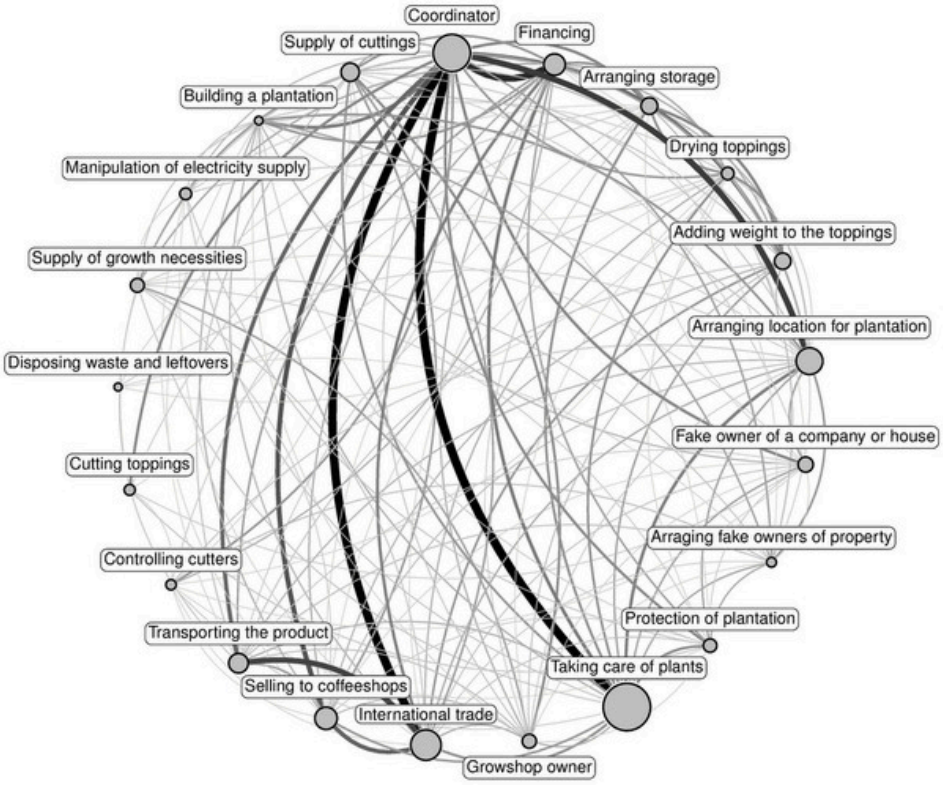
Big Data: technologie verkenning voor het Ministerie van Veiligheid & Justitie

Tony Busker, Jan Kroon, Mortaza Shoae Bargh



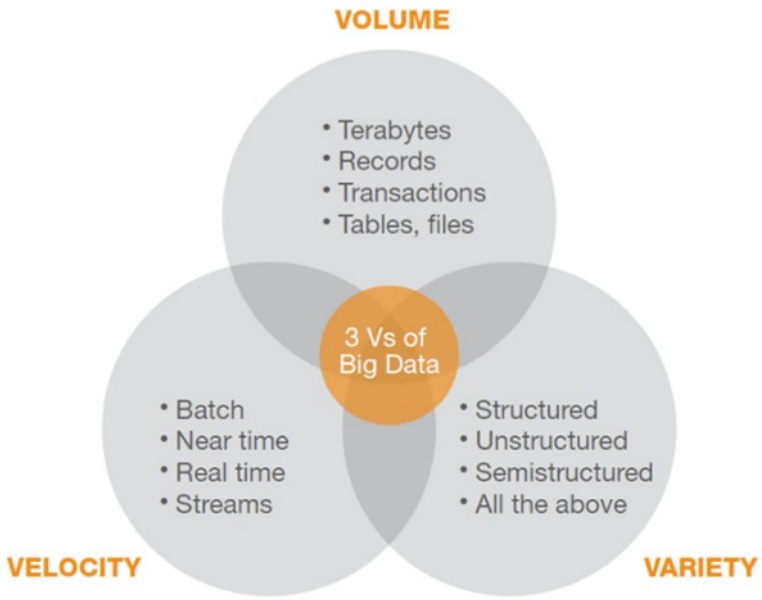
Instituut:	Kenniscentrum Creating010
Auteur(s):	Tony Busker, Jan Kroon, Mortaza Shoae Bargh
Reviewer(s):	Ahmad Omar,
Functie auteur(s):	Lectoren en Docent-Onderzoekers
Verantwoordelijk directeur:	Paul Rutten
Datum:	31 augustus 2015
Status:	Definitief

Versiegeschiedenis

Versie	Status	Datum	Opmerkingen
0	Interviews experts	Jun. 2014	Verslag expert interviews van Nathalie Stembert
1	Werkversie met actiepunten	Nov. 2014	Tekstbijdragen van Tony Busker, Jan Kroon, Paul Rutten. Interne review Mortaza Shoa Bargh. Ter informatie verstuurd naar opdrachtgever.
2	Gereed voor finale review	Dec. 2014	Ter review verstuurd naar opdrachtgever Frank Willemsen
3	Gereed voor interne review	Jun. 2015	Naar aanleiding van bespreking met opdrachtgever op dinsdag 14 april 2015 zijn algemene, beschouwende teksten verwijderd. Huidige data processing capabilities en dashboards WODC en een concreet Big Data aktieplan toegevoegd.
4	Gereed voor finale review	Jun. 2015	Ter review verstuurd naar opdrachtgever Frank Willemsen
5	Definitieve versie	Aug. 2015	Review commentaar verwerkt.

Inleiding

We leven in het tijdperk van **Big Data**. Het fenomeen is niet alleen groot, maar ook complex, gevarieerd en vaak real-time. De potentiele waarde van Big Data wordt als enorm ingeschat. Big Data is “the next big thing”, “de nieuwe olie” [Hinssen 2015]. Er wordt gesproken over Big Data als er sprake is van de drie V’s: (1) *Volume*, grote hoeveelheden, (2) *Variety*: grote variëteit in data formats en (3*) Velocity*, hoge snelheid van gegevensverwerking. Het gaat dus om het verwerken en analyseren van grote hoeveelheden data, in verschillende formaten die met hoge snelheid worden gegenereerd en geanalyseerd omdat mogelijk de waarde in tijd snel afneemt. Soms wordt er een vierde V als kenmerk toegevoegd: (4) *Veracity*, het waarheidsgehalte. Doordat de gegevens uit soms minder betrouwbare bron komen, of gegevens gekoppeld zijn met minder dan 100% betrouwbaarheid moet er eigenlijk ook expliciet aangegeven worden hoe betrouwbaar de analyses zijn die gemaakt worden.



Men spreekt ook van Big Data wanneer de hoeveelheid gegevens zó omvangrijk en divers is dat ze niet te beheren is met de gebruikelijke middelen, zoals conventionele databases of DataWareHouses. De hoeveelheid data groeit snel, verandert voortdurend en is diffuus en ongestructureerd van aard. In verband met big data wordt ook vaak gesproken over complexiteit, waarde en relevantie. Er bestaat een duidelijke behoefte om te willen weten wat we met al die data kunnen en wat we ermee zouden willen, opdat we er niet voor niets veel geld en energie insteken. Big data is daarmee eigenlijk een grote belofte die erop wacht ingelost te worden. Alleen is nog niet helemaal duidelijk waarom en hoe precies.

Om de waarde van Big Data te kunnen ontdekken en benutten en ermee te kunnen experimenteren en leren, zijn in ieder geval nieuwe tools, technieken en skills nodig. Die zijn nu nog onvoldoende of in het geheel niet voor handen. McKinsey voorspelt in ‘Big Data: the next frontier for innovation, competition and productivity’ [McKinsey 2011] dat de Verenigde Staten in 2018 een tekort telt van 140 tot 190 duizend mensen met ‘deep analytical skills’ die nodig zijn om de oogst van Big Data binnen te halen. Hierbij gaat het bijvoorbeeld om wiskunde, statistiek en machine learning. Ook

circuleren er tal van bedragen over de mogelijke toegevoegde waarde van Big Data voor de mondiale economie.

Onderzoeksvragen

Bij het Ministerie van Veiligheid en Justitie leeft een aantal vragen over de mogelijkheden en eventueel toekomstig gebruik van Big Data voor de eigen praktijk van voorbereiding en ontwikkeling van beleid, maar ook met het oog op de productieve toepassing van Big Data in de praktijk van preventie en binnen de uitvoering van taken in de context van de strafrechtketen. Die vragen zijn leidend in de technische verkenning die is uitgevoerd voor dit ministerie, waarvan de resultaten in dit rapport zijn vastgelegd.

1. Welke type big data en welke daarbij passsende databronnen zijn relevant en bruikbaar voor het werkterrein van het Ministerie van Veiligheid & Justitie?

2. Welke infrastructuur, tools en skills die bijvoorbeeld worden toegepast en relevant zijn binnen wetenschappelijk onderzoek en binnen big data onderzoek in het bedrijfsleven zijn mogelijk relevant voor de toepassing binnen het werkterrein van het Ministerie van Veiligheid en Justitie?

3. Wat is het belang van 'klassieke' begrippen als validiteit en betrouwbaarheid in de context van big data analyses en toepassingen? Gelden ze op dezelfde wijze als in de bestaande onderzoekspraktijk of krijgen ze binnen big data een andere betekenis en invulling?

4. Hoe zouden specifieke toepassingen van de mogelijkheden van big data voor het Ministerie van Veiligheid & Justitie er uit kunnen zien?

5. Is het reëel om bestaande vormen van onderzoek, waaronder bestaand monitoring onderzoek van het Ministerie van Veiligheid & Justitie te vervangen door big data toepassingen of zijn ze eerder complementair aan de bestaande vormen en praktijk?

6. Wat kan het Ministerie van Veiligheid & Justitie op de korte termijn doen om werk te maken van toepassing van big data toepassingen, waaronder real-time analytics?

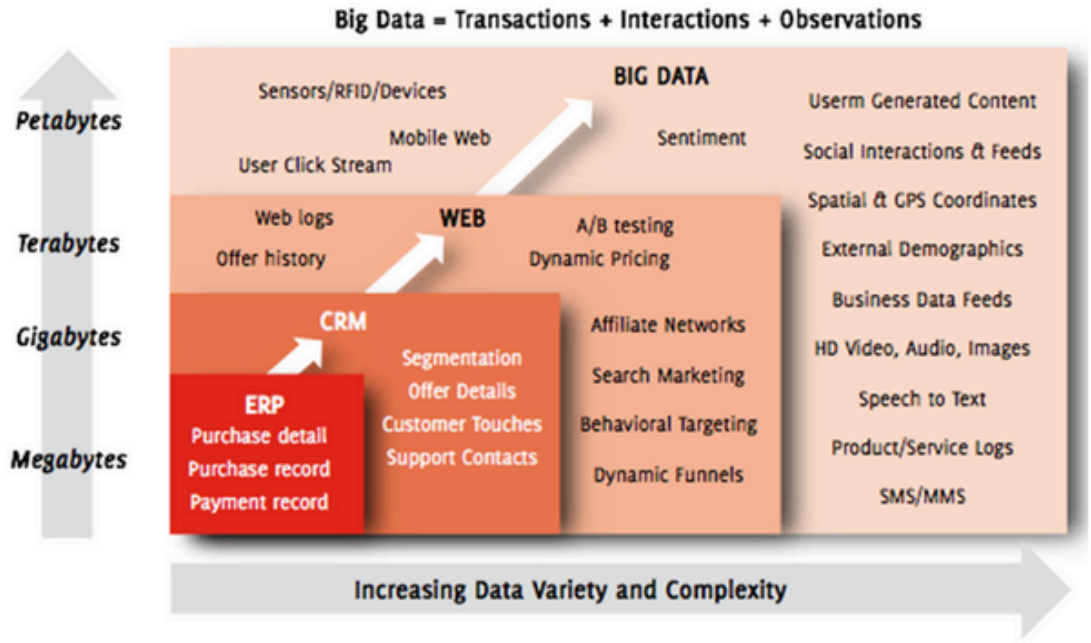
Technische aspecten van Big Data

Wat is nieuw?

Veel van wat nu onder de vlag van 'Big Data' wordt gepresenteerd is niet nieuw en werd eerder Data Analytics, Business Intelligence, Data Warehousing of Data Mining genoemd.

Toch is 'Big Data' zeker geen *"oude wijn in nieuwe zakken"*. Hoewel de analytics methoden bekend zijn, is de data waarop die analyses uitgevoerd worden revolutionair veranderd en is ook de manier waarop de analyses worden uitgevoerd radicaal nieuw.

Big Data Bronnen



Al sinds de jaren '80 zijn bedrijven bezig hun processen te automatiseren met Enterprise Resource Planning (ERP) systemen. Het Nederlandse Baan was één van de pioniers op dit gebied. SAP, JD Edwards, Microsoft en Oracle zijn andere ERP leveranciers. In ERP systemen worden alle bedrijfsprocessen integraal geautomatiseerd, waardoor allerlei gegevens als productiecijfers, verkoopcijfers, voorraden en levertijden online beschikbaar kwamen voor analyse en rapportages. Je zou deze gegevens *Transactions* kunnen noemen, zoals transacties tussen het bedrijf en de buitenwereld en tussen bedrijfsonderdelen onderling.

