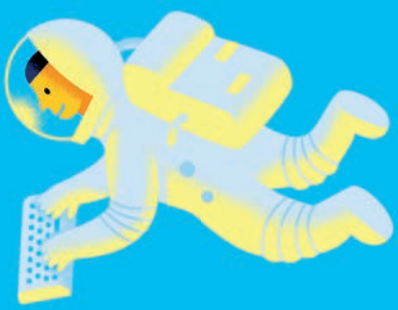
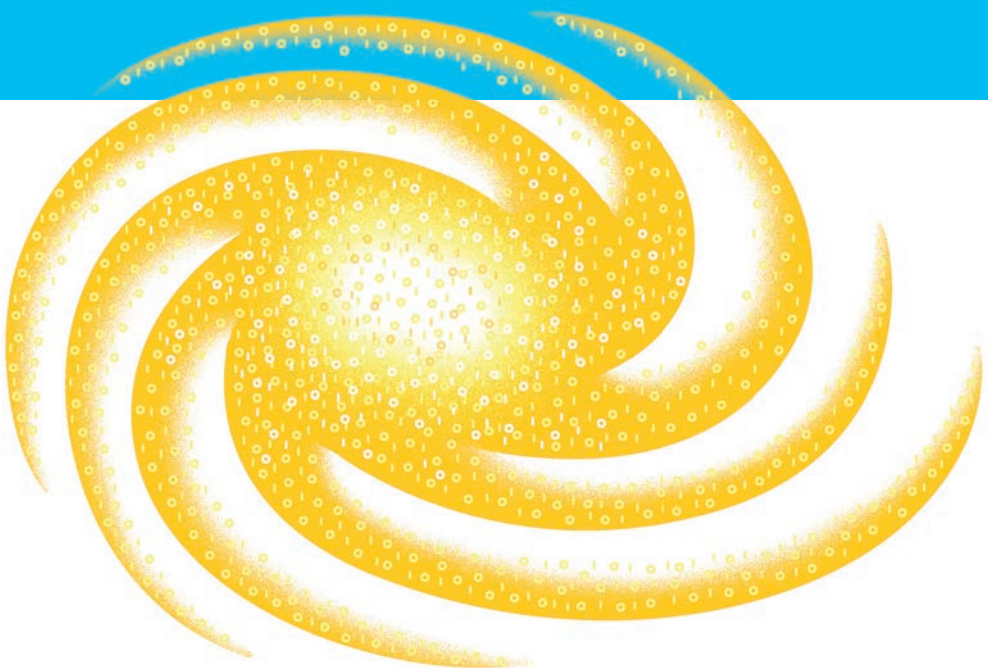
Big Data

Tom Breur





AutomatiseringGids

ag

ACADEMIC SERVICE

Big Data – De nieuwe goudkoorts?

Voor u ligt alweer de tweede uitgave in de AG-reeks, een serie boeken over actuele ICT- en business-thema's die door Academic Service in samenspraak met de redactie van de AutomatiseringGids wordt samengesteld. De boeken in deze reeks worden geschreven door deskundige auteurs, die allen expert op een vakgebied of thema zijn. Het is de bedoeling dat we circa vier boeken per jaar uitgeven, en na *Business Logic Management* (dat vorige maand verscheen) is er nu dan *Big Data*.

De onderwerpen mogen dan verschillend zijn, de opzet is gelijk. Elk boek bevat achtergronden, visie, een blik op trends en toekomst, maar is daarnaast ook praktisch en helder geschreven, en we bieden u met behulp van onze checklisten en stappenplannen concrete handreikingen om met het geleerde aan de slag te gaan. De praktische en lokale voorbeelden maken de behandelde onderwerpen inzichtelijk en vergroten de herkenbaarheid en toepasbaarheid voor u.

We zien het als onze 'missie' om u goed te informeren, verbinden en inspireren met betrouwbare, maatgevende en praktische informatie. We hopen dat we daar, ook middels dit boek van Tom Breur, in slagen en wensen u veel leesplezier, en succes!

Arjan Kors, hoofdredacteur AutomatiseringGids

Marcel Roozeboom, uitgever Academic Service

Big Data

De nieuwe goudkoorts?

Tom Breur

ACADEMIC SERVICE

Meer informatie over deze en andere uitgaven kunt u verkrijgen bij:
Sdu Klantenservice
Postbus 20014
2500 EA Den Haag
tel.: (070) 378 98 80
www.sdu.nl/service

Academic Service is een imprint van Sdu Uitgevers bv.

Omslagontwerp: Studio Polkadot
Zetwerk: Redactiebureau Ron Heijer, Markelo
Druk- en bindwerk: Drukkerij Wilco, Amersfoort

ISBN 978 90 125 8567 5
NUR 980

Inhoud

1 Inleiding

We leven in een tijdperk met een ongekende groei in datavolumes. Momenteel wordt het totaal geproduceerde volume aan data in het afgelopen jaar geschat op zo'n twee zettabytes. Om een indruk te krijgen van dit enorme volume: de megabytes en de gigabytes kennen we inmiddels goed genoeg. Vervolgens heb je de terabytes, de petabytes, de exabytes en dan pas de zettabytes. Volgt u het nog? Een twee met 21 nullen: 2 000 000 000 000 000 000 000.

De jaarlijkse groei in datavolumes bedraagt zo'n 60% (bron: onderzoeksbureau IDC). Dat komt neer op grofweg een vertienvoudiging elke vijf jaar. En het eind is nog lang niet in zicht. Er is geen enkel signaal dat suggereert dat die groei de komende tijd minder zal worden, of zelfs zal afvlakken. Goed nieuws voor de verkopers van datastorage.

Uiteenlopende managementboeken zoals *Competing on Analytics* (Davenport & Harris, 2007), *Super Crunchers* (Ayres, 2007) en bijvoorbeeld *Data Driven* (Redman, 2008) hebben de toon gezet bij een breed lezerspubliek. In een vorig jaar gepubliceerd rapport, bestempelt de McKinsey Global Institute 'Big Data' als 'the next frontier for innovation, competition and productivity'. De strategische waarde van data is bij senior management onder de aandacht gekomen.

Concurrentie speelt zich niet meer af op nationaal of regionaal, maar steeds vaker op globaal niveau. Hierdoor neemt concurrentiedruk en efficiëntie van markten toe. Dit zorgt er dan weer voor dat schaalgrootte en 'slim' uitnutten van grote hoeveelheden gegevens een cruciale rol speelt in het verwerven en behouden van *duurzaam* concurrentievoordeel. In *The World is Flat* (Friedman, 2005) doet de auteur uit de doeken hoe deze trend zich in de komende jaren alleen nog maar verder zal uitbreiden.

'Big Data' is een hype die ook tot de bestuurskamer is doorgedrongen. Succesverhalen van organisaties die 'business analytics' een centrale rol in hun bedrijfsstrategie toekennen. Neem Google bijvoorbeeld. Waar hebben zij hun succes aan te danken? Een superieur zoekalgoritme? Google AdWords? Of zou het komen omdat zij de slimste computerexperts aan zich weten te

binden, waarmee ze beter en sneller dan hun rivalen (Yahoo, Bing) hun enorme clusters van computers aan het werk zetten?

