoorde beslissingen. Samen bepalen we zo de concrete invulling van de komende drie onderzoeksnotities na deze kick-off over Big Data.

In veel organisaties staat momenteel de uitdaging centraal om het relevante klantgedrag en de consequenties daarvan op een zo rijk mogelijke manier in kaart te brengen en vandaaruit te sturen. Dit is de kern van *Social Business Analytics*, het hoofdthema van de tweede Big Data-onderzoeksnotitie van VINT uit deze reeks van in totaal vier.

Literatuur en illustraties

- Anderson, C. (2008): 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete'
- Appro Supercomputer Solutions (2012): 'From Sensors to Supercomputers (Part 1)'
- Appro Supercomputer Solutions (2012): 'From Sensors to Supercomputers (Part 2)'
- Credit Suisse Equity Research (2012): *The Apps Revolution Manifesto. Volume 1: The Technologies*
- Economist Intelligence Unit/SAS (2011): *Big Data: Harnessing a Game-changing Asset*
- Frost & Sullivan (2011): *Big Science > Big Data > Big Collaboration – Cancer Research in a Virtual Frontier*
- Gartner (2012): *Information Management Goes 'Extreme': The Biggest Challenges for 21st Century CIOs*
- Harbor Research (2012): 'Smart Systems Drive New Innovation Modes'
- Hortonworks (2012): '7 Key Drivers for the Big Data Market'
- IBM (2011): *Big Data Success Stories*
- IBM Data Management (2012): 'Big Data Governance: A Framework to Assess Maturity'
- IDC/SAS (2011): *Big Data analytics: Future architectures, Skills and roadmaps for the CIO*
- Leadership Council for Information Advantage/EMC (2011): *Big Data: Big Opportunities to Create Business Value*
- McKinsey Global Institute (2011): *Big Data: The Next Frontier for Innovation, Competition, and Productivity*
- Mehta, C. (2012): '4 Big Data Myths – Part II'
- MIT Sloan Management Review/IBM Institute for Business Value (2010): *Analytics: The New Path to Value*
- Sumser, J. (2012): 'Big Data: The Questions Matter Most'
- The 451 Group (2010): 'Total data: "bigger" than big data'
- UN Secretary-General (2011): *Global Pulse*
- Wolfram, S. (2011): 'Jeopardy, IBM, and Wolfram|Alpha'
- World Economic Forum (2012): *Big Data, Big Impact: New Possibilities for International Development*
- Yared, P. (2012): 'Big Data may be hot, but "little data" is what makes it useful'

Helderheid creëren met Big Data

www.sogeti.com/vint/r1q1

Vraag 1

Winnen feiten het definitief van intuïtie?

www.sogeti.com/vint/r1q2

Vraag 2

Hoe koppelt u realtime Big Data aan de operationele sturing van uw bedrijf?

www.sogeti.com/vint/r1q3

Vraag 3

Wat is de beste aanpak om voor Big Data een gewillig oor bij het management te vinden?

www.sogeti.com/vint/r1q4

Vraag 4

Wat is de belangrijkste nieuwe spelregel voor organisaties ten aanzien van Big Data?

www.sogeti.com/vint/r1q5

Vraag 5

In hoeverre is Big Data een oplossing op zoek naar een probleem?

www.sogeti.com/vint/r1q6

Vraag 6

Hoeveel privacy bent u bereid op te offeren om een optimale service te krijgen?

www.sogeti.com/vint/r1q7

Vraag 7

Kunt u met Big Data de toekomst beter voorspellen?

**Doe mee aan onze
Big Data-discussie op
[www.sogeti.com/vint/
bigdata/questions](http://www.sogeti.com/vint/bigdata/questions)**



SOGETI

VINT | Vision • Inspiration • Navigation • Trends

Over VINT

Alle ontwikkelingen volgen op IT-gebied is voor veel organisaties een zware opgave. Vaak staan nieuwe IT-mogelijkheden immers ver af van het primaire bedrijfsproces. Bronnen die deze ontwikkelingen inzichtelijk en pragmatisch benaderen, door ook het mogelijke gebruik te belichten, zijn dun gezaaid. VINT geeft invulling aan die koppeling tussen bedrijfsprocessen en nieuwe IT.

In elke rapportage over een verkenning die het instituut heeft uitgevoerd, zoekt VINT het juiste midden tussen feitelijke beschrijving en beoogde toepassing. Op die manier inspireert VINT organisaties om nieuwe technologie in beschouwing te nemen of zelfs te gaan gebruiken.

vint.sogeti.com

vint@sogeti.nl