In de jaren '90 gingen bedrijven ook klantcontacten registreren in Customer Relationship Management (CRM) systemen. Siebel, PeopleSoft, Salesforce.com en PerfectView zijn bekende CRM leveranciers. Met CRM systemen kan een bedrijf de relatie met haar klanten beheren, en werken aan klantbehoud en optimale verkoop aan bestaande klanten (up-sell, cross-sell). Initieel waren de contacten vooral persoonlijke verkoopbezoeken of contacten per brief of telefoon.

Toen het World Wide Web, rond 2000, doorbrak bij het grote publiek kwamen er gegevens over klikgedrag op websites en webshops bij. Ranking in zoekresultaten van Google werd een issue. De opgeslagen gegevens bestaan inmiddels al lang niet meer alleen uit *Transactions.* Ook alle

Interactions met klanten en bezoekers worden geregistreerd.

Vanaf circa 2010 hebben Social Media een grote vlucht genomen. Er vinden nu ook op andere dan de bedrijfswebsite of webshop voor het bedrijf relevante activiteiten plaats: berichten op Twitter, posts op Facebook, messages in WhatsApp. Overheidsdiensten maken grootschalig gebruik van camerabewaking. Sensoren mobiele telefoons, fitbits, smart watches en auto's zorgen voor nog meer, en vrijwel altijd real-time stromen van data. Je zou kunnen zeggen dat de gegevens behalve *Transactions* en *Interactions* nu ook *Observations* van velerlei soort omvatten.

Transactions	88%
Log data	73%
Events	59%
Emails	57%
Social Media	43%
Sensors	42%
External Feeds	42%
RFID scans or Point Of Sale data	41%
Free-form text	41%
Geospatial data	40%
Audio	38%
Still Images or Video	34%

In [IBM 2014] heeft IBM onderzocht welke databronnen op dit moment aangeboord worden door bedrijven en instellingen die met Big Data bezig zijn (Het onderzoek omvatte ruim 800 bedrijven in 95 landen).

Behalve dat het type van de opgeslagen data van karakter is veranderd, is ook de hoeveelheid data enorm toegenomen. In het ERP tijdperk hadden we nog te maken met MegaBytes (10^6 = 1 miljoen Bytes), een hoeveelheid gegevens die ook toen al op een harddisk paste. In het huidige Big Data tijdperk werken we met PetaBytes (10^15 Bytes). Bovenstaande plaatje geeft eigenlijk een totaal verkeerd beeld. Big Data lijkt in het figuur qua oppervlakte ongeveer 20 keer zo groot te zijn als ERP data. In werkelijkheid is het een miljard keer zo groot! Traditionele manieren om data te verwerken zijn dan ook niet meer toereikend!

Relevante 'Big Data' bronnen voor het Ministerie van Veiligheid & Justitie

De oorsprong van gegevens die gebruikt worden in de praktijk van Big Data is heel divers en in principe eindeloos. Het is onmogelijk om een volledig overzicht van bronnen van gegevens over transacties, interacties en observaties op te stellen. Wel kunnen we aangeven wat op dit moment de meest gebruikte Big Data bronnen zijn. Hieronder volgt een overzicht met een korte typering, mede gebaseerd op de gesprekken met experts. We kunnen databronnen typeren aan de hand van inhoudelijke kenmerken, maar ook aan de hand de wijze waarop ze worden verzameld en de 'locaties' waar ze te vinden zijn.

Online informatie (tekst, beeld, geluid, interactief)

De hoeveelheid informatie die via het web gepubliceerd wordt is gigantisch. Daarbij gaat het om tekstuele informatie, audio, audiovisuele informatie en allerlei interactieve toepassingen, zoals games. Deze informatie is een belangrijke grondstof voor het genereren van Big Data. Daarbij gaat het zowel om het analyseren van patronen in de grote informatiezee die het web is als het zoeken naar outliers die juist de uitzondering op de brede patronen vormen en om alle denkbare vormen die

daartussen liggen. De verzameling van dit soort data gebeurt met zogenaamd web crawls. Het web wordt systematisch en automatisch afgegraasd op basis van een gerichte zoekvraag. Daarbij gaat het om allerlei soorten gegevens, om webdocumenten als krantenartikelen, maar ook verbindingen tussen online data bronnen (hyperlinks), annotaties van organisaties en bedrijven, video, audio etc. Het oogsten van data op basis van tekstuele informatie is momenteel nog dominant, vanwege de complexiteit van het oogsten en interpreteren van andere vormen van informatie.

Online zoekgedrag

Ook allerlei door mensen gegenereerde data kunnen worden geoogst en geanalyseerd. Hieronder verstaan we data die voortkomen uit gedrag van mensen gericht op een ander doel dan het produceren van informatie bedoeld voor intermenselijke communicatie. Een bekend voorbeeld daarvan is online zoekgedrag van mensen. Zoekgedrag is een bijzondere vorm van mens-computer interactie, die kan worden bijgehouden en geanalyseerd met behulp van statische querylogs. We hebben hiervoor al aangegeven dat menselijk zoekgedrag op het internet een belangrijke bron is voor voorspellingen en trendonderzoek.

Menselijk gedrag in de openbare ruimte

Behalve via online zoeken, genereren mensen informatie door andere vormen van gedrag die op welke wijze dan ook geregistreerd en opgeslagen wordt. Op dat moment wordt gedrag omgezet in data en worden die data, mits in grote hoeveelheden opgeslagen, (tot grondstof voor) Big Data. Gedrag geregistreerd in de openbare ruimte door bijvoorbeeld camera's, zorgt voor relevante data. Het belang van de openbare ruimte als databron zal in de toekomst met de groei van de toepassing van sensoren, verder in belang toenemen. Een minder directe manier van het genereren van data op basis van menselijk gedrag is het registreren van data uitgezonden door mobiele apparaten zoals mobiele telefoons, tablets en navigatiesystemen. Mensen dragen die bij zich of ze zitten in hun voertuigen en gebruiken ze bijna voortdurend. Door het uitgezonden signaal kan het verplaatsingspatroon van individuen vrij nauwkeurig worden vastgesteld, maar is het ook mogelijk bewegingen van groepen en massa's te monitoren of te onderzoeken. Dat laatste is bijvoorbeeld relevant voor 'crowd control' bij grootschalige evenementen.

Vooralsnog blijkt het echter niet eenvoudig om grote hoeveelheden van verschillende typen data te analyseren. Daar ligt nog een behoorlijke uitdaging. Bij de analyse van camerabeelden bijvoorbeeld is het nu nog vaak nodig om de menselijke interpretatie in te schakelen om te kunnen vaststellen '... of mensen aan het feestvieren of aan het vechten zijn.' Geautomatiseerde patroonherkenning is een belangrijk onderzoeksveld om hier vorderingen te kunnen maken. Een van de geïnterviewde experts gaf het voorbeeld van het gedrag van zakkenrollers: " Zij gaan in de rij staan en vlak voordat ze dan de beurt zijn dan gaan ze weer achter in de rij staan". Het automatisch herkennen van patronen, maar ook het herkennen van gezichten en expressies, is van essentieel belang om gebruik te kunnen maken van deze vorm van ongestructureerde data. Bij de andere hier beschreven vormen van data is het toekennen van betekenis aan en het trekken van conclusies uit de gegenereerde data een belangrijke uitdaging. Hoe moet verplaatsingsgedrag uitgelegd worden en wat is er mee te verklaren?

Menselijke uitingen en communicatie

Een al langer gebruikte bron van informatie, in het bijzonder in het kader van opsporingsactiviteiten door Justitie is informatie uit communicatie tussen mensen: inhoud van telefoongesprekken en schriftelijke communicatie. Met de komst van digitalisering zijn daar nieuwe vormen bijgekomen, in eerste instantie e-mail, later ook communicatie via (semi-openbare digitale media ook wel aangeduid als social media. De door mensen doelbewust geproduceerde en voor communicatie bedoelde informatie die via de sociale media wordt gedeeld is niet in alle gevallen de meest optimale bron

voor preventie en opsporing, maar biedt wel oneindig veel andere mogelijkheden. Sociale media zijn bijvoorbeeld een waardevolle bron voor sentiment analyse, bijvoorbeeld als het gaat over gevoelens van veiligheid en onveiligheid. Er zijn bovendien mogelijkheden om de inhoud van de te analyseren uitingen te verbinden aan andere data over de personen die de bron van de boodschappen zijn. Dat varieert van locatiegegevens op data die direct te herleiden zijn uit technische componenten van de boodschap (GPS locaties, timestamps en hash tags) en de afzender, tot het achterhalen van de leeftijd en geslacht van de persoon die een bericht post. Op basis van deze toepassingen ontstaan alternatieven voor bepaalde vormen van publieksonderzoek. Het is dan ook niet verwonderlijk dat deze mogelijkheden volop worden benut voor onderzoek naar producten, diensten en merken, met name het gebruik en de positionering ervan.

Interne en ingevorderde data

Organisaties en bedrijven beschikken zelf ook over data die traditioneel niet gezien werden als mogelijke bron van informatie voor het beter begrijpen en beheersen van processen, taken, klanten en gebruikers. Daar hebben we in de vorige paragraaf ook al aan gerefereerd. Dat geldt ook voor het WODC en nog meer voor het Ministerie van Veiligheid en Justitie. Daarbij gaat het bijvoorbeeld over data uit diverse onderzoeksmonitoren. Daarnaast beschikt het ministerie over grote hoeveelheden data over zeer uiteenlopende dingen. Er is uitgebreide informatie over verschillende justitiële cases, zoals duur, hoeveelheid en aantal betrokken rechtbanken. Ook is er informatie over aangiftes, misdrijven, opsporing (telefoon taps, location tracking), profielen van daders, menselijk inzet enzovoort. Onder andere vanuit het besef dat data kunnen dienen in de optimalisatie van processen is het project 'Verbetering prestaties in de strafrechtketen'. Een van de deelprojecten is 'Uitvoeringsketen strafrechtelijke beslissingen'. Dat beoogt de tenuitvoerlegging van straffen zo goed en efficiënt mogelijk te laten verlopen. Onderdeel daarvan is het beter gebruik van data. Daartoe is een nieuw datacentrum opgezet dat draait bij het Centraal Justitiële Incassobureau, het CJIB. Gegevensuitwisseling tussen het CJIB, Dienst Justitiële Inrichtingen (DJI), reclassering en het Openbaar Ministerie (OM) kan zorgen voor belangrijke inzichten en bijdragen aan beoogde optimalisaties. Interne data kunnen, indien van toepassing, aangevuld en verrijkt worden met ingevorderde data op basis van de bevoegdheden die bij wet aan het Ministerie zijn toegekend. Daarbij kan het gaan om computers, mobieltjes of harde schijven die zijn ingenomen bij verdachten of om gegevens over gebruik van mobiele communicatie apparatuur. Dit soort gegevens kunnen in theorie in combinatie met gegevens binnen andere onderdelen van het overheidsapparaat, bijvoorbeeld de Belastingdienst, interessante inzichten genereren. Om data uit te wisselen tussen overheidsinstanties moet er op dit moment nog een specifieke vraag en een onderzoeksplan worden ingediend. Bij toestemming kunnen er geanonimiseerde data gedeeld. Mogelijk kunnen daarin ook combinaties gemaakt worden met andere vormen van data die we hiervoor hebben besproken. Dat is bijzonder complex en in sommige gevallen aan restricties gebonden op basis van bijvoorbeeld privacy wetgeving. Het debat daarover wordt momenteel uitgebreid gevoerd als onderdeel over de discussies over de voorgestelde Datawet.