Dichter bij huis hebben we in Nederland een partij als Cool Blue gezien. Een internetshop die heel klantgericht websites inricht, hoofdzakelijk voor de verkoop van elektrische apparatuur. In sneltreinvaart heeft Cool Blue de markt veroverd door slim in te spelen op klantbehoeften, waarbij ze zich in hoge mate laten leiden door data. Eerst marktonderzoek om te bepalen in welke markt(en) ze willen concurreren, en vervolgens elke muisklik van bezoekers registreren om te leren wat mensen willen, waar ze voor willen betalen, en minstens zo belangrijk: hoeveel. Bol.com is ook al zo'n voorbeeld van een internet-retailer die heel snel inspeelt op veranderende klantbehoeften.

De wens vanuit de business om over steeds meer, en steeds rijkere (lees: ook *ongestructureerde*) data te beschikken, staat op gespannen voet met de traditionele machinerie waarmee we business intelligence tot dusver hebben uitgerust. Er is behoefte aan meer flexibele, beter schaalbare architectuur. Traditionele relationele databasemanagementsystemen (RDBMS-en) blijken in dat opzicht een aantal beperkingen te hebben die met name bij (zeer) grote datavolumes en ongestructureerde gegevens voor problemen zorgen.

Het lijkt nauwelijks toeval dat ontwikkelingen binnen Google aan de basis staan van een belangrijk deel van de huidige Big Data-innovaties. Nieuwe Big Data-technieken zoals MapReduce en GFS (Google File System) zijn bij hen ontwikkeld (Open Source!) om meer resultaten te halen uit de bergen van data waar zij over beschikken. Het is op zijn minst opmerkelijk dat *vrijwel alle* 'Big Data'-oplossingen onder een Open Source-model tot stand zijn gekomen. Als software zelf niet je kerncompetentie is, kiezen steeds meer organisaties ervoor om gebruik te maken van het Open Source-model van ontwikkeling.

De business intelligence-markt (BI-markt) werd de afgelopen decennia gedomineerd door leveranciers van relationele databasemanagementsystemen. Die hebben hun werk steeds prima gedaan, en dat doen ze nog steeds. Maar zij hebben ook een belangrijk nadeel: de mogelijkheid om het relationele model op te schalen over (zeer) grote clusters van servers, is uiterst beperkt. Wat je wilt is een architectuur die (bijna) lineair kan meegroeien met toenemende volumes van data. Het relationele model is hier niet geschikt voor, de kosten voor opschalen nemen op een gegeven moment exponentieel toe, waardoor er een praktisch plafond is aan volumes die je met dergelijke systemen bedrijfseconomisch verantwoord kunt beheren.

Dat fenomeen had tot gevolg dat businesscases voor toepassingen met zeer grote hoeveelheden aan (met name ongestructureerde) data lastig te maken zijn. De hardware blijkt dan al snel eenvoudigweg te duur. En net zoals het water achter een dam op zoek gaat naar nieuwe stromen, ontstonden langzaam maar zeker alternatieve oplossingen voor bestaande schaalbaarheidsproblemen.

Door de informatie-explosie op het internet, en vergaande 'digitalisering' van onze maatschappij, produceren we onvoorstelbare hoeveelheden data. Maar data is nog geen informatie. Het 'opwerken' van data tot actiegerichte inzichten in de 21e eeuw is het equivalent van de industriële revolutie in 19e eeuw. Tussen 1850 en 1900 draaide alles om toegang tot grondstoffen en materialen, en de (logistieke) capaciteit om deze snel aan- en af te voeren. In de 21e eeuw wordt de concurrentieslag gestreden door steeds meer en betere data te benutten, en dus moet opslag en verwerking zo economisch mogelijk gebeuren.

'Traditionele' RDBMS-en (gebaseerd op SQL) leggen het steeds vaker af tegen nieuwe(re) technologie zoals Apache Hadoop en NoSQL-databases. Met name als de volumes erg groot zijn, en als er een mix verwerkt moet worden van gestructureerde en ongestructureerde gegevens. SQL (Structured Query Language) was in BI met afstand de belangrijkste programmeertaal van de afgelopen decennia. Geen wonder, want we werkten vrijwel alleen met gestructureerde data. NoSQL-oplossingen kennen ook nog wel hun beperkingen, maar daarover later meer.

Een kenmerk van de recente groei in data is dat deze van voornamelijk gestructureerd is uitgebreid naar goeddeels ongestructureerde of semigestructureerde informatie. Voorheen waren data vooral CRM- en transactiegeoriënteerd: de dialoog tussen klant en bedrijf genereerde het leeuwendeel van de data. In het nieuwe, Big Data-tijdperk, zijn het niet langer alleen mensen, maar vooral *machines* onderling die verantwoordelijk zijn voor de explosieve groei in data.

Denk aan in- en uitchecken met je ov-chipkaart. Afleiden van reislengte, bijwerken van saldo, berekenen van tijd sinds overstap, nieuwe routeprijs berekenen, etc. Al deze bewerkingen vinden achter de schermen plaats tijdens een ogenschijnlijk 'eenvoudige' ov-reis. Of denk aan de verbinding tussen zendmast en mobiele telefoon. Meerdere keren per seconde peilt de telefoon de aanwezigheid van signalen, om voortdurend aan te sluiten op de sterkste zender. Of de massale invoering van RFID-chips. Etc. Al deze

processen onttrekken zich grotendeels aan onze waarneming, maar zij genereren achter de schermen wel enorme hoeveelheden data.

1.1 De digitale samenleving

Computers zijn niet meer weg te denken uit ons dagelijks leven. Steeds meer processen worden ondersteund door computers en/of digitale technologie. En al die elektronica laat een spoor van data na. Van een hotelkamer die zich met een keycard laat bedienen, tot en met de handscanner in uw plaatselijke supermarkt. De lussen in de weg die verkeer registreren, uw chipkaart die reisbewegingen vastlegt, we leven zo langzamerhand in een elektronische jungle.

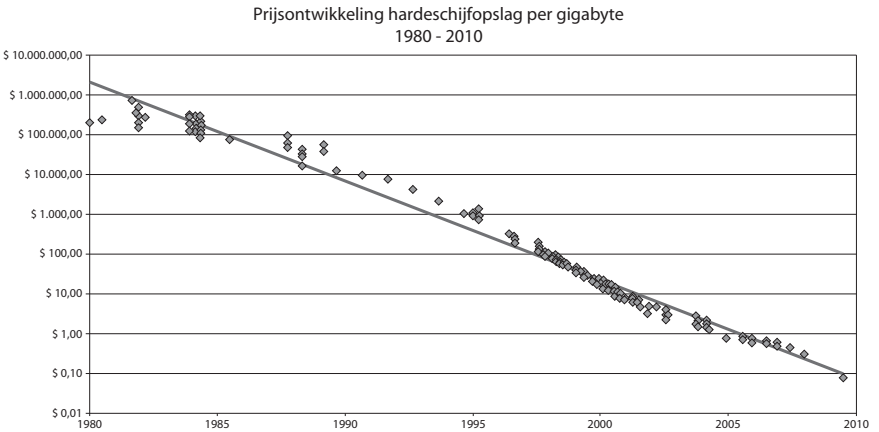
En dan is er het internet, waar elke muisklik wordt vastgelegd. Cloud-oplossingen en Netbooks illustreren dat beschikbaarheid van internetverbinding steeds algemener is geworden. Telecomleveranciers denken op dit moment al na over 4G-oplossingen die de traditionele kabel- en koperverbindingen kunnen gaan vervangen. We stevenen op korte termijn af op een toekomst met alomtegenwoordig draadloos internet, een toekomst die nog meer innovaties mogelijk zal maken, en *nog meer* data zal produceren.

Merk op dat een belangrijk deel van deze groei in datavolumes niet meer door mensen, maar door machines (computers) wordt gegenereerd. Daar vindt de meest spectaculaire groei in volume plaats, samen met audio-, en vooral videotechnologie. Op steeds meer plaatsen worden (bewakings) camera's geplaatst. Registreren van video is een ding, maar soms wil je die beelden later nog analyseren.

De laatste tijd zijn een paar nare gevallen van openbaar geweld en mishandeling (in de nachtelijke uren…) op internet te zien geweest. Op 4 januari 2013 werd iemand in het centrum van Eindhoven, ogenschijnlijk zonder enige aanleiding, vreselijk in elkaar geschopt en geslagen. Camerabeelden circuleerden al snel op internet, hevige verontwaardiging in de (sociale) media, en al snel leidde dit tot honderden tips en de identificatie en aanhouding van de geweldplegers (sommige verdachten woonden in België, en konden niet onmiddellijk worden gearresteerd).

Op een vergelijkbare manier gebruikt de politie camerabeelden bij geweldpleging en overvallen, en vaak met succes. Alle camera's die her en der zijn bevestigd, en 24 x 7 opnames maken, leiden tot een enorm volume aan opgeslagen gegevens. Analyse van al die beelden is (nog steeds) een deels handmatige activiteit, maar ook hier zijn al algoritmes ontwikkeld die personen en gezichten in beelden kunnen onderscheiden.

Als je bedenkt dat het project Google Glass zijn pilotfase in gaat, en je bedenkt hoeveel data al die Google-brillen zullen verzamelen, dan begrijp je waarom datavolumes exponentieel blijven groeien. Dit is mede mogelijk geworden door de scherpe daling in prijzen voor gegevensopslag (zie figuur 1.1).



Bron: Matt Komorowski, 2009

Figuur 1.1 Ontwikkeling in opslagkosten per gigabyte

Door deze spectaculaire prijsdalingen, is het steeds gemakkelijker om te besluiten gegevens te bewaren 'voor het geval dat...' En doordat steeds meer data, en steeds meer historie beschikbaar komt, worden er nieuwe *secundaire* toepassingen voor al die gegevens ontdekt. Een vicieuze cirkel. Hoe meer gegevens we opslaan, hoe meer toepassingen we gaan ontdekken voor die data.

Behalve de prijs van opslag, is ook de rekenkracht van computers in de afgelopen decennia spectaculair toegenomen. Toen wetenschappers in de jaren zestig de Cray-supercomputer ontwikkelden, deed men een profetische prognose: op termijn zou men met dertien van deze machines de gehele Verenigde Staten van alle ooit benodigde rekenkracht kunnen voorzien. Tja, vandaag de dag heeft een gemiddelde spelcomputer aanzienlijk meer rekenkracht dan die Cray destijds...

Er zijn heel veel manieren om rekenkracht van computers te meten, maar bij wijze van proxy kijken we even naar de prijs van transistors, de elementaire bouwstenen van (digitale) computers. Sinds de Tweede Wereldoorlog zijn alle computers digitaal, dus daarmee hebben we 'een' meetlat waarmee we doorheen de tijd een vergelijking kunnen maken.

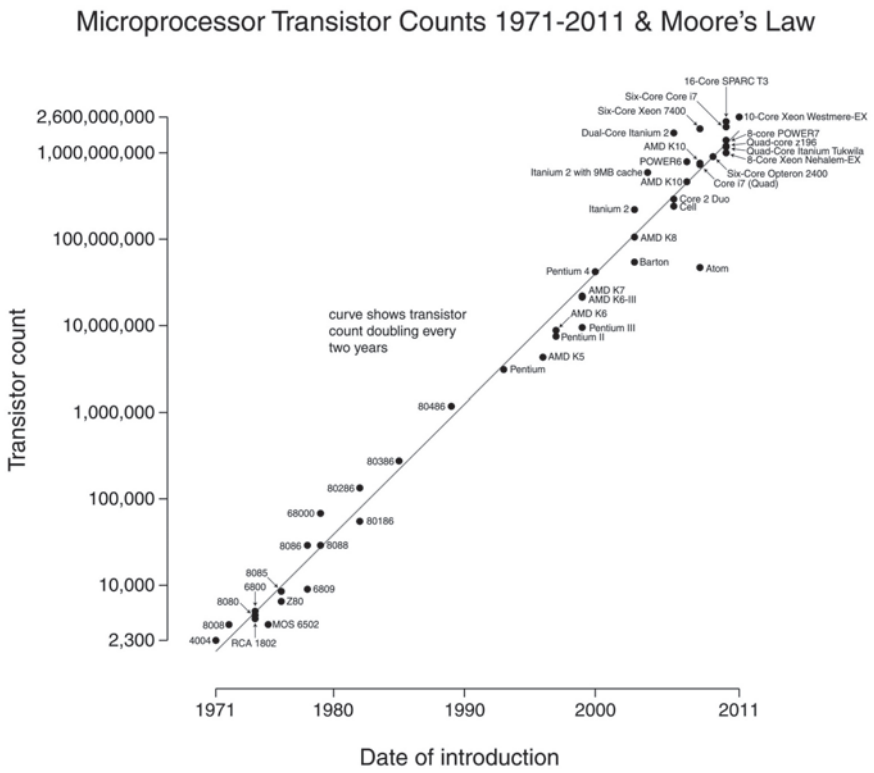
In dit verband wordt vaak melding gemaakt van de Wet van Moore (Moore's Law). Deze is vernoemd naar een van de (mede)oprichters van Intel, Gordon E. Moore, en werd in 1965 (!) beschreven. Destijds werd al duidelijk dat het aantal transistors dat in ic's (integrated circuits, tegenwoordig noemen we dat chips) werd ingebouwd, grofweg elke achttien maanden verdubbelde.

In de praktijk (in de natuur) zijn er weinig processen die lange(re) tijd exponentieel (blijven) groeien. Want processen die exponentieel groeien, worden meestal door een of andere tegenkracht weer afgeremd. En de werking van die tegenkracht doet de groei op termijn weer afvlakken. Als er lelies in een vijver woekeren, elke dag verdubbelen, en ze bedekken de hele vijver na dertig dagen, hoe lang duurt het dan voordat ze de vijver voor de helft bedekken? Inderdaad, 29 dagen. Veel mensen zouden intuïtief een lager getal noemen (probeer het maar!), omdat het menselijk brein dit soort exponentiële groei maar moeilijk kan bevatten.

In de natuur zou die groei bijvoorbeeld kunnen afvlakken doordat er een schaarste aan voedsel of zonlicht optreedt. Vergelijkbare tegenkrachten remmen ook meestal technologische groei (op den duur) weer af. Nochtans blijkt de Wet van Moore al bijna een halve eeuw op te gaan. En deze spectaculaire groei heeft voor computers, de snelheid van sensors, resolutie van digitale camera's, en nog veel meer aspecten van onze hedendaagse technologie een belangrijke innovatieve rol gespeeld.