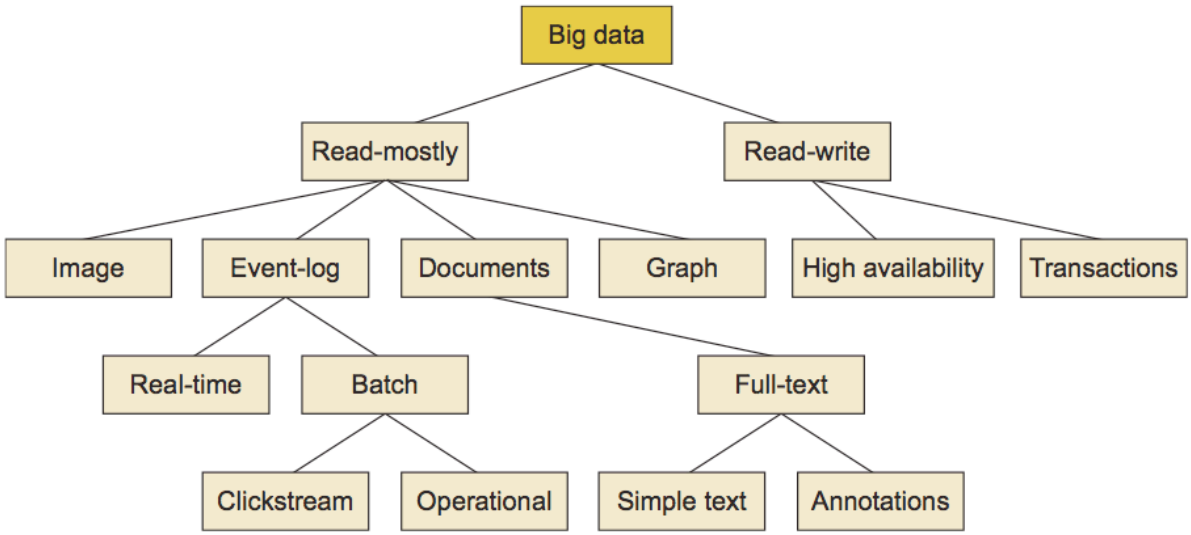
Traditionele databronnen

Het WODC en het Ministerie van Justitie werken nauw samen met bijvoorbeeld het CBS en externe bureaus bij het (laten) uitvoeren van onderzoek. In die gevallen gaat het om uitkomsten van onderzoek op basis van meer traditionele onderzoeksmethoden als survey-onderzoek, en data uit meer formele bronnen, bijvoorbeeld de Kamer van Koophandel, Kadaster of het bevolkingsregister. Er wordt volop gespeculeerd over de vraag of Big Data geheel of gedeeltelijk de plaats kan innemen van bijvoorbeeld het survey-onderzoek. Daar komen we in het volgende hoofdstuk nog op terug. Voor een ander deel van het onderzoek, in het bijzonder de meer structurele data over demografie, voorzieningen, infrastructuren en geografie geldt dat hun waarde toeneemt wanneer ze gecombineerd worden met de data uit andere bronnen, die we hiervoor hebben besproken.

Tenslotte willen we hier ook nog wijzen op de mogelijkheden om op traditionele wijze, op papier vastgelegde informatie te introduceren in het domein van Big Data. In paragraaf 2.4 verwezen we al naar de informatie die vervat is in de almaar in omvang groeiende papieren dossiers waar rechters mee geconfronteerd worden. Door deze informatie te digitaliseren wordt het mogelijk deze conventioneel opgeslagen gegevens op een Big Data manier te analyseren, om op die manier complexiteit te reduceren en verbindingen met andere Big Data te leggen. [Zie onder meer voor methodieken van digitale 'objectherkenning' in documenten: http://www.doi.org/].

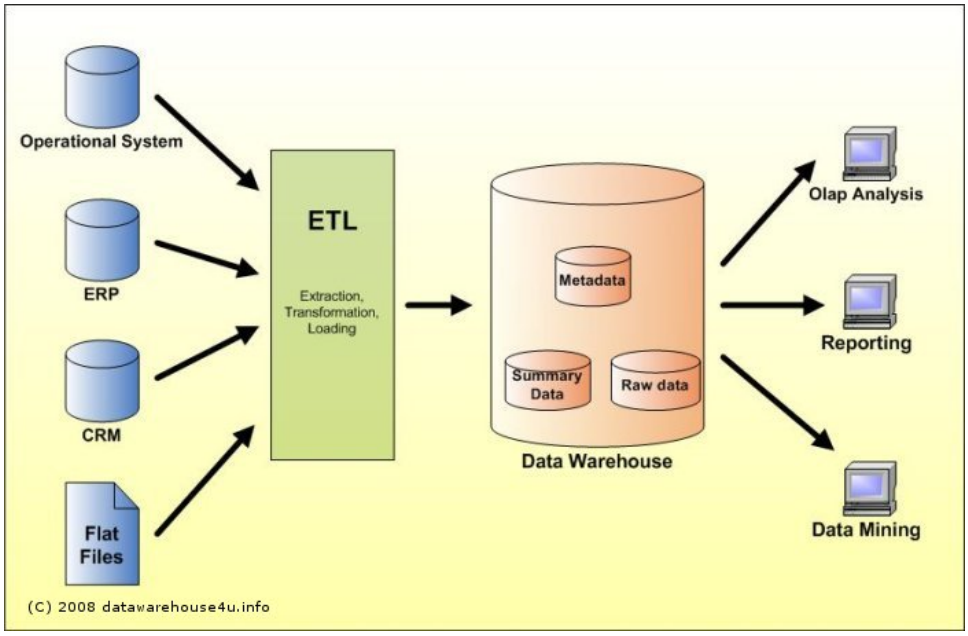
Big Data data types en type databases

Onderstaande figuur laat zien hoe de type data uit de verschillende mogelijke databronnen zich tot elkaar verhouden.



Bron illustratie: www.wikipedia.com

Een groot deel van deze data zijn goed te modelleren in een traditionele tabel-structuur, en zijn relationele (SQL) databases geschikt om analyses uit te voeren. De traditionele DataWareHouse architectuur bestaat dan ook voornamelijk uit relationele databases met verschillende rollen tussen processen die de data bewerken en verplaatsen.



Bron illustratie: www.datawarehouse4u.info

In bovenstaande figuur staan links de operationele systemen van een bedrijf. Het ETL proces (Extraction, Transformation, Loading) zorgt ervoor dat er regelmatig gegevens van deze systemen gekopieerd worden naar het DataWareHouse. Zonodig worden gegevens hierbij ge-üniformeerd, zodat bij praktische problemen als bijvoorbeeld klantgegevens in een CRM systeem onder klantnaam zijn opgeslagen en in een ERP systeem gegevens over dezelfde klant onder klantnummer zijn gerangschikt het systeem alle gegevens relateerd aan dezelfde klant. Dit ETL proces voltrok zich doorgaans 's-nachts zodat de volgende dag alle gegevens netjes in het DataWareHouse beschikbaar waren voor verdere bewerking en analyse.

De twee grote denkers over het DataWareHouse concept zijn Bill Inmon [Inmon 2002] en Ralph Kimball [Kimball 2002]. Bill Inmon introduceerde de term DataWareHouse en zijn ideeën in 1990. Hij zag het bouwen van een DataWareHouse als een enorm project, dat top-down voor de hele onderneming alle relevante gegevens voor besluitvorming moest verzamelen en ordenen. Welke klanten leveren het meest winst op? Welke klanten stappen over naar de concurrentie en waarom?

Ralph Kimball propageerde in 1996 een alternatieve, bottom-up aanpak: klein beginnen met *Data Marts* die alleen gegevens bevatten die voor een bepaalde afdeling van een bedrijf relevant zijn. Door verschillende Data Marts te koppelen met een *Enterprise Data Bus* ontstaat op organische wijze een DataWareHouse, was zijn idee.

Overeenkomst tussen beide strategieën is dat gegevens die belangrijk zijn voor beslissers vanuit de operationele systemen gekopieerd worden naar een apart systeem, het DataWareHouse, en daar worden opgeslagen en geanalyseerd. De operationele systemen zijn bedoeld *"to run the business".* Het DataWareHouse is bedoeld *"to optimize the business processes".* Een DataWareHouse is eigenlijk een *Decision Support System* voor de bestuurders en/of beleidsmakers.

Ralph Kimball versus Bill Inmon Comparison

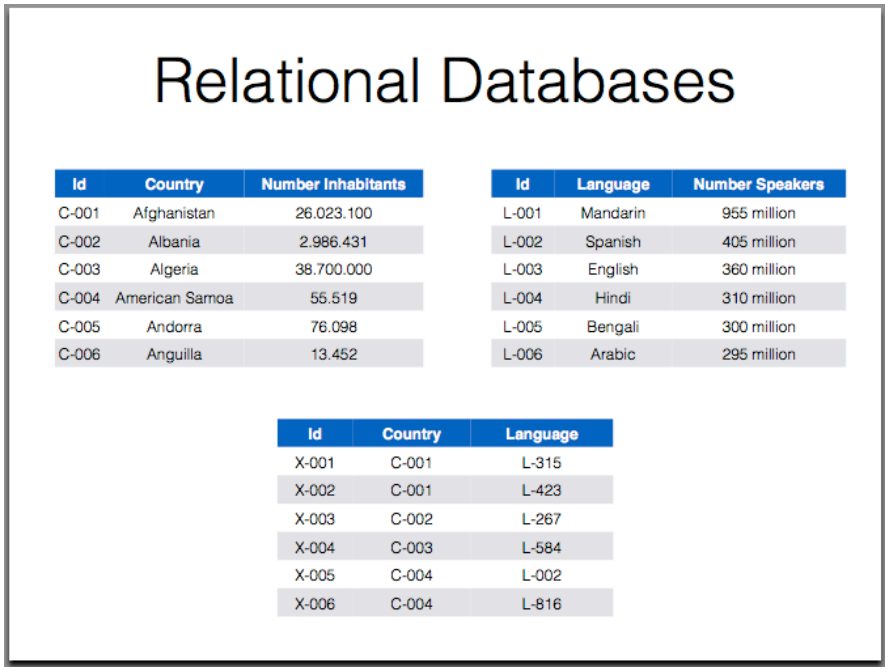
	Kimball	Inmon
Need	Immediate	Longer time scale
Drive	Business areas	Enterprise
Budget	Smaller budget	Larger budget
Requirements	Volatile	More stable and growing
Customer	User base	Corporate
Sources	Stable	Changeable
Startup cost	Lower	Higher
Projects	Same cost as start up	Cheaper than start up

Bron illustratie: http://www.slideshare.net/mikejf12/data-warehouse-inmon-versus-kimball-2

Relationele (SQL) Databases

In relationele databases worden gegevens opgeslagen in tabellen. Om te voorkomen dat er veel data wordt gerepliceerd, wordt er een database ontwerp gemaakt met verschillende tabellen die naar elkaar kunnen verwijzen.

Stel dat we de volgende gegevens willen opslaan van alle ambtenaren op een ministerie: naam, afdeling, functie, email-adres, mobiele-telefoonnummer, zakelijk-postadres. Als we dit in één grote tabel doen (zoals bij Excel mogelijk is), bevatten heel veel regels hetzelfde zakelijk-postadres. Het is handiger twee tabellen te maken: een ambtenaren-tabel en een afdelingen-tabel. In de afdelingen-tabel staat dan per afdeling slechts één maal het zakelijk-postadres vermeld. Vanuit de ambtenaren-tabel verwijst het veld ‘afdeling’ naar een regel in de ambtenaren tabel.

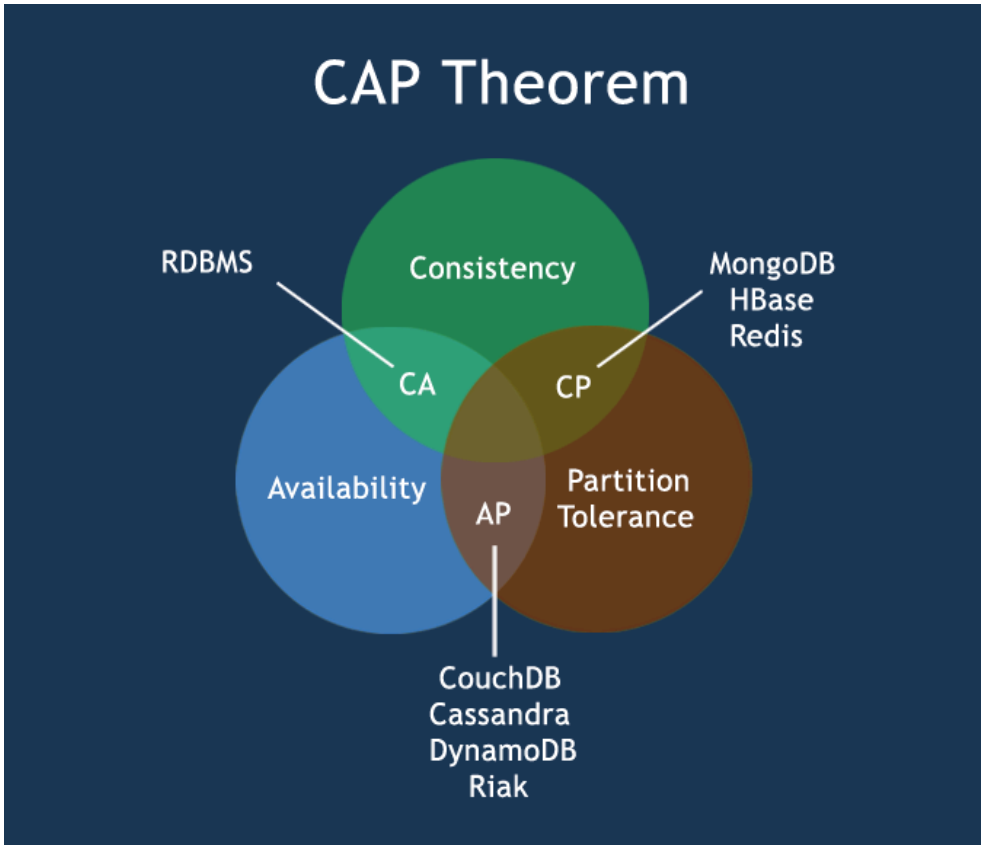


Voorbeeld van een SQL query is:

```
SELECT * FROM Countries WHERE Language = 'English'
```

Om deze opdracht uit te kunnen voeren zal de database weer enkele tabellen moeten samenvoegen, en uit de grote join-tabel een selectie van regels maken waar 'English' in de Language kolom staat. Deze joins nemen veel tijd en veel geheugenruimte in beslag. Als je erg veel data in korte tijd wilt verwerken, worden de joins de bottleneck en loopt het hierop vast. In jargon wordt dit de *'join-bomb'* genoemd.
Een tweede probleem van relationele databases is dat een tabel niet over meerdere computers (nodes) verdeeld kan worden. Dit legt beperkingen op aan de hoeveelheid data die verwerkt kan worden, en aan de snelheid waarmee dat gebeurt. Deze eigenschap wordt *'partition intolerance'* genoemd.
Bij Big Data heeft een belangrijk deel v deze data een structuur die minder geschikt is om in een relationele (SQL) database op te slaan, en zijn andere typen database (NO-SQL: Not Only SQL) meer geschikt.