Als direct gevolg van de Wet van Moore en de duizelingwekkende snelheid waarmee digitale technologie innoveert en verbetert, worden we links en rechts ingehaald door de realiteit. De eerste mobiele telefoons die met camera's werden uitgerust, leken een modegril. Onderzoek destijds toonde ondubbelzinnig aan dat consumenten geen behoefte hadden aan mobiele telefoons die met een camera werden uitgerust. Op basis van de resultaten concludeerden marktonderzoekers dat we nooit (voldoende) voor dergelijke features zouden willen betalen, 'omdat niemand behoefte heeft aan een camera met een dergelijk lage resolutie'. Vandaag de dag hebben (vrijwel) alle telefoons een camera. En de meeste met meer dan vijf megapixels…

Zowel de groei in aantallen transistors, als de daling in prijs voor gegevensopslag, verlopen langs een *logaritmische* schaal. Deze exponentiële groei heeft voor een enorme stuwing in technologische innovatie gezorgd. Zodra die grenzeloze mogelijkheden duidelijk worden, zijn er ondernemers die kansen ruiken om hier op in te spelen. Smartphones, AppStores, gepersonaliseerde on-line aanbevelingen, het gemak dient de mens.



Bron: Wikipedia

Figuur 1.2 Groei van het aantal transistors per ic volgens de Wet van Moore

Vandaag de dag heeft ongeveer een derde van de wereldbevolking toegang tot internet. De groeiende penetratie van breedbandinternet heeft het mogelijk gemaakt om steeds vaker interactieve applicaties aan te bieden, audio en video te 'streamen', waardoor weer een nieuwe vicieuze cirkel in werking treedt om gebruik te maken van de nieuwe mogelijkheden die dit biedt.

In 2012 wist Jason Spero (hoofd van Google Mobile Sales) te melden dat er op dat moment zo'n 1 miljard mensen het internet gebruikten via een mobiel apparaat (smartphone, tabletcomputer, etc.), en ook de software die men hiervoor gebruikt, laat een gedetailleerd digitaal spoor na. Als ik gewoontegetrouw een reisplanner gebruik om trein- en bustijden op te zoeken, zal de reeks aan zoekopdrachten een opmerkelijk nauwkeurig beeld verschaffen van mijn 'whereabouts'.

In een wereld waar computers onze kwaliteit van leven helpen verbeteren, zullen we alsmaar meer data produceren. Deze data zijn nodig om primaire

processen te ondersteunen, maar bieden vooral ten behoeve van secundaire analyse bijna grenzeloze mogelijkheden.

1.2 Data en (predictive) analytics

Data worden opgeslagen omdat ze voor iemand, ergens, waarde vertegenwoordigen. Een elektronische sleutel bleek 'handiger' in hotels, dan de ouderwetse sleutel met baard. Alhoewel de elektronische kaartsleutel op de eerste plaats dienst doet zoals elke andere sleutel, laat hij ook een elektronisch spoor na. Behalve het primaire doel om de deur te openen, kun je door de gegevens over gebruik op te slaan, wellicht ook het een en ander te weten komen over het gebruik van de kamer. En over voorkeuren van de betreffende gast.

Merk op dat er steeds een primair proces is waarvoor data worden verzameld (kamerdeur openen/sluiten), en (mogelijk) een secundair proces dat ook gebruikmaakt van dezelfde (of afgeleide) data ten behoeve van analytische doeleinden (bijvoorbeeld waak-/slaaptijden van de hotelgast bepalen). Het is met name die laatste categorie, secundaire toepassingen van reeds beschikbare data, die zo enorm aan het groeien is, en een sleutelrol speelt in Big Data.

Systemen worden gebouwd met een bepaald gebruiksdoel voor ogen. Maar als het systeem er eenmaal is, zullen creatieve geesten nieuwe mogelijkheden bedenken om de reeds beschikbare data voor alternatieve doeleinden nuttig te maken. Door systemen te koppelen, neemt dat aantal mogelijkheden exponentieel toe. En naar mate er meer toepassingen worden bedacht, zal dit commerciële geesten weer inspireren.

Het onderzoeksbureau IDC heeft becijferd dat er voor het eerst in de geschiedenis meer data *over* ons worden verzameld, dan we zelf genereren gedurende ons leven. En de verwachting is dat die trend alleen maar verder zal doorzetten. Naarmate bedrijven (en overheden) meer toepassingen van die gegevens benutten, neemt die vraag naar afgeleide gegevens alsmaar verder toe.

Onze Rijksoverheid, bijvoorbeeld, maakt bij de planning van onze infrastructuur gebruik van een combinatie van Big Data zoals die voortkomt uit het netwerk van lussen in de weg, en informatie over bijvoorbeeld geplande werken aan de weg. Op die manier probeert men de filedruk binnen de perken te houden, zelfs als er werk aan de weg plaatsvindt waarvoor rijbanen moeten worden gesloten.

Amsterdam beschikt behalve over een rondweg ook over een digitale ring: op camera's worden de nummerborden van alle auto's vastgelegd die vanaf de ring de stad binnenrijden. Dit Automatic Number Plate Registration System (ANPR) wordt gebruikt om toe te zien dat er geen (vervuilende) vrachtwagens de binnenstad in rijden. Maar behalve vrachtwagens, zou je natuurlijk alle geregistreerde voertuigen kunnen volgen. Het is een kwestie van tijd voor daar toepassingen voor worden bedacht.

Behalve het gebruik van navigatieapparatuur (zoals TomTom), wordt er nu ook al nagedacht om gegevens van mobiele telefoons te benutten bij het registreren van files. De dichtheid van mobiele telefoons is erg hoog in Nederland, en elk apparaat heeft een unieke (IMEI-)code, waarmee zendmasten – bij voldoende dichtheid – tot op een of enkele meters de exacte locatie van het toestel kunnen bepalen.

Bij forensisch onderzoek worden nu al telefoonmastgegevens gebruikt om plaatsbepaling te doen. Sinds 1 september 2009 is er een nieuwe Wet bewaarplicht telecommunicatiegegevens in werking getreden, die aanbieders van telefonie en internet verplicht om zes tot twaalf maanden gegevens over gebruik van hun klanten op te slaan. Justitie doet in voorkomende gevallen een beroep op deze gegevens ten behoeve van wetshandhaving en bestrijding van terrorisme.

In al deze gevallen worden gegevens voor een primair proces verzameld (telefoonmasten zenden *primair* om een gesprek tot stand te brengen), en pas daarna worden secundaire toepassingen *afgeleid* van die data. Kennis of inzicht komt tot stand in functie van het 'digitale spoor' dat achterblijft doordat zo veel van ons handelen vandaag de dag door computers wordt geregistreerd.

In de gevallen die we tot dusver hebben beschreven, ging het steeds over het beschrijven van gebeurtenissen in het verleden. Maar behalve rapporteren over wat er *is* gebeurd, is er ook vaak behoefte aan prognoses over wat er *zal gaan* gebeuren. Als je een organisatie louter en alleen bestuurt op basis van historische informatie, is het als het ware alsof je in het verleden leeft.

Nu is het verleden een hele goede en doorgaans betrouwbare raadgever. Als ik wil weten wat voor weer het morgen zal zijn, is een blik uit het raam een hele aardige raadgever. Waarschijnlijk lijkt het weer morgen veel op het weer van vandaag, maar niet altijd.