Alternatieve No-SQL databases kunnen wel over meerdere computers (nodes) verdeeld worden, maar hebben dan soms weer als nadeel dat niet gegarandeerd kan worden of alle gegevens consistent zijn of dat bepaalde gegevens soms tijdelijk niet beschikbaar zijn.



Bron illustratie: www.w3resource.com

Zoals wel vaker het geval is, bestaat een oplossing met alleen maar voordelen niet. Je zult moeten kiezen welke eigenschappen: Consistentie (Consistency), Beschikbaarheid (Availability) of Mogelijkheden om gegevens over meerdere systemen te verdelen (Partition Tolerance) de doorslag geven. Dit dilemma wordt het CAP Theorem genoemd.

In praktijk betekent dit dat bij Big Data vaak gewerkt wordt met een combinatie van SQL en NO-SQL databases.

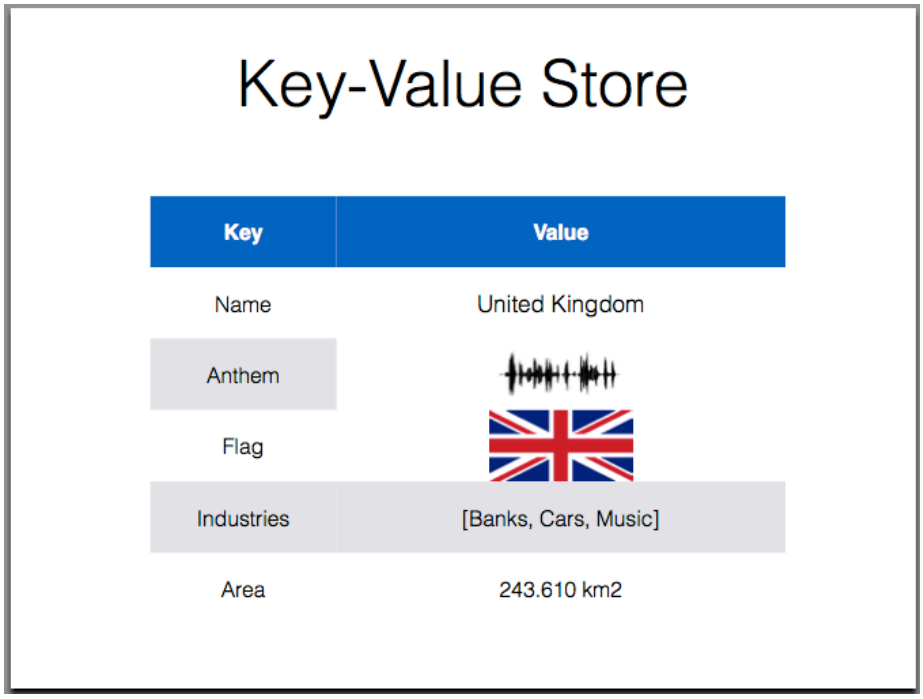
NO-SQL (Not Only SQL) Databases

NO-SQL is een niet-relationeel database management systeem wat op een aantal belangrijke onderdelen verschilt van een traditionele relationele database management systeem. Ze zijn ontworpen om zeer grote hoeveelheden data op te kunnen slaan en te bewerken. Voor deze systemen is het in veel gevallen niet nodig om een fixed schema aan te leggen. Om grote hoeveelheden te kunnen bewerken worden join-operaties vermeden. Het schalen van deze opslagsystemen is meestal door het bijplaatsen van meer servers (horizontaal schalen).

Wanneer een applicatie miljoenen queries per seconde moet kunnen verwerken kan dat alleen bereikt worden door meer servers aan de infrastuctuur toe te voegen. Dit kan op een goedkoop en relatief eenvoudige wijze bereikt worden met NoSQL. Dit in tegenstelling tot de meer complexe schaalbaarheid van traditionele SQL databases. Op dit moment gebruiken vooral de grotere websites de volledige functies van NoSQL mogelijkheden. Bijvoorbeeld: Facebook gebruikt duizenden servers met daarop Cassandra. Bol.com gebruikt Hadoop om dagelijks de recommendations te bepalen.

In het landschap van NoSQL zijn verschillende opslagsystemen beschikbaar met elk hun eigenschappen. Deze eigenschappen maken deze opslagsystemen geschikt voor een type probleem. Hieronder zullen we een aantal van deze systemen bespreken.

Key-Value Opslagsystemen (Key-Value Store)



In een Key-Value opslagsysteem heeft de data geen vast formaat. De data (value) wordt benaderd met behulp van een string (key). Deze systemen beschikken over een zeer eenvoudige interface om met de opgeslagen informatie te werken. Het datamodel is niet meer dan een (key, value) pair met als basisbewerkingen: insert(key,value), fetch(key), update(key) en delete(key). De waarden worden opgeslagen als een BLOB (Binairy Large Object). De Key-Value store heeft geen kennis van de inhoud van de opgeslagen waarde. De applicaties die de Key-Value store gebruiken zijn verantwoordelijk voor de inhoud van de waarde. Een Key-Value store kan gezien worden als een tabel met één kolom met de primaire sleutel en één kolom met de waarde.

De implementatie is efficient, schaalbaar en fault-tolerance. De records worden op basis van hun sleutel gedistribueerd over verschillende nodes (servers).

Voorbeelden van Key-Value opslagsystemen zijn: Amazon Dynamo, ...

Document-based Opslagsystemen (Document Stores)

Een document opslagsysteem lijkt in veel opzichten op een Key-Value store waarbij de value een document is. Het formaat van het document is vaak in JSON (JavaScript Object Notation), BSON (Binary JSON), XML of een ander semi-gestructureerd formaat. Naast de basisbewerkingen zoals omschreven bij de Key-Value store is het ook mogelijk values op te ophalen (fetch) op basis van de inhoud van het document.

In tegenstelling tot de Key-Value store kan de value in een document store weer meerdere 'name/value pairs' bevatten. Eén enkele kolom kan dus weer honderden van deze waarden bevatten die per rij weer kunnen verschillen van inhoud. De records kunnen verschillende structuren bevatten.

De waarden in deze kolom zijn doorzoekbaar in Document stores. De structuur van een document kan 'on the fly' aangepast worden door het toevoegen of wijzigen van attributen van het document.

Een voorbeeld van een JSON document is:

```
{
   "_id" : ObjectId("939abc49de39e09cd38ba5c8"),
   "voornaam" : "....",
   "achternaam" : "....",
   "adres" : {
              "straat" : "...",
              "plaats" : "...",
              "postcode" : "..."
    }
}
```

Document opslagsystemen zijn goed in het opslaan van onregelmatige (semi-gestructureerde) data. In een RDBMS zou deze data leiden tot een omvangrijk aantal nul-waarden in de records van de database tabellen. Document stores worden veelal toegepast in Big Data-omvang verzamelingen van tekstdocumenten, emailberichten, XML documents, maar ook gedenormaliseerde (geaggregeerde) data.
Voorbeelden van Document opslagsystemen zijn: CouchDB, MongoDB, en SimpleDB.

Column-based opslagsystemen (Column stores)

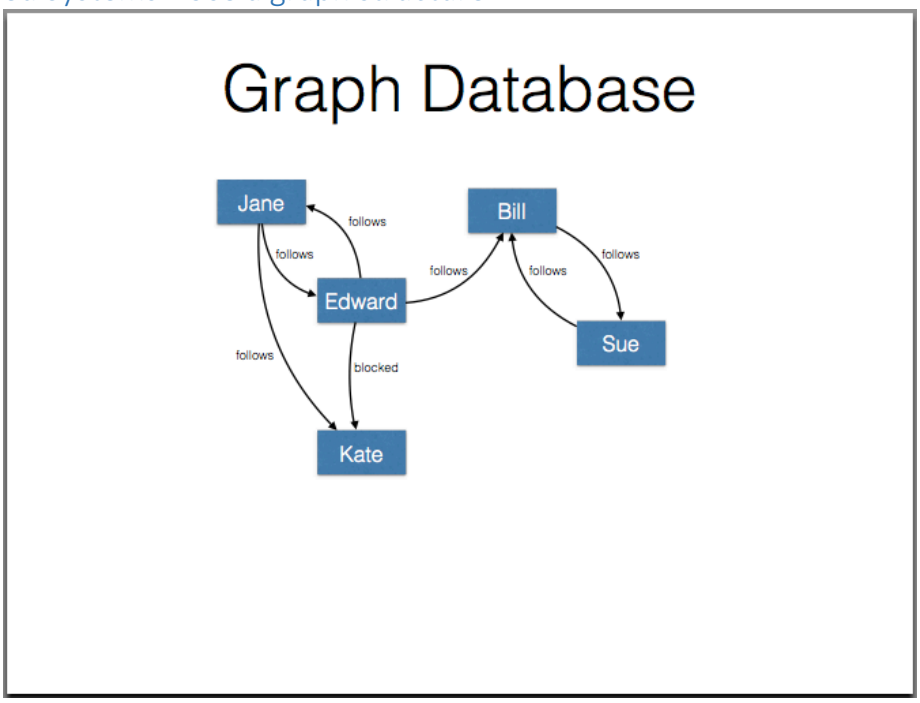
Pionier van de Column-based opslagsystemen is Google's BigTable. Essentieel verschil van Big Table en andere column-based systemen met normale relationele databases is dat een record met gegevens niet opgeslagen wordt als een rij in een tabel maar als een kolom.

Op het eerste gezicht lijkt dit misschien weinig verschil te maken, maar als je weet dat een database gegevens rij voor rij op disk opslaat en weer wegschrijft maakt dit een groot verschil. Stel dat je zoekt in een gigantisch bestand met scooter-rijders naar Willem Holleeder. Als van elke scooter-rijder honderden gegevens bekend zijn (die dus in een rij met honderden velden zijn opgeslagen), worden er tijdens het zoeken talloze gegevens onnodig in het geheugen geladen. Als dezelfde data in een Column-based systeem zou zijn opgeslagen, geeft de eerste read meteen alle namen van de personen in de database, en kun je meteen uitvinden of Willem Holleeder daar tussen staat.

De kolom is de kleinste eenheid van data en is in feite een tuple met naam, waarde en timestamp. De timestamp wordt gebruikt om de validiteit van de waarden te bepalen. Gedistribueerde opslagsystemen kunnen geen consistency garanderen, immers beschikbaarheid is belangrijker (CAP-theorie). De store gebruikt timestamps om vast te stellen welke data in de opgeslagen nodes (servers) up-to-date zijn. Dit type opslagsystemen zijn goed in:

- Gedistribueerde opslag van data en in het bijzonder data met versies (timestamps).
- Grootschalig, batch-georiënteerde data processing: conversie, parsen, sorteren, complexe algoritmische bewerkingen, etc.
- Business Intelligence (exploratieve en predictieve analyse)

Graph-based Systems - Use a graph structure



Een graph database is een database die gebruik maakt van graafstructuren (grafentheorie) met knooppunten (nodes), ribben (edges) en eigenschappen (properties) om data op te slaan en te representeren. Alle data in een graph database bevat een directe verwijzing naar aangrenzende data, waardoor geen indexen nodig zijn om relaties te bepalen (index-free adjacency). Elk opslagsysteem dat index-free adjacency biedt is in feite een graph database.

In het algemeen zijn graph database *zeer* geschikt voor problemen waar de relatie tussen data belangrijker is dan de data zelf. Denk hierbij aan het doorlopen van complexe (sociale) netwerken, genereren van afhankelijkheden, patroonherkenning, forensisch onderzoek.

Omdat de data anders gestructureerd zijn, werkt ook SQL als standaard taal om zoekopdrachten te formuleren, niet meer. Er worden daarom alternatieve zoektalen ontwikkeld. De Graph Database Neo4J werkt bijvoorbeeld met zogenaamde Cypher queries.