Organisaties die hun bedrijfsvoering vooral laten afhangen van het verleden, nemen echter een risico. Zij leggen hun toekomst in de waagschaal door de aanname dat de toekomst zal lijken op het verleden. Het is een beetje alsof je rijdt in een auto met een geblindeerde voorruit, waarbij je stuurt door in de achteruitkijkspiegel te kijken. Vaak (lang) gaat het goed, *tot* er een scherpe bocht in de weg opdoemt…

Analistenfirma Gartner is stellig overtuigd van de toekomst van zogenaamde 'predictive analytics'. Zij voorspellen dat in 2020 driekwart van alle BI-gebruikers de beschikking zal hebben over predictieve functionaliteit. Vandaag de dag is dat minder dan een derde. Gartner verwacht dat dit zal groeien naar de helft in de komende twee jaar.

Dit soort predictieve analyse is niet nieuw, verre van dat. Technieken zoals Forecasting, Predictive Modeling, en Optimisation kennen we al jaren. Door de (veel) grotere beschikbaarheid van data zijn er echter veel meer toepassingen van dit soort algoritmes in beeld gekomen. We passen bestaande, bekende technologie toe op nieuw beschikbare data. Weinig nieuws.

Wat wel (tamelijk) nieuw is, zijn toepassingen die dergelijke voorspellingen *inbouwen* in applicaties die primaire processen ondersteunen. Men spreekt in dit verband van 'embedded analytics', algoritmes en rekenmodellen die worden ingebouwd in operationele applicaties (primaire systemen). Voorspellingen zijn dan niet langer het (exclusieve) domein van data-analisten, maar worden daarmee veel breder inzetbaar.

Een andere term die in dit verband wordt gebezigd is 'decisioning': zonder (of met minimale) tussenkost van gespecialiseerde data-analisten geven we predictieve technologie in handen van proceseigenaren. Die eigenaren van het businessproces (zonder al te veel kennis van statistiek of datamanagement) gebruiken deze voorspellingen vervolgens om heel kort-cyclisch betere beslissingen te nemen.

Neem als voorbeeld de bezetting van personeel in een callcenter (contactcenter). Elke dag zijn er drukke en minder drukke momenten. De callcentermanager wil voorkomen dat de wachttijden oplopen en er mensen ophangen terwijl ze nog in de wacht staan. Dat is niet klantvriendelijk. Maar diezelfde manager wordt ook geacht 'oordeelkundig' (en dat betekent meestal zuinig) om te gaan met de inzet van personeel. Je wilt elke dag, de hele dag, goed bereikbaar zijn. En je wilt zo min mogelijk personeel aanwezig hebben, zeker geen personeel dat 'idle' is.

Dat is best een lastig optimalisatieprobleem, maar alle gegevens die nodig zijn om hier een keuze in te maken, zijn in principe bekend, of in ieder geval *kenbaar*. Je weet op elk moment van de dag wat de wachttijd is, en je weet hoeveel medewerkers, en met welke skills (vaak kunnen niet alle medewerkers op alle 'werkstromen' worden ingezet) aanwezig zijn. Op basis van de beoogde serviceniveaus kies je dan een 'optimale' bezetting van je callcenter.

Het optimalisatiemodel dat die berekening uitvoert, hoeft door de callcentermanager niet 'gekend' te worden. Hij moet opgeven wat de doelstelling is in termen van wachttijden, serviceniveaus, etc. De callcentermanager kan ook inschatten in welke mate de bezetting kan worden op- en neergeschaald. Het is gebruikelijk bij dat soort afdelingen dat veel medewerkers een zogenaamd min/max-contract hebben. Zij werken in principe parttime (met garantie voor een minimum aantal werkuren toegezegd), maar kunnen als de situatie daarom vraagt, worden 'uitgenodigd' meer uren te werken.

Op deze manier kan een redelijk complex rekenmodel *direct* worden ingezet bij de planning van de werkpatronen van medewerkers. Managers kunnen zo redelijk ad hoc besluiten wie zijn verzoek voor een vrije dag gehonoreerd zal krijgen, en wie gevraagd zal worden om 'over te werken'. De schommelingen in volumes van telefoongesprekken worden zo door middel van algoritmes afgestemd op beschikbare menskracht, zonder dat de manager feitelijk statistische berekeningen uitvoert. Het patroon in belvolumes, leidt tot (kort-cyclische) schommelingen in bezetting.

De teneur om steeds vaker bedrijfsstrategie te sturen op basis van voorspellingen beschouwt Gartner als een 'game changer', zij spreken in dit verband van 'Pattern based strategies'. Bedrijven die beter dan hun concurrenten, in (near) real-time, in staat zijn om dit soort patronen te signaleren en te benutten, zullen efficiënter en klantvriendelijker kunnen zijn.

1.3 Big Data en NoSQL

De term Big Data is historisch gezien ongeveer tegelijk in zwang gekomen met de opkomst van NoSQL-databaseoplossingen. Vandaar dat het niet gek is dat deze twee begrippen zo sterk met elkaar worden geassocieerd. Gezien het recente belang dat gehecht wordt aan zogenaamde NoSQL-toepassingen in het kader van Big Data, is het goed om stil te staan bij de rol van deze oplossingen.

Big Data en NoSQL-databases (een afkorting van Not *only* SQL-databases) lijken welhaast onlosmakelijk met elkaar verbonden te zijn. Waarom is dat? De schaalbaarheidsproblemen waar we met RDBMS-en tegenaan lopen, worden in de kern veroorzaakt doordat deze 'traditionele' (relationele) systemen vasthouden aan het ACID-principe: Atomicity, Consistency, Isolation en Durability. Elke transactie (databasebewerking) voldoet in een RDBMS altijd aan alle vier deze principes.

Bedrijven die zeer grote hoeveelheden data proberen te verwerken, merken dat hun traditionele (relationele, SQL-)oplossingen zich maar matig laten opschalen. Reken- of opslagcapaciteit verdubbelen gaat waarschijnlijk nog wel, en daarna dan nóg eens verdubbelen misschien ook. Maar al redelijk 'snel' wordt een plafond bereikt en nemen de kosten voor een grotere en snellere databaseoplossing (veel) sneller toe dan de rekenkracht. Dat is het schaalbaarheidsprobleem (of –fenomeen) dat NoSQL-oplossingen proberen te adresseren, door kostenefficiënt alternatieven te bieden.

De 'transactie' geldt traditioneel als elementaire eenheid van verwerking. In de NoSQL-wereld laten we het ACID-principe los om schaalbaarheidsproblemen het hoofd te bieden. Je zult dan moeten accepteren dat je soms compromissen moet sluiten ten opzichte van (de voordelen van) relationele SQL-verwerking. Met name de relatief 'dure' insert-updates (een bepaald type SQL-transactieverwerking) zitten je bij verwerking van zeer grote hoeveelheden data 'in de weg'.

De elementaire eenheid van verwerking wordt bij NoSQL daarom een *deelverzameling* van de transactie, om optimaal van parallel verwerken te kunnen profiteren. Maar daar kleven ook belangrijke nadelen aan. Wie regelmatig actief is op social media (en giganten zoals Facebook of Twitter maken veelvuldig gebruik van NoSQL-oplossingen), zal het zeker opvallen dat er soms aperte inconsistenties zichtbaar zijn, zoals volgordeverwisselingen, tijdelijke verdwijningen, of dubbele records. Algemeen principe is dat je met alle NoSQL-oplossingen op ten minste een van de vier ACID-principes moet inleveren ten behoeve van betere performance. Of dat verlies in consistentie gerechtvaardigd is, blijft natuurlijk een businesskeuze.

Interactieve games zoals bijvoorbeeld FarmVille op Facebook trekken zo'n 15 miljoen gebruikers per dag. Veel transacties daar zijn inherent collaboratief, en niet per se sequentieel. En dat geldt voor meer 'Web 2.0'-toepassingen. Als je al die transacties volgens ACID-principes zou proberen te verwerken, heb je ongelofelijk dure hardware nodig. Aangezien het toch slechts 'een spelletje' betreft, waarom zou je dan niet wat compromissen

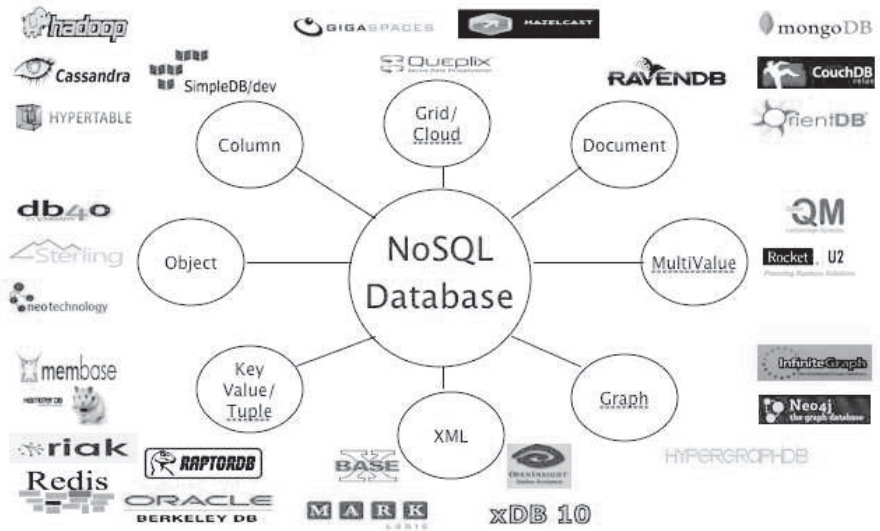
(bijvoorbeeld inconsistenties) accepteren? Andersom: als verwerking door traditionele hardware een vereiste zou zijn, wordt het eenvoudigweg te moeilijk om een rendabele businesscase (meer daar over in hoofdstuk 3) rond te krijgen.

Op zich vinden we consistentie nog steeds belangrijk, maar deze hoeft niet per se *op elk moment* in de tijd te worden gewaarborgd. We kiezen dan voor consistentie *op den duur*. Daarom zie je ook zo vaak in dat soort toepassingen dat er tussen 'refreshes' op je scherm kleine afwijkingen optreden. Voorbeelden van nieuwe technologie die hiervoor ontwikkeld is, zijn 'entangled queries' of 'declarative data-driven coordination'. Maar hoe dan ook, er is geen 'free lunch': door de transactie als eenheid van verwerking op te geven, winnen we (soms dramatisch) aan performance, maar leveren we ook wat in op ten minste een van de vier ACID-principes.

NoSQL-oplossingen worden ingezet waar de traditionele, relationele systemen ons onvoldoende schaalbaarheid (en soms kostenefficiëntie) bieden. Van oudsher zijn we gewend om *in principe* gebruik te maken van relationele technologie. Maar telkens wanneer volumes werkelijk de pan uit rijzen, zou je kunnen bekijken of er een NoSQL-alternatief (equivalent) voorhanden is, dat als vervanger zou kunnen fungeren. Gezien de volwassenheid (of eigenlijk het gebrek daaraan) van NoSQL-oplossingen zul je vanuit beheeroptiek meestal als eerste naar 'gewone' SQL-oplossingen kijken.

Als er een duidelijke IT-behoefte is om meer gegevens, sneller te verwerken, is het normaal dat je naar alternatieven op zoek gaat. De NoSQL-wereld biedt vandaag de dag een heel scala aan alternatieven voor uiteenlopende Big Data-toepassingen. Er zijn speciale toepassingen voor video en beeldmateriaal. Er zijn speciale toepassingen voor tekst, voor Grafen (netwerken, zoals bijvoorbeeld LinkedIn of Facebook), voor ongestructureerde, of semigestructureerde datastromen die snel veranderen (denk aan internet), etc.

Interessant genoeg zijn dit vrijwel allemaal Open Source-oplossingen: kennelijk is het softwareontwikkelmodel dat hier achter zit, waarbij grote groepen ontwikkelaars samen bijdragen aan de ontwikkeling en het debuggen van nieuwe versies, een krachtig mechaniek. Dit collaboratieve model leidt ertoe dat op een heel organische manier de functionaliteit die veel mensen belangrijk vinden (of nodig hebben) ook het eerst aan nieuwe versies wordt toegevoegd. Zelfs softwaregiganten als Microsoft, Oracle, SAP, etc., zijn niet of nauwelijks in staat gebleken NoSQL-innovaties voor te blijven.



Figuur 1.3 Bekende NoSQL-oplossingen

In figuur 1.3 zie je een bescheiden selectie van sommige wat beter bekende NoSQL-oplossingen. Maar zodra dit boek verschijnt is dit overzicht alweer achterhaald, want als gevolg van het Open Source-model achter dit ecosysteem van ontwikkelaars, worden steeds nieuwe platforms gelanceerd, al naar gelang behoefte.

Graph databases zijn een toepassing die specifiek gericht is op netwerken: overal waar mensen gekoppeld zijn aan elkaar zoals bij LinkedIn of Facebook, kun je die structuur beschrijven als een stelsel van personen en koppelingen daartussen. Wiskundigen spreken van knopen (de individuen), en lijnen of zijden, die al of niet een richting hebben. Leonhard Euler was een van de vroege eminente wiskundigen die met zijn grafen-theorie baanbrekend werk verrichtte.

Euler beschreef het probleem van de zeven bruggen van Koenigsberg, en maakte daarbij gebruik van grafen. Het idee dat je zoiets als 'routeplanning' op een hele nieuwe, wiskundige manier kunt benaderen, bleek later veel meer toepassingen te hebben. Aanbevelingen op LinkedIn of Facebook worden op een vergelijkbare manier 'berekend': door te kijken naar de relaties tussen mensen, en hier wiskundige bewerkingen op los te laten, kan een algoritme 'voorspellen' wie je mogelijk nog meer kent, maar die nog niet in jouw netwerk zitten.

Voor logistieke problemen zoals routebepaling is het soms niet mogelijk alle mogelijke 'paden' door te rekenen. Een grafen-model kan dan helpen om een hoop van de mogelijke routes (analytisch) 'tegen elkaar weg te strepen' om zodoende een veel kleiner aantal routes daadwerkelijk door te hoeven rekenen. Als telefoongesprekken of internetverbindingen door een netwerk van switches en hubs geleid moeten worden, spelen vergelijkbare problemen. Ook hier is onwaarachtig veel rekenkracht vereist, en dus kunnen algoritmes die het probleem vereenvoudigen, enorm waardevol zijn.

Het aantal mogelijke toepassingen voor NoSQL-platforms is schier eindeloos. Nog elke dag worden nieuwe toepassingen gevonden en zien nieuwe oplossingen het daglicht. Dit is een opwindende tijd, die een beetje doet denken aan het begintijdperk van de introductie van computers: veel nieuwe oplossingen zijn experimenteel, en nog niet helemaal uitgerijpt. Dit maakt dat (veel) IT-afdelingen huiverig zijn om te gaan werken met nieuwe technologie die zijn bestaansrecht nog niet echt heeft kunnen bewijzen. Alleen de toekomst kan uitwijzen wat de 'winnaars' van deze technologische evolutie zullen zijn.