Een voorbeeld van een Cypher query is:

```
MATCH (Jane) –[:follows]->()->[:follows]->(fof)
RETURN fof
```

Met deze opdracht kunnen de ‘friends of friends’ van ‘Jane’ gevonden worden.

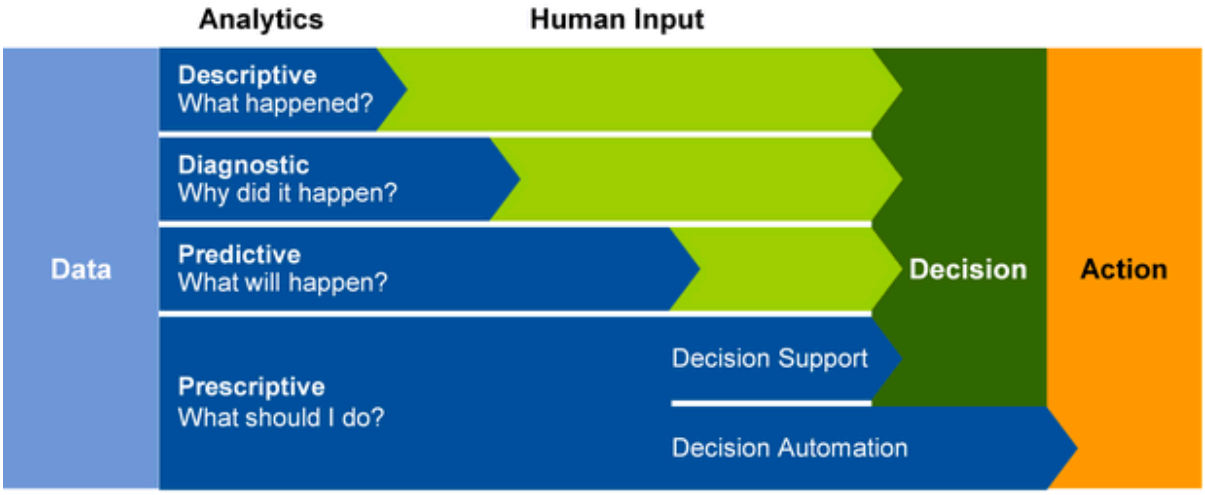
Voorbeelden van graph database systemen zijn: Pregel en Neo4J.

De nadelen van NO-SQL systemen

Er bestaat geen gestandaardiseerde universele taal voor deze systemen zoals SQL voor relationele system. Elk NO-SQL product werkt op een andere wijze. Relationele database zijn volwassen, dit in tegenstelling tot NO-SQL wat pas sinds een jaar of zes beschikbaar is. Daarnaast zijn relationele systemen onderdeel van een groot “ecosysteem” en zijn vele tools beschikbaar.

Big Data Analyse Methoden

Big Data Analytics betreft methoden voor het analyseren en verwerken van grote hoeveelheden en verschillende typen data om informatie en kennis te verwerven uit deze data. Het idee is dat mensen op basis van deze informatie en kennis beter beslissingen kunnen nemen (fact-based, evidence-based), en daardoor effectievere acties kunnen uitzetten.



Bron figuur: [Gartner 2013]

In bovenstaande figuur worden vier soorten Analytics onderscheiden op basis van het doel dat nagestreefd wordt:

Beschrijvend (Descriptive)	"Wat is er precies gebeurd?" Het doel van *Beschrijvende Analytics* is een goed beeld van de werkelijke situatie te geven, zodat een mens op basis hiervan een goed besluit kan nemen om bepaalde verbeteracties uit te zetten.
Verklarend (Diagnostic)	"Waarom is dat gebeurd?" Het doel van *Verklarende Analytics* is een verklaring te geven (of meerdere mogelijke verklaringen) van de oorzaak van een ontstane situatie. Een mens kan op basis van deze inzichten vaak een beter besluit nemen en vervolgacties definiëren, dan alleen op basis van *Beschrijvende Analytics*.
Voorspellend (Predictive)	"Wat zullen gevolgen van een bepaalde actie zijn?" Het doel van *Voorspellende Analytics* is verschillende oplossings-scenario's te kunnen doorrekenen (What if …?"), voordat een besluit wordt genomen. Een mens kan op basis van deze exercitie nog betere besluiten nemen, omdat de gevolgen van een eventuele ingreep meegenomen kunnen worden.
Voorschrijvend (Prescriptive)	"Wat kan ik het best doen?" Het doel van *Voorschrijvende Analytics* is een advies te geven over de beste oplossing voor een bepaald probleem. Er zijn twee typen Voorschrijvende Analytics tools te onderscheiden: (1) Decision Support Systemen en (2) Automatische Beslissings Systemen. Bij Decision Support Systemen is er nog steeds een mens betrokken, die geadviseerd door het systeem een besluit neemt. Bij de tweede soort systemen speelt het menselijke beoordelingsvermogen geen rol

meer. (Alle relevante kennis is gecodeerd in algoritmen. Dit gebeurt bijvoorbeeld bij Algo-trading systemen in de aandelenhandel.)

In zekere zin is ook het doel van data mining om kennis te extraheren uit grote hoeveelheden data. Alleen wordt bij data mining hoofdzakelijk gestructureerde data geanalyseerd. Voor de analyse wordt veelal gebruikt gemaakt van technieken uit de statistiek, kunstmatige intelligentie en machine learning. In de loop der jaren was er ook een behoefte om kennis te extraheren uit ongestructureerde data. Dit heeft geleid tot het gebied van de tekst mining. Om uit teksten bruikbare kennis te extraheren wordt veelal gebruikt gemaakt van data mining technieken in combinatie met technieken uit de natuurlijke taalverwerking. Bij natuurlijke taalverwerking worden methoden en technieken ontwikkeld voor de analyse van verschijnselen in syntaxis en semantiek van een natuurlijke taal.

Omdat big data zich op het snijvlak van real-time, multimedia en conventionele large databases bevindt, is het niet zo verwonderlijk dat bij het analyseren van big data gebruik wordt gemaakt van methoden en technieken uit deze gebieden. In deze paragraaf beschrijven we in het kort een aantal bekende technieken die bij de verschillende typen database gebruikt worden.

Bij **real-time databases** worden grote hoeveelheden gestructureerde data gegenereerd die in korte tijd verwerkt dienen te worden. Data gegenereerd door sensoren vallen hieronder. Bijvoorbeeld om de route van een vliegtuig te tracken worden data uit verschillende sensoren verzameld en verwerkt teneinde de positie van een vliegtuig te kunnen bepalen. Hiertoe worden technieken uit multi-sensor data fusie gebruikt. Stel dat we twee metingen hebben. In het algemeen zijn dit twee verschillende waarden en ligt de "waarheid" in het midden. De taak van het fusie algoritme is om dit midden zo goed mogelijk te kiezen. Hierbij maakt een goed algoritme gebruik van een betrouwbaarheidsschatting van beide waardes; de meest betrouwbare waarde krijgt de zwaarste weging in een gewogen gemiddelde. Naast metingen wordt ook gebruik gemaakt van al bekende kennis over een fenomeen om data optimaal te fuseren. Bekende algoritmen voor data fusie zijn Kalman filters, Bayesiaanse netwerken en technieken gebaseerd op de Demster-Shafer theorie. Een van de mogelijkheden die Big Data biedt is het uitbuiten van informatie over de posities van objecten. Nu de positie van een voertuig nauwkeurig is te bepalen met behulp van multi-sensor data fusie, kunnen er nieuwe typen vragen worden beantwoord met conventionele database technieken. Stel dat u zich in de auto bevindt en het systeem in uw auto weet dat u een sterke voorkeur heeft voor La Place restaurants. Als u nu kenbaar maakt dat u honger heeft dan kan het systeem voor u uitrekenen wat de dichtstbijzijnde restaurants zijn en de dichtstbijzijnde La Place. Voort kan het systeem u naar het gewenste restaurant navigeren.

Bij **multimedia technologie** is het doel om verschillende vormen van elektronische gegevens, zoals tekst, beeld en geluid, geïntegreerd te verwerken en te presenteren. Naast technieken om de gegevens op te slaan, worden verschillende bestaande technieken uit het veld van patroonherkenning, data mining en statistiek op maat gesneden om vragen betrekking hebbend op de data, te beantwoorden. Bijvoorbeeld gevraagd zou kunnen worden om alle doelpunten van de wereldkampioenschappen voetbal 2014 te selecteren uit een verzameling van beelden. Om een dergelijke vraag te beantwoorden kan het begrip doelpunt mathematisch gedefinieerd worden maar ook een selectie gemaakt worden van alle beeldfragmenten waar er veel gejuicht wordt. Multi-media technieken bestaan uit een verzameling van bekende clustering en classificatietechnieken die op maat zijn gesneden. Bekende clustering- en classificatietechnieken zijn beslisbomen, regressie, neurale netwerken, genetische algoritmen, enz.

Bovenstaande type databases en technieken hebben voortgeborduurd op de theorieën en technieken van de conventionele large databases. Nadat de opslag en bevragingen van

gestructureerde data goed begrepen waren is de behoefte ontstaan om naar kennis te speuren in de grote hoeveelheden data die in databases waren opgeslagen. Om dit systematisch te doen, zijn ruwweg vier stappen te onderscheiden bij mining van de data. Bij de eerste stap dient men vast te stellen naar welke soort kennis men op zoek is. Bij de tweede stap moeten de data die her en der aanwezig zijn en een rol kunnen spelen bij het opbouwen van de gezochte kennis, bij elkaar worden gebracht; het liefst in enkele data warehouse. Tijdens de derde stap wordt naar nuttige kennis gezocht in de data met behulp van een verzameling algoritmen. Tenslotte, wordt in de laatste stap, validatiestap, vastgesteld of de gevonden kennis in overeenstemming is met de werkelijkheid. Voor het minen van de data in databases zijn een groot aantal technieken uit de statistiek, machine learning en kunstmatige intelligentie op maat gesneden voor mining doeleinden, zoals regressietechnieken, kernel density functies, beslisbomen, neural netwerken, hill climbing, simulated annealing enz.

Het is de verwachting dat Big Data Analytics de bestaande verzameling van technieken uit real-time technologie, data mining, natuurlijke taalverwerking en multimedia technologie verder zal avanceren en integreren. Deze ontwikkeling zal er voor zorgen dat er vragen kunnen worden beantwoord en toepassingen ontstaan die voorheen niet mogelijk waren.

Hieronder zetten de besproken Analystics methoden nogmaals op een rij:

- Summary Statistics: maten voor midden en spreiding, histogrammen, scatter diagrammen, maten voor samenhang twee datasets (correllation)
- Optimization Models: Operations Research / Linear Programming (Simplex Method), Probabilistic models… (Monte Carlo), ...
- Segmentation, K-Means Clustering (“bepalen van verkoopbevorderende marktsegmentatie uit verkoopgegevens”)
- Prediction Models, Forecasting: Regression models, (Training of a Model, Machine Learning), Feed current parameters in model, Naive Bayes (bag of words method) “sentiment analysis of tweets”, Ensemble models, Recommendations: “People who ordered this also bought …”
- Outlier Detection (“filteren Credit Card transacties op mogelijke fraude gevallen”)
- Visualisation, Network Graphs: (“ontdekken patronen door visuele inspectie datasets”)

Nogmaals, dit zijn geen fundamenteel nieuwe methoden. Wel zijn voor deze methoden nieuwe algoritmen bedacht, die geschikt zijn voor parallel processing op clusters van computers. In de volgende paragraaf zetten we uiteen hoe een Big Data infrastructuur is opgebouwd, en wat dit voor eisen stelt aan de vorm van de algoritmen

Big Data infrastructuur

De IT-infrastructuur, die je nodig hebt om Big Data analyses te kunnen maken, bestaat grofweg uit drie delen:

1. Data opslag (storage)
2. Data verwerking (processing)
3. Analytics tools (inclusief visualisatie)

Traditioneel bestond de ‘Data opslag’ van een Data Warehouse voornamelijk uit relationele databases, waaruit met SQL queries gegevens konden worden gehaald. Bij Big Data bestaat de ‘Data opslag’ uit meerdere soorten opslagsystemen. Naast relationele (SQL) databases zijn er ook NO-SQL databases Key-Value Stores, Document Databases en Graph Databases.