1.4 Big Data en Hadoop

Big Data en Hadoop lijken in Nederland dusdanig sterk met elkaar verweven dat je bijna de indruk zou krijgen dat je er zonder Hadoop-cluster niet meer bij hoort. Let wel, om vermoedelijk sociologische redenen is dit vooral een Nederlands fenomeen. Apache Hadoop is een Open Source-softwareproject, met een breed scala aan producten dat met name in de BI-wereld zijn nut heeft bewezen.

Hadoop lijkt (vooral binnen Nederland!) zo'n beetje de 'industry benchmark'-technologie voor Big Data geworden. Uiteenlopende toepassingen die zeer grote hoeveelheden data gebruiken, leunen (zwaar) op Hadoop. Naar verluidt heeft Facebook momenteel het grootste Hadoop-cluster draaien: 21 petabyte opslag, in één HDFS-cluster (HDFS = Hadoop Distributed File System). Meer dan 2000 machines die ieder zo'n 12 terabyte aan (gecomprimeerde!) data toevoegen *per dag*. Duizelingwekkende getallen, ook voor degenen die wel vaker met hele grote datasets te maken hebben. Voor wat perspectief: bedenk dat er in Nederland maar weinig datawarehouses te vinden zijn met meer dan 5 à 10 terabyte.

Eigenlijk is Hadoop niet één product, maar eerder een eco-systeem aan producten, waarvan er een aantal in veel opzichten sterke verwantschap vertonen met 'traditionele' BI-tools. Behalve de 'kern', het Hadoop Distributed File System, heb je daarnaast MapReduce, Pig, Hive, HBase, etc., maar

ook Mahout, ZooKeeper, en HCatalog. Met name Hive (data-warehousing), HBase (database) en HCatalog (metadata, samen met Hive), bieden voor BI-specialisten vertrouwde functionaliteit.

Naast deze 'mainstream' Hadoop-producten is er de laatste tijd een heel scala aan nieuwe opties beschikbaar gekomen. Ambari, Hue, Flume, Oozie en Chukwa bieden functionaliteit om Hadoop/HDFS makkelijker te beheren. Dat gebeurt door laagdrempeliger interfaces en scheduling. Met Impala wordt het mogelijk om via SQL toegang te krijgen tot HDFS en Hive, wat belangrijk is als je gegevens uit het Hadoop-cluster naar een 'traditioneel' RDBMS wilt overbrengen. Met Mahout (data mining) en R (Open Source-data-analyse) is er daarnaast functionaliteit voor meer geavanceerde data-analyse.

Hadoop laat zich ook integreren met traditionele (relationele SQL) BI-tools, zodat je van twee walletjes kunt snoepen. De flexibiliteit, snelheid, en kracht van Hadoop, in combinatie met de betrouwbaarheid en robuustheid van relationele BI-platforms. Soms zal de extractie van gegevens uit bronsystemen sneller (en dus goedkoper) kunnen met Hadoop, waarna een subset van de gegevens uit het bronsysteem verder 'gewoon' relationeel wordt verwerkt in een traditioneel datawarehouse (zie volgende paragraaf), bijvoorbeeld met Impala.

Hadoop is geïnspireerd door ontwikkelingen als MapReduce van Google en Google File System (GFS). Binnen het ecosysteem (Hadoop Common) vallen producten als HBase, Hive en ZooKeeper, respectievelijk een database-equivalent, een datawarehouse en een coördinatiedienst voor gedistribueerde applicaties. HBase is de Hadoop-variant die in meerdere opzichten vergelijkbaar is met BigTable van Google (ook NoSQL). HDFS (Hadoop File System) is het equivalent van GFS. Met deze producten samen heb je het geraamte waarmee een NoSQL-equivalent voor een datawarehouse ontwikkeld zou kunnen worden.

Een belangrijk verschil tussen traditionele (relationele, RDBMS) datawarehousesystemen en vergelijkbare functionaliteit in een Hadoop-omgeving is dat een groot gedeelte van 'het geraamte' in Hadoop zelf ontwikkeld moet worden. Dit heeft (grote) voordelen qua flexibiliteit: je kunt zo maximaal de performancemogelijkheden van een NoSQL-oplossing gebruiken. De keerzijde daarvan is ten eerste dat dit (veel) meer technische kennis vergt. Het is alsof je een kant-en-klare motor (RDMS) vergelijkt met een meccanodoos. Die laatste optie biedt meer vrijheid en creativiteit, en kan daardoor 'be-

ter' gebruikmaken van de mogelijkheden van de gebruikte hardware (later meer hierover).

Hadoop-ecosystemen worden aangestuurd met Java, een buitengewoon laagdrempelige programmeertaal. Dat lijkt op zich een voordeel. Doordat er volop 'resources' (= mensen, programmeurs) beschikbaar zijn, is de bemensing van dit soort projecten eenvoudiger, en beter. Er is echter ook een keerzijde aan deze redenatie.

Er zijn mensen die beweren dat de reden waarom applicaties op een Applecomputer betrouwbaarder (minder bugs) en gebruiksvriendelijker zijn, te maken heeft met het feit dat de daarvoor gebruikte programmeertaal (Objective C) juist veel minder toegankelijk is. Doordat de leercurve voor het ontwikkelen van Apple-applicaties veel steiler is dan de leercurve voor vergelijkbare applicaties op een Windows-PC, zal de eerste categorie een meer 'elitaire' groep van ontwikkelaars vergen. Resources met meer ervaring, die meer tijd en moeite hebben moeten besteden om hun vak te leren.

Dat wil overigens niet zeggen dat mensen die toepassingen voor een Windows-PC ontwikkelen per definitie minder competent zijn. Het minimaal vereiste ingangsniveau is lager, *waardoor het mogelijk is* dat er door minder ervaren, minder bekwame programmeurs 'werkende' (maar kwalitatief inferieure) toepassingen kunnen worden opgeleverd. En juist omdat voor een Hadoop-toepassing zo veel programmeerkennis nodig is, en er meer 'zelf' ontwikkeld moet worden, is het risico op zogenaamde legacy-code groter.

Onder legacy-code verstaan we programma's die min of meer doen wat er beoogd wordt, maar die om uiteenlopende reden ofwel nog niet helemaal 'af' zijn, ofwel minder onderhoudbaar en minder veranderbaar zijn door de lagere kwaliteit van de programmacode. Iedereen die wel eens een computerprogramma van een ander heeft moeten wijzigen, weet hoe groot de verschillen zijn tussen door ervaren programmeurs 'netjes' geschreven code, en 'rommelige' code afkomstig van minder competente schrijvers. De laatste categorie is een veelvoud duurder om te onderhouden en te veranderen.

Zodra de mogelijkheden om de bestaande oplossing te veranderen of uit te breiden, steeds moeilijker of duurder worden, spreken we van legacy-code. Het is alsof de voorgaande programmeurs een (negatieve) erfenis hebben nagelaten. Vervangen is na verloop van tijd lastig omdat documentatie vaak ontbreekt. De betrokkenen die de oplossing hebben gebouwd, werken tegenwoordig wellicht ergens anders en na verloop van tijd 'durft niemand meer aan de code te komen'. Totdat er iets *moet* veranderen…

Dit zijn de grootste risico's van het werken met een relatief onvolwassen platform als Hadoop, gebruikmakend van een zeer laagdrempelige programmeertaal zoals Java. Alhoewel Hadoop vaak gekozen wordt om bedrijfseconomische redenen (hardware voor Big Data-initiatieven zou anders te duur worden), moet er wel degelijk rekening mee worden gehouden dat er relatief meer 'met de hand' geprogrammeerd moet worden, en dat vaak ook nog door relatief dure data-scientists. Meer over de nieuwe discipline van 'data-science' in paragraaf 1.6.

Hadoop is nog dermate jong dat er nog wel een aantal upgrades voor nodig zullen zijn vooraleer een grote(re) groep van BI-gebruikers en organisaties hier hun vertrouwen in zullen stellen. Zoals met veel technische innovaties, zijn er early adopters die de weg wijzen. En zoals met veel onvolwassen technologie hebben veel van de early adopters een 'first mover *dis*advantage.' Tot deze weerbarstige techniek verbetert, kleven er nog een aantal belangrijke nadelen aan het werken met Hadoop.

In de traditionele BI-systemen zijn er tal van voorzieningen die beheer van security mogelijk maken. Wie heeft toegang tot welke gegevens? Welke lees- en schrijfrechten heeft die gebruiker? Welke encryptie gebruik je, waar, wanneer, voor welke velden? Wie kan (mag) query's aftrappen, en hoe lang mogen die lopen? Hoeveel (welk percentage) van de beschikbare rekenkracht mogen ze gebruiken, etc.

In een tijd waarin steeds meer primaire processen gebruikmaken van informatie die verkregen is uit het datawarehouse, stijgen de 'eisen' die we stellen aan beschikbaarheid van die informatie. Traditionele BI-systemen hebben een significant hogere 'up-time' dan NoSQL- en Hadoop-oplossingen. Hadoop gebruikt de 'NameNode' voor de toewijzing van bestandslocatie, een beetje vergelijkbaar met NTFS- en FAT-technologie op een (Microsoft) Windows-computer. Deze NameNode is een zogenaamd 'single point of failure', en zorgt er (soms) voor dat gegevens eventjes niet beschikbaar zijn.

Helaas is hardware die zo nu en dan de geest geeft, een 'fact of life' en sterker nog: vooral in NoSQL-oplossingen wordt *bewust* gekozen voor goedkope(re) hardware *omdat* die parallelle hardware in het geval van defecten weer vervangen kan worden. Nou juist dit parallel schakelen van voordelige componenten maakt NoSQL-oplossingen zo veel goedkoper dan zwaargewicht relationele systemen.

Lagere beschikbaarheid is een prijs die je betaalt voor de bedrijfseconomische voordelen van NoSQL-oplossingen die kunnen draaien op (veel!)

goedkopere hardware dan traditionele RDBMS-en die zijn 'opgevoerd' voor topprestaties qua snelheid en volume. Je kunt die beschikbaarheidsproblemen nog enigszins beheersen door iets robuustere hardware te kiezen waar specifiek je NameNode op draait. Maar hoe dan ook: deze techniek is wat minder volwassen, en dit soort van kinderziektes hoort daar een beetje bij. Niets voor niets, of zoals de Engelsen zeggen: 'There is no such thing as a free lunch.'

Een aantal van de grotere partijen die Hadoop-technologie aanbieden om 'Enterprise class'-oplossingen mee te realiseren zijn Cloudere, Greenplum, Hortonworks, of ParAccel, om er een paar te noemen. Gezien de snelle ontwikkelingen in dit veld is deze lijst waarschijnlijk alweer verouderd als dit boek verschijnt. Al deze partijen hebben ook oplossingen die een of andere koppeling met 'traditionele' SQL mogelijk maken.

Praktisch gesproken zijn er enkele voor de hand liggende domeinen waar de volumes van gegevensverwerking dusdanig hoog zijn dat NoSQL-oplossingen in beeld komen: opslag van internetgegevens (voor websites die veel bezoekers hebben), RFID-verwerking, en locatie- en telecomgegevens. Als strikte (nagenoeg 100%) accuratesse te allen tijde geboden is, kiest men niet zo vaak voor NoSQL. Het is een kwestie van tijd (een of enkele jaren) voor de meeste van deze zorgpunten zullen worden geadresseerd in nieuwe NoSQL-oplossingen.

Anno 2013 denken we aan Hadoop-toepassingen (met hun inherente nadelen) bij toepassingen die in de orde van grootte van 5-10 terabyte moeten werken. Daaronder zijn er dikwijls nog kostenefficiënte alternatieven. Daarboven, zeg vanaf tientallen terabytes, zul je (vrijwel) zeker NoSQL-oplossingen in overweging moeten nemen. Door de voor- en nadelen tegen elkaar af te wegen, zal per situatie bekeken moeten worden wat de meest rationele keuze is.

1.5 De relatie tussen NoSQl en business intelligence

Big Data zijn een autonome ontwikkeling: er zijn businesstoepassingen die om opslag en beheer van (zeer) grote hoeveelheden data vragen. Traditionele platforms kunnen niet of nauwelijks voldoen aan de eisen betreffende performance (snelheid) en kostenstructuur (te duur, en onvoldoende schaalbaar) van deze *bedrijfsinnovaties*. En dus kiezen business-stakeholders voor NoSQL-oplossingen. Dikwijls zonder IT of BI daar (in een vroeg stadium) bij te betrekken.

AutomatiseringGids

Big Data

Big Data is dé business intelligence (BI) trend van de laatste tijd en lijkt dit jaar helemaal 'door te breken'. De continue introductie van nieuwe technologie, de dalende kosten voor gegevensopslag en het alsmaar krachtiger worden van computers, maken het mogelijk extreme gegevensvolumes op een betaalbare manier te verwerken. Hoe ga je als bedrijf om met deze schat aan informatie? Welke kansen biedt het? En hoe gebruikt u de data vervolgens voor het behalen van bedrijfsdoelstellingen?

De auteur neemt u in dit boek mee in het universum van Big Data. U maakt kennis met de laatste technologische mogelijkheden en trends, maar vooral ook met zaken als privacy, return on investment en de kwaliteit van data en data-analyse. Daarnaast helpt dit boek bij het opstellen van een goede datastrategie. Door middel van tips en stappenplannen helpt de auteur u ook direct praktisch op weg.

Doelgroep
Dit boek is bestemd voor managers, marketeers, business intelligence specialisten en IT managers. Kortom: voor iedereen die door de aard van zijn werk (mogelijk) te maken heeft met grote hoeveelheden data, en wil weten hoe die optimaal uitgenut kunnen worden om bedrijfsdoelstellingen te helpen realiseren.

Tom Breur studeerde aan het Massachusetts Institute of Technology en is werkzaam als consultant. Hij is gespecialiseerd in datamining en –analyse, en marktonderzoek en heeft werkervaring bij zowel profit als non-profit organisaties. Daarnaast is hij veelgevraagd spreker op congressen en seminars.

Deze uitgave verschijnt in de AG-boekenreeks die in samenwerking met AutomatiseringGids wordt samengesteld. Eerder verscheen in deze reeks de uitgave Business Logic Management (ISBN 9789012585545).

ISBN 978 90 12 58567 5
NUR 980

ACADEMIC SERVICE

9 789012 585675 >

www.academicservice.nl