Bij Big Data is de verwerking van data revolutionair anders dan bij traditionele DataWareHouses. De hoeveelheid gegevens en verschillende datatypes zorgen ervoor dat bestaande methoden en tools

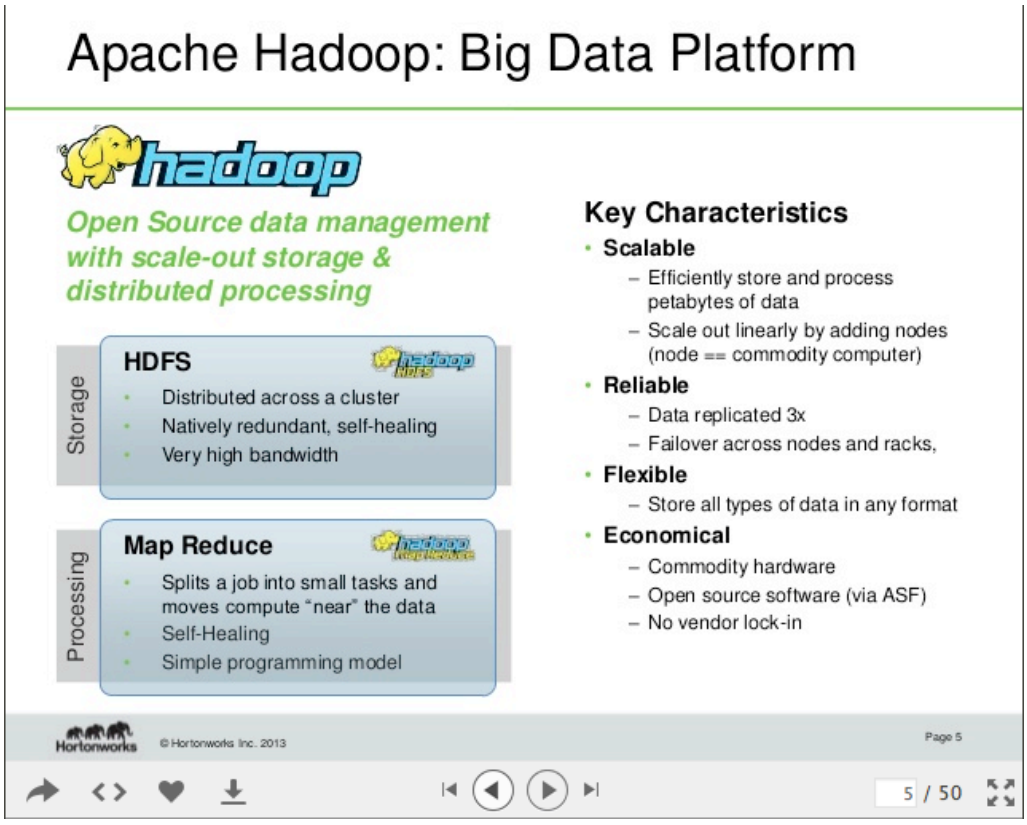
niet meer voldoen. Bij Internetbedrijven als Yahoo, Google en Facebook is gepionierd met nieuwe methoden. Uiteindelijk is hieruit Open Source Project **Hadoop** voortgekomen: een infrastructuur voor het massaal parallel verwerken van data.

Voor de Analytics laag wordt bij Big Data nog steeds zoveel mogelijk gebruik gemaakt van de vertrouwde Reporting tools en Data Mining tools van Data Warehouse oplossingen. Dit zijn tools waarmee Data Scientists en Business Analysts vertrouwd zijn. Belangrijk verschil is wel dat regelmatig verschijnende rapporten over een afgelopen periode, vaak vervangen worden door real-time dashboard met de actuele stand van zaken.

De grote vernieuwing zit dus in de manier waarop dat verwerkt worden, en Hadoop is hierin richtingbepalend geweest. Vandaar dat we in deze paragraaf eerst wat dieper ingaan op Hadoop.

De basis van Hadoop is **HDFS**: Hadoop Distributed File System. Dit is technologie om grote hoeveelheden data verspreid over een groot aantal computers (*nodes* in Hadoop terminologie) op te slaan. De essentie van Hadoop is dat niet data uit een database naar de computer getransporteerd wordt, maar dat *computing* naar de computers gaat waar data al aanwezig is.

Op HDFS draait de **Map-Reduce** laag. Map-Reduce is uitgevonden bij Google en voor het eerst beschreven in [Dean e.a. 2004]. In feite is het standaard infrastructuur software die zorgt dat de computing klus wordt verdeeld over de beschikbare nodes, op zo'n manier dat elke node zonder met andere nodes informatie uit te wisselen, geheel zelfstandig een deel van het werk kan doen.



Bron illustratie: [Hortonworks 2013]

Dit opdelen van de klus in parallelle werkzaamheden wordt geprogrammeerd in de **Mapper**. In praktijk is dezelfde data vaak op minstens 3 nodes aanwezig, en wordt de Mapper klus dus minstens drie maal uitgevoerd. Dit maakt Hadoop bestand tegen het uitvallen of vastlopen van een enkele node. (Stel dat je een cluster hebt met 1000 nodes, waarvan voor elke node de *mean time between failures* drie jaar is. Dat wil zeggen dat je mag verwachten dat elke computer circa drie jaar zonder problemen werkt voordat ie een keer hapert. Dan is de verwachting dat er vandaag één van de nodes hapert: 1000*(1/(3*365)) en dat is bijna gelijk aan 1. Er gaat dus vrijwel zeker elke dag één node stuk!)

Als een mapper klus klaar is, wordt het resultaat gesorteerd en daarna opgestuurd naar de **Reducer.** De Reducer verzamelt de gesorteerde tussenresultaten en berekent het eindresultaat.

De *flow of execution* in Hadoop lijkt op lopende band werk, waarbij een aantal stadia altijd hetzelfde zijn (en je dus niet meer hoeft te programmeren) en de Mapper en de Reducer altijd maatwerk blijven, afhankelijk van de klus.

Intermezzo: Uitgewerkt Map-Reduce Voorbeeld	
Doel:	Stel je wilt zo snel mogelijk weten welke woorden voorkomen in *Het Wetboek van Strafrecht,* en welke woorden het vaakst voorkomen. Je hebt geen computers tot je beschikking maar wel een enorme groep nauwkeurig werkende ambtenaren.
Input:	Drie exemplaren van Het Wetboek van Strafrecht 
Mappers:	De opdrachtgever scheurt alle pagina's uit de drie wetboeken. Elke ambtenaar krijgt één wetboek pagina, en een stapeltje blanco A4-tjes. De opdracht luidt: Lees de wetboek pagina woord voor woord. Als je een nieuw woord tegenkomt, pak dan een nieuw A4-tje. Schrijf het woord bovenaan het A4-tje, en schrijf het pagina nummer op dit A4-tje. Als je een woord tegenkomt dat je al eerder hebt verwerkt, schrijf het pagina nummer dan bij op het bestaande A4-tje. Als je klaar bent, lever je A4-tjes dan in bij de Sorter. De ambtenaar die pagina 247-248 verwerkt produceert dus net zoveel A4-tjes als er verschillende woorden op zijn pagina staan. Op het elk A4-tje staat, dus een woord en een lijst cijfers, bijvoorbeeld: 'Vatbaar': 247, 247, 247, 247, 248, 248 'voor': 247, 247, 247, 247, 247, 247, 248, 248, 248, 248, 248, 248, 248, 248, 248, 248 'verbeurdverklaring': 247, 247, 247, 247, 248, 248

Sorter (Shuffle):	De sorter neemt de A4-tjes in ontvangst, en maakt stapeltjes op alfabetische volgorde van de woorden die bovenaan de A4-tjes staan. A4-tjes met hetzelfde woord komen op één stapeltje te liggen Elk stapeltje A4-tjes met eenzelfde woord wordt vervolgens doorgegeven aan een Reducer.
Reducers:	De reducer gaat het aantal paginanummer dat op de A4-tjes is vermeld tellen. De reducer van 'verbeurdverklaring' komt bijvoorbeeld uit op 178 keer. De Reducer schrijft dit aantal op een kaartje, met daarachter het woord. Dus: 178: 'verbeurdverklaring' en geeft dit door aan de opdrachtgever.
Opdrachtgever:	De opdrachtgever neemt de kaartjes van de Reducers in ontvangst, en sorteert ze op volgorde van grootte van hoog naar laag. Op de bovenste kaartjes staan de meestvoorkomende woorden.
	Opdracht voltooid!

Eigenlijk kun je deze basis van Hadoop alleen gebruiken als je kunt programmeren in een taal als Java, Perl of Python. Omdat veel data analysten dit niet kunnen of willen, zijn er rond Hadoop een aantal modules gemaakt zodat Hadoop aangestuurd kan worden vanuit voor data analysten vertrouwde omgevingen.

Deze modules (en/of bijbehorende talen) zijn:

- HBASE een gedistribueerde NO-SQL database die op HDFS draait
- PIG een eenvoudige scripting taal om MapReduce jobs te specificeren (woordgrapje: Pig Latin betekent Potjeslatijn).
- HIVE een *Data Warehouse* view op Hadoop, dat gewoon in SQL kan worden aangestuurd.
- FLUME een software framework om Hadoop met data te vullen
- SQOOP een connectivity tool om data van externe databases en/of data warehouses van en naar Hadoop te transporteren
- MAHOUT een Data Mining library
- ETL, Extract Transform Load tools

Eigen ervaringen

Bij onze experimenten hebben we in eerste instantie algoritmen getest op een *single node Hadoop cluster* op een laptop van één van onze onderzoekers. Als het algoritme oplevert wat we willen, kan het zonder enige wijziging gedraaid worden op een veel groter Hadoop cluster als bijvoorbeeld Google of Amazon leveren (in de Cloud).

Het is natuurlijk ook mogelijk om zelf een Hadoop cluster te bouwen. Onderstaande figuur toont hoe de hardware infrastructuur in principe kan worden opgebouwd. Als vertrouwelijkheid van gegevens belangrijk is, zal deze hardware achter een firewall staan zodat er wel (real-time) gegevens van Internet kunnen worden gehaald, maar de omgeving niet vanaf Internet te benaderen is.

Alternatief voor Hadoop: Splunk

Naast de gratis Open Source software van Hadoop is **Splunk** een grote speler voor Big Data toepassingen. Hoewel Splunk een proprietary en behoorlijk duur product is, wordt het toch massaal gebruikt in het bedrijfsleven. Een verklaring zou kunnen zijn dat bedrijven liever betalen voor een

product waarmee ze meteen aan de slag kunnen, en waarvoor support geleverd wordt, dan zelf te gaan pionieren met Open Source. Bij universiteiten is uiteraard Hadoop populairder (weinig geld, tijd genoeg).

Doordat Hadoop nu ook ondersteund wordt door IBM en gespecialiseerde bedrijven als Hortonworks, verwachten we dat Hadoop ook bij bedrijven meer en meer zal worden toegepast.

Splunk was oorspronkelijk een oplossing voor data mining uit log files. Met Splunk kun je monitors bouwen, computer log files doorzoeken en het bevat tools voor het analyseren en visualiseren van deze logfiles.

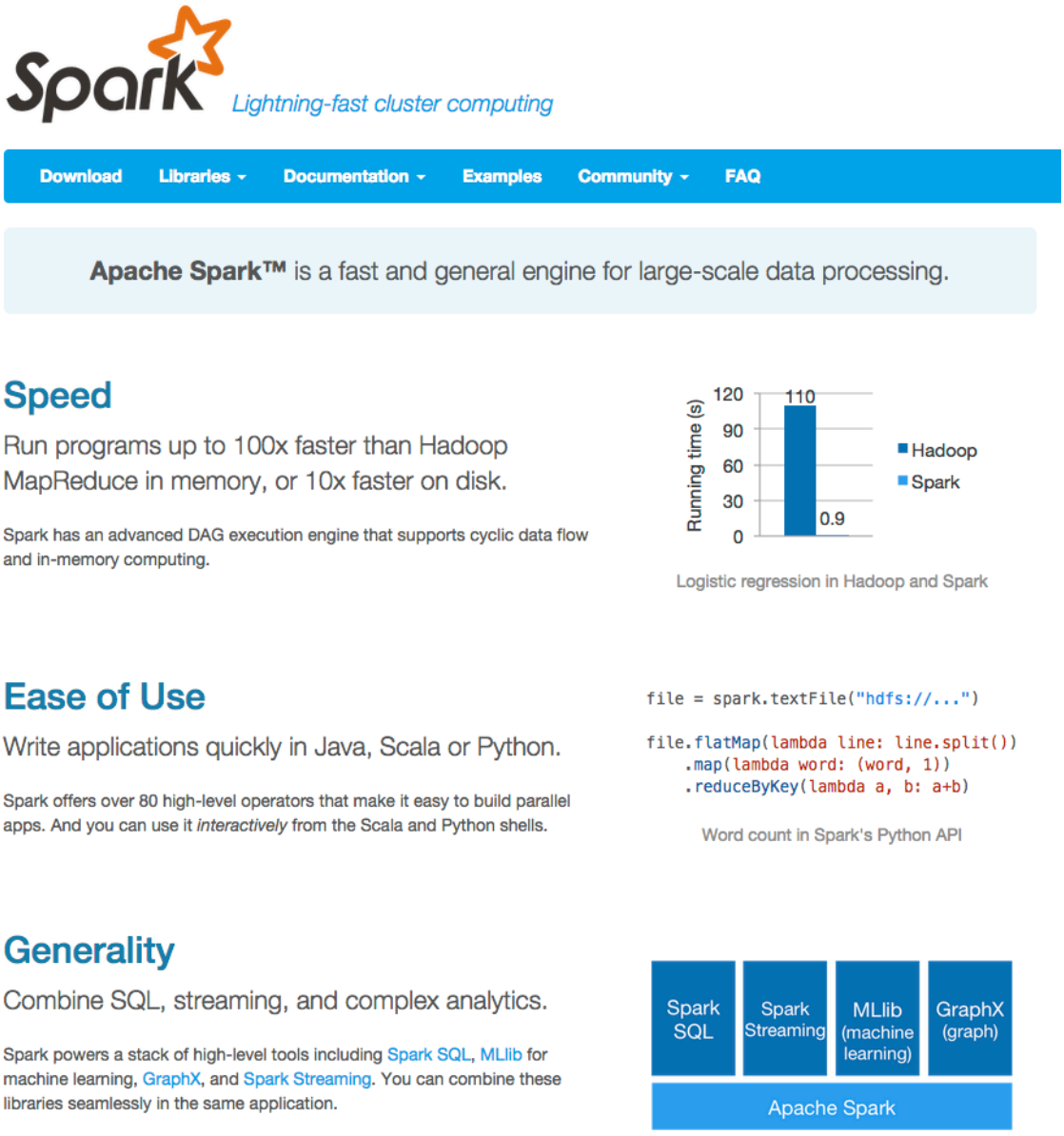
Van Splunk is ook een versie verschenen die 'in the cloud' draait: **Splunk Storm** genaamd. Recent is er ook een versie verschenen van Splunk die op de Hadoop infrastructuur draait: **Hunk** genaamd.

Het bedrijf achter Splunk had in 2013 circa USD 200 miljoen omzet, maar draait nog steeds aanloopverliezen.

Recente ontwikkelingen: Spark

Nu Hadoop redelijk uitontwikkeld is, en in praktijk ingepast wordt in Big Data Analytics oplossingen van grote leveranciers als IBM, Oracle, SAP en Microstrategy is er in de Open Source community alweer een nieuwe veelbelovende ontwikkeling gestart: Spark!

Spark maakt gebruik van dezelfde HDFS gedistribueerde opslag en verwerking van gegevens als Hadoop. Het heeft echter een revolutionair nieuwe aanpak voor de Map-Reduce laag. Bij Spark worden er nauwelijks meer gegeven van disk gelezen of naar disk weggeschreven, maar vindt vrijwel alle processing *in-memory* plaats. Hierdoor wordt een enorme snelheidsverbetering gehaald. Spark Map-Reduce is circa 100 keer sneller dan traditionele Hadoop Map-Reduce.



Bron illustratie: spark.apache.org

Samenvatting

Er zijn afgelopen vijf jaar revolutionaire ontwikkelingen geweest in (1) de hoeveelheid en varieteit van beschikbare data en (2) methoden om NO-SQL data op te slaan en data massaal parallel te verwerken. Voor beslissers bij bedrijven of instellingen en de Data Scientists die voor hen werken hebben deze ontwikkelingen veel nieuwe toepassingen mogelijk gemaakt.

Wat er veranderd is voor een Data Scientist is dat:

- Traditioneel beschikte een Data Scientist over alle, netjes in tabellen geordende data. Bij Big Data zijn de gegevens vaak incompleet en 'messy' (rommelig).

- Traditioneel had een Data Scientist de Data Sets goed onder controle. Bij Big Data is het beheren van de Data Sets door omvang, real-time karakter en omdat ze overal vandaan

kunnen komen een uitdaging op zich.

- Traditioneel beantwoordden Data Scientists vooraf gestelde vragen met behulp van de gegevens. Bij Big Data worden patronen gezocht: wat vertelt de data ons?

- Traditioneel werden de rapportages gebruikt om prestaties in het verleden te meten. Bij Big Data worden bedrijfsprocessen vaak real-time bestuurd door de analyses.

Voor een beslisser bij een bedrijf of instelling is veranderd, dat er veel meer analyses gedaan kunnen worden, waardoor in principe betere besluiten kunnen worden genomen. In sommige gevallen kan de besluitvorming beter geheel overgelaten worden aan het systeem, en kan de beslisser zich laten omscholen tot Data Scientist.

Ontwikkelingen bij Ministerie van Veiligheid & Justitie

Het Wetenschappelijk Onderzoeks- en Documentatie-Centrum (WODC) van het Ministerie van Veiligheid & Justitie heeft ruime ervaring in het verwerken van gegevens uit diverse bronnen bij Politie en Justitie tot overzichten waarin beleidmakers kunnen zien wat het effect is van beleidswijzigingen.

In [Choenni 2011] wordt bijvoorbeeld gerapporteerd over een DataWareHouse dat gebouwd is om inzicht te krijgen hoe misdadigers door de aangifte-, opsporings-, berechtings en straf-processen lopen. Er worden gegevens geput uit operationele systemen voor opsporing bij politieregio's (HKS), voor vervolging door het Openbaar Ministerie (COMPAS), voor berechting door de rechtbanken (JDS) en voor strafoplegging door Justitiële Inrichtingen (TULP).

Doordat deze systemen volstrekt gescheiden zijn, hebben ze allemaal hun eigen wijze van persoonsregistratie. De persoonsgegevens uit de verschillende bronnen worden gekoppeld door algoritmes, die met grote waarschijnlijkheid bepalen dat het om dezelfde persoon gaat. Er wordt hierbij een nauwkeurigheid van 93% gehaald. In 7% van de gevallen gaat het dus niet goed, en wordt bijvoorbeeld een veroordeling van een vader gecombineerd met een celstraf van zijn zoon. Dit illustreert dat ook al bij DataWareHousing de *Veracity*, het waarheidsgehalte van gegevens en analyses eigenlijk expliciet zou moeten worden meegenomen!

De architectuur van het gebouwde DataWareHouse bestaat uit drie lagen:

1. Data op orde (DBMS'en)
2. Data Analytics (algoritmen)
3. Duiding (visualisatie)

Bij het bouwen van het DataWareHouse is expliciet gekozen voor een dader-gecentreerde aanpak. De indeling van tabellen in databases in laag 1 is toegesneden op deze keuze. Het is daarom niet zo eenvoudig om om basis van dezelfde gegevens analyses te doen over bezetting van cellen in diverse Penitiaire Inrichtingen, of vergelijking van efficiëntie van rechtbanken in verschillende regio's. Een DataWareHouse bouwen is veel werk, waarbij keuzes gemaakt worden waardoor flexibiliteit van soorten analyses afneemt.

In [de Vries ??] wordt gerapporteerd over een systeem dat gebouwd is om informatie uit alle politieregio's over meerdere jaren te kunnen vergelijken, in het perspectief van algemene demografische ontwikkelingen in Nederland. De basis-informatie is zowel afkomstig uit politie-databases met aangiftes en opsporings-gegevens, als uit antwoorden op surveys die in politieregio's zijn gehouden.

Vaak krijgen gegevens krijgen pas betekenis als je ze kunt vergelijken: waarom nemen inbraken in de ene regio toe, terwijl ze in de andere regio's afnemen? Stijgende misdaadcijfers zijn niet alarmerend als de bevolking nog harder stijgt. Focus bij dit project was dan ook om een dashboard te maken, waarmee beleidsmakers interactief overzichten en vergelijkingen konden maken tussen verschillende geografische regio's. Er worden daarom ook gegevens van CBS gebruikt, om ontwikkelingen in misdaadcijfers te relateren aan algemene demografische ontwikkelingen.

Ook bij dit project is DataWareHouse technologie toegepast, in dit geval met drie lagen:

1. Scripts om gegevens uit operationele databases te halen.
2. Data op orde (DBMS'en met een apart systeem met meta-data, gegevens over de verzamelde gegevens)

3. Dashboard (duiding, met interactie mogelijkheden).

Er is dus al behoorlijk wat kennis en ervaring opgebouwd bij het WODC met het bouwen van *Decision Support Systems* voor beleidsmakers en beslissers.

Tot nog toe zijn echter:

1. De gegevensbronnen beperkt gebleven tot gegevens van politie en justitie zelf, danwel betrouwbare gegevens van het CBS.
2. De verwerkingssnelheid laag, hetgeen acceptabel is omdat vooral gewerkt wordt met historische gegevens.
3. De systemen zijn toegesneden op een specifieke toepassing: volgen van daders door het systeem of analyseren van regionale verschillen.

Het zou interessant zijn om ook ervaring op te gaan doen met Big Data technologie. Bijvoorbeeld om meer real-time gegevens te verwerken: Wat is de relatie tussen inbraken en het weer? Wat gaat er mis bij grootschalige evenementen, en kunnen we kans op misdaad verkleinen door crowd management of extra eisen te stellen voor het afgeven van vergunningen?

Verder is het interessant om ook gegevens uit andere, wellicht minder betrouwbare bronnen te gebruiken. In mondeling overleg hebben we begrepen dat het ministerie een abonnement heeft bij Coostoo om Social Media uitingen te kunnen volgen. Dat is een uitstekend begin, maar welke inzichten kunnen verkregen worden als we die gegevens kunnen combineren met misdaadcijfers? Welke mogelijkheden onstaan als locatie-gegevens van TomTom of telecom providers gebruikt worden, of gegevens over verdachte transacties van banken?

Tenslotte zou Big Data technologie nieuwe inzichten kunnen verschaffen, onverwachte patronen kunnen detecteren. In theorie kan dat ook al met behulp van Data Mining technieken in een DataWareHouse omgeving, maar in praktijk zijn de gevens dan al zo geselecteerd en voorgestructureerd dat er weinig nieuwe inzichten uitkomen. Met Big Data technologie kun je vrijer en sneller werken, ook met minder gestructureerde gegevens. Dit biedt mogelijkheden om snelle analyses te doen naar bijvoorbeeld: Welke activiteiten van gevangenen tijdens hun straf hebben invloed op kans op recidive? Of: Welke werkwijze zorgt voor sneller doorlopen van de strafrecht-keten?

Conclusies en aanbevelingen

De volgende vragen vormden het startpunt voor dit verkennend onderzoek:

1. Welke type big data en welke daarbij passende databronnen zijn relevant en bruikbaar voor het werkterrein van het Ministerie van Veiligheid & Justitie?

2. Welke infrastructuur, tools en skills die bijvoorbeeld worden toegepast en relevant zijn binnen wetenschappelijk onderzoek en binnen big data onderzoek in het bedrijfsleven zijn mogelijk relevant voor de toepassing binnen het werkterrein van het Ministerie van Veiligheid en Justitie?

3. Wat is het belang van 'klassieke' begrippen als validiteit en betrouwbaarheid in de context van big data analyses en toepassingen? Gelden ze op dezelfde wijze als in de bestaande onderzoekspraktijk of krijgen ze binnen big data een andere betekenis en invulling?

4. Hoe zouden specifieke toepassingen van de mogelijkheden van big data voor het Ministerie van Veiligheid & Justitie er uit kunnen zien?

5. Is het reel om bestaande vormen van onderzoek, waaronder bestaand monitoring onderzoek van het Ministerie van Veiligheid & Justitie te vervangen door big data toepassingen of zijn ze eerder complementair aan de bestaande vormen en praktijk?

6. Wat kan het Ministerie van Veiligheid & Justitie op de korte termijn doen om werk te maken van toepassing van big data toepassingen, waaronder real-time analytics?

Relevante Databronnen

In het algemeen kunnen we de volgende databronnen onderscheiden:

Transactions
Log data
Events
Emails
Social Media
Sensors
External Feeds
RFID scans or Point Of Sale data
Free-form text
Geospatial data
Audio
Still Images or Video

Groot verschil met het tijdperk van voor Big Data is:

- dat lang niet alle data meer uit eigen systemen komen, en de kwaliteit van de data daardoor wisselend zal zijn
- dat veel data niet meer in een traditionele relationele (SQL) database past, en daardoor alternatieve opslagsystemen toegepast moeten worden

- dat er heel veel meer data is, en in sommige gevallen data niet meer opgeslagen kan worden maar binnenstroomt en binnen luttele seconden haar waarde verliest.

Meer specifiek voor het beleidsterrein Veiligheid & Justitie is ons advies niet alleen te kijken naar nieuwe databronnen maar ook naar bestaande aangiftes en in het verleden in beslag genomen administraties. Met de nieuwe methoden kunnen deze traditionele databronnen wellicht een schat aan informatie opleveren.

Infrastructuur en Opslagsystemen

De technologie voor gegevensopslag en gegevensverwerking is nog volop in ontwikkeling. Het is inmiddels wel duidelijk dat alleen traditionele SQL-databases niet meer volstaan. Het is sowieso nodig om voor bepaalde type gegevens key-value stores, document-databases of graph-databases te benutten. Voordeel van deze databases is dat de opslagstructuur beter aansluit bij de manier waarop de gegevens in-memory verwerkt worden. Hierdoor wordt een belangrijke snelheidswinst bij wegschrijven en teruglezen van gegevens gehaald.

Voor verwerking van massale hoeveelheden gegevens worden HDFS clusters van duizenden nodes ingezet. Op deze HDFS infrastructuur kunnen verschillende soorten Map-Reduce software draaien. De trend, ingezet door Spark, is dat de gegevens in-memory verwerkt worden. De bij Hadoop Map-Reduce gebruikelijke opslag op disk is vrijwel achterhaald, want te langzaam.

Technologiekeuze is lastig omdat tientallen start-up bedrijven slimme, maar vaak nog onbewezen, technologie ontwikkelen om een deel van de Big Data Analytics te realiseren. Het lijkt erop dat de grote IT vendors (IBM, HP, SAP) nog even wachten met overnames, totdat duidelijker is voor welke aanbieders de markt gaat kiezen.

Voor het Ministerie van Veiligheid & Justitie is het daardoor nu ook niet opportuun een definitieve keuze te maken. Ervaring opdoen in pilots, en geen onomkeerbare beslissingen nemen is de beste strategie op dit moment.

Validiteit en Betrouwbaarheid van Big Data.

De onderzoeksvraag naar validiteit en betrouwbaarheid van Big Dat is erg goed en uiterst relevant. Op dit moment is het onmogelijk hierop een finaal antwoord te geven. De Big Data methoden zijn niet ontworpen door wetenschappers met het doel de betrouwbaarheid van hun conclusies aan te geven, maar door online verkopers om meer boeken te verkopen of meer advertenties op hun zoekmachine. Dat Big Data in praktijk werkt is overduidelijk. Een wetenschappelijke theorie om Big Data analytics te vergelijken met statistisch betrouwbaar onderzoek (“Als we een betrouwbaarheid van 98% willen, betekent dit dat de steekproef zo groot moet zijn.”) is er nog niet. Bij Big Data zijn de gegevens vaak ook niet gericht verzameld om een bepaalde vraag te beantwoorden, maar wordt uitgegaan van de beschikbare gegevens en worden daar explorerend goede vragen bij gezocht. Bedrijven als Google en Amazon kunnen aan hun omzet zien of het voor hun werkt, maar dat is uiteraard geen wetenschappelijke methode.

Mogelijke toepassingen voor Ministerie van Veiligheid & Justitie

We zien een heel scala aan mogelijke toepassingen voor het ministerie van Veiligheid & Justitie.

Het interessants op korte termijn, lijkt ons het toepassen van de nieuwe analyse technieken op bestaande gegevensverzamelingen, als historische aangiftes en strafdossiers en in beslag genomen administraties. Dit levert geen onzekerheden ten aanzien van validiteit of betrouwbaarheid, en biedt

waarschijnlijk interessantere inzichten dan het analyseren van openbare databronnen dat al door veel andere bedrijven en organisatie wordt gedaan.

Verder lijkt het ons interessant om de nieuwe generatie apps van IBM en Apple [IBM 2015], die gebruik maken van Watson Big Data analytics in the cloud uit te proberen met fake gegevens. Bijvoorbeeld *Case Advice* bij reclasseringsmedewerkers en *Incident Aware* bij politie. Door pilots met deze al beschikbare apps te doen kan, zonder lange, risicovolle ontwikkeltrajecten kan mogelijk snel draagvlak gecreëerd worden voor Big Data oplossingen. Big Data jobs draaien op Cloud platform raden we af, omdat er dan het gevaar bestaat dat de controle over privacy-gevoelige data verloren wordt.

Tenslotte adviseren wij om een op Big Data feeds gebaseerd dashboard te ontwikkelen met relevante externe bronnen als Twitter, Facebook, CBS gegevens, weervoorspellingen en nieuwsberichten om zelf ervaring op te doen met Big Data technologie. Voorlopig zal dit dashboard reguliere onderzoeken niet vervangen, maar wellicht wel een interessante 'second opinion' geven van wat er in maatschappij gebeurt.

Vervangen bestaande onderzoeken

Op langere termijn zullen sommige Big Data analyses wellicht de huidige survey-gebaseerde onderzoeken overbodig maken. Op dit moment is het daar echter nog te vroeg voor. Er zijn veel kanttekeningen te zetten bij de validiteit en betrouwbaarheid van massaal verkregen gegevens. Het is goed te realiseren dat de Big Data vaak bijeengeraapt is, en niet specifiek verzameld om een bepaalde onderzoeksvraag te beantwoorden. De grotere hoeveelheid compenseert wellicht deels de verminderde geldigheid, maar op dit moment zijn ze nog niet te vergelijken met zorgvuldig samengestelde steekproeven en speciale vragenlijsten opgesteld voor een specifieke onderzoeksvraag.

Real-time Analytics

Het is zeker mogelijk om bepaalde, regelmatig verschijnende onderzoeksrapporten te complementeren met continue metingen en presentatie van resultaten op een dashboard. Door verschillende typen data near-real-time te presenteren komen beleidsmakers of onderzoekers wellicht op ideeën voor samenhang tussen verschillende fenomenen. Deze inzichten kunnen dan weer aanleiding tot gericht onderzoek zijn.

Belangrijk aspect hierbij is goed rekening te houden met de privacy van burgers en de vertrouwelijkheid van bepaalde zakelijke gegevens

De volgende stap…

Voor het ministerie van Veiligheid en Justitie openen deze ontwikkelingen een heel scala aan toepassingsmogelijkheden, zowel bij preventie, opsporing, rechtspraak en reclassering. Aandachtspunten zijn de vertrouwelijkheid van de gegevens waarmee gewerkt wordt en de in hoeverre de statische betrouwbaarheid van conclusies voldoende is om oordelen op te baseren. We adviseren daarom ook om in gang gezette, gerichte onderzoeken (monitors) niet stop te zetten, maar parallel te draaien aan nieuwe experimenten met Big Data uit interne of externe bron. Juist de vergelijking tussen de twee onderzoeksmethoden kan nieuwe inzichten verschaffen over validiteit en betrouwbaarheid van de uitkomsten.

Ons advies is om sowiezo praktisch aan de gang te gaan met Big Data toepassingen. Er zijn recent een aantal op Big Data analytics gebaseerde apps gelanceerd door IBM en Apple, waarmee ervaring kan worden opgedaan. Daarnaast zou een experimenteel systeem kunnen worden gebouwd om

gegevens uit eigen of externe bronnen te verwereken in een real-time dashboard van de veiligheids-situatie in Nederland. Gezien de vertrouwelijkheid van de gegevens waarmee Veiligheid &Justitie werkt is het aan te bevelen deze experimenten op een eigen Hadoop cluster uit te voeren. Op deze manier zelf ervaring opgebouwd worden met Big Data technieken, en resultaten vergelijken met lopende traditionele onderzoeken.

Concreet kan begonnen worden met een kleinschalige Hadoop omgeving met bijvoorbeeld 8 nodes (Intel servers met Linux OS) waarop Hadoop wordt geïnstalleerd, verbonden met een terrabit switch. Op deze manier wordt voorkomen dat in te vroeg stadium gekozen wordt voor systeem van een leverancier met een complete oplossing. Door wat kleine pilot projecten te draaien kan ervaring worden opgedaan, en eerste succesjes worden behaald. Bijvoorbeeld door veel sneller dan een bestaande DataWareHouse omgeving tot inzichten te komen (Ervaring bij Bol.com is dat klus die vroeger de hele nacht op Oracle SQL-databases draaide, nu in enkele minuten op Hadoop wordt gedraaid.), of compleet nieuwe vermoedens te toetsen danwel patronen te herkennen in beschikbare gegevens.

Als Spark ontwikkelingen zo snel gaan dan Hadoop ingehaald wordt, kan eenvoudig overgestapt worden van het de ene Open Source oplossing naar de andere.

Consequentie is wel dat er medewerkers zijn die beschikken over kennis en vaardigheden om de servers te installeren, te beheren en Hadoop jobs te programmeren. Eventueel kan dat ook in samenwerking met Hogeschool Rotterdam gebeuren, in afstudeerprojecten of onderzoeksprojecten met Creating010.

Bronnen

Boeken

- [Breur 2014] Tom Breur “Big Data - De nieuwe Goudkoorts?” 2014
- [Easley e.a. 2010] David Easley & Jon Kleinberg “Networks, Crowds and Markets: Reasoning about a Highly Connected World” Cambridge University Press 2010
- [Foreman 2014] John W. Foreman “Data Smart - Using data science to transform information to insight”
- [Hinssen 2015] Peter Hinssen “The network always wins”, McGraw Hill publications
- [Inmon 2002] W.H. Inmon “Building the Data Warehouse – third edition”, John Wiley Computer Publishing
- [Kimball e.a. 2002] Ralph Kimball, Margy Ross “The Data Warehouse Toolkit – The Complete Guide to Dimensional Modelling - second edition”, John Wiley Computer Publishing
- [Mayer e.a. 2013] Viktor Mayer-Schönberger & Kenneth Cukier “De Big Data revolutie - hoe de data explosie al onze vragen gaat beantwoorden”
- [McKinney 2013] Wes McKinney “Python for Data Analysis”, O’Reily
- [Robinson e.a. 2013] Ian Robinson, Jim Webber & Emil Eifrem “Graph Databases” O’Reilly 2013
- [Silver 2012] Nate Silver “The Signal and the Noise - Why so many predictions fail, but some don’t”
- [White 2012] Tom White “Hadoop: The Definite Guide”, 3rd edition, O’Reilly 2012.

Wetenschappelijke artikelen, Onderzoeksrapporten

- [Bongers 2014] Frank Bongers. “Big data: een ontdekkingsreis voor bestuurders en onderzoekers. Working paper voor het politicologenetmaal 2014, Maastricht 12 en 13 juni 2014”
- [Choenni 2011] Sunil Choenni, Ronald Meijer “From police and judicial database to an offender-oriented data warehouse”, Proceedings iadis international conference e-society 2011 edited by Piet Kommers and Pedro Isaias
- [Cohen 2013] Job Cohen e.a. “Twee werelden: hoofdrapport commissie ‘Project X’ Haren” http://www.rijksoverheid.nl/documenten-en-publicaties/kamerstukken/2013/03/08/brief-tweede-kamer-aanbieding-rapport-van-de-commissie-haren.html
- [Dean e.a. 2004] Jeffrey Dean and Sanjay Ghemawat “MapReduce: Simplified Data Processing on Large Clusters”
- [De Vries e.a. ??] Diederik W. de Vries, Sunil Choenni, Erik Leertouwer “A Data Warehouse Approach to Public Safety Monitoring”
- [Duyn e.a. 2014] Paul A.C. Duijn, Victor Kashirin & Peter M.A. Sloot “The relative ineffectiveness of criminal network disruption” http://www.nature.com/srep/2014/140228/srep04238/full/srep04238.html#t1 february 2014
- [Wang 2003] Avery Li-Chun Wang “An Industrial-Strength Audio Search Algorithm” http://ee.columbia.edu

Technische Documentatie, Manuals

- Hadoop docs “HDFS (Hadoop Distributed File System) Federation” http://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-hdfs/Federation.html
- Hadoop docs “MapReduce NextGen a.k.a. MapReduce v2 a.k.a. YARN (Yet Another Resource Negotiator)” http://hadoop.apache.org/docs/r2.4.1/hadoop-yarn/hadoop-yarn-site/YARN.html
- MongoDB docs http://docs.mongodb.org/manual/

- Neo4J docs http://neo4j.com/docs/stable/
- Spark docs https://spark.apache.org/

Beleidsnotities

- EU “Open Big Data in Europe” http://tech.eu/features/381/open-big-data-in-europe/
- NYU en NHS “The Open Data Era in Health and Social Care” http://thegovlab.org/nhs/

Nieuwsberichten, Overzichtsartikelen, Consultancy White Papers, Expert Interviews

- [Brown e.a. 2011] Brad Brown, Michael Chui & James Manyika “Are you ready for the era of ‘Big Data’? in McKinsey Quarterly, October 2011
- [Dijkstra 2014] Hanne Dijkstra. ‘Met big data is een uitbraak van ebola te voorspellen. Financieel Dagblad, 21 augustus, 2014
- [EMA e.a. 2013] EMA & 9Sight Consulting “Operationalizing the Buzz: Big Data 2013” november 2013
- [Gartner 2013] Gartner “Extend Your Portfolio of Analytics Capabilities
- [IBM 2014] IBM Global Business Services “Analytics: The real-world use of big data”
- [IBM 2014] IBM Global Business Services “The new hero of big data and analytics: The Chief Data Officer”
- [IBM 2014]] IBM Global Business Service “Analytics: The speed advantage”
- [IBM 2015] http://www.ibm.com/mobilefirst/us/en/mobilefirst-for-ios/
- [McKinsey 2011] McKinsey & Company “ Big Data: The next frontier for innovation, competition and productivity”
- [Warwick 2014] “Big Data Security Analytics Still Immature, Say Security Experts”, Computer Weekly http://www.computerweekly.com/news/2240230864/Big-data-security-analytics-still-immature-say-security-experts