Inhoud

1 Inleiding

1.1 Waarom?

Dit boek geeft een introductie in Big Data, op dit moment een belangrijk onderwerp. Veel bedrijven en instellingen hebben de indruk dat hun voortbestaan voor een belangrijk deel afhangt van het feit of zij in het komende decennium Big Data kunnen gebruiken. Gevolg is dat veel bedrijven en instellingen zo veel mogelijk gegevens verzamelen van zo veel mogelijk acties en transacties. Toch is slechts het verzamelen van alleen gegevens net zoiets als het vergaren van heel veel geld. Op zich is geld, net als heel veel gegevens, niets waard. Wie veel geld heeft kan er letterlijk in zwemmen, maar dan nog heeft het geen waarde. Maar er zijn wel meer overeenkomsten tussen geld en Big Data. Het is net zo lastig om aan veel geld te komen als aan veel (goede) gegevens. Bovendien heeft geld waarde omdat je er dingen voor kunt kopen en Big Data heeft waarde omdat er veel (verborgen) informatie in te vinden is.

Om de (verborgen) informatie uit Big Data te halen wordt een hele serie gereedschappen aangeboden, die allemaal zonder uitzondering volgens de leverancier de allerbeste zijn. De vraag blijft wel, welke van het aangeboden gereedschap voor uw bedrijf en voor uw toepassing nu echt het

Afbeelding 1.1: Dagobert Duck die liever zwemt in zijn geld dan dat hij het nuttig gebruikt.

beste is. Een andere vraag is natuurlijk: wie mogen de Big Data gaan onderzoeken en wie hebben toegang tot de resultaten? Daarnaast is het de vraag op welke manier de resultaten van alle onderzoeken en bewerkingen gepresenteerd moeten worden. Ook hier geldt dat dit niet alleen afhankelijk is van de resultaten zelf, maar ook van het publiek waarvoor de presentaties gemaakt worden.

Het waarom van dit boek is dat het een overzicht geeft van de (on)mogelijkheden van het onderzoeken en verwerken van Big Data en richting geeft aan het gebruik ervan. Dit boek geeft de gemiddelde gebruiker van de resultaten van analyses van Big Data houvast aangaande de inhoud, betekenis en achtergrond van de informatie waarmee gewerkt wordt. Met die kennis is het ook mogelijk om een indruk te krijgen van het nut en het belang van Big Data voor de eigen organisatie.

1.2 Waarover?

Dit boek gaat over het verzamelen van Big Data, over de manier waarop Big Data verwerkt kan worden en op welke manieren Big Data kan spreken. Nu kan elke stap tot in de kleinste details omschreven worden, maar dat is niet de weg die hier gevolgd wordt. In dit boek wordt elke stap beschreven vanuit het gezichtspunt van de mensen die (dagelijks) werken met de resultaten van het onderzoeken van Big Data en die de uitkomsten daarvan vertalen naar nieuwe plannen, nieuwe producten of een nieuwe aanpak voor bestaande problemen. Daarom wordt er in dit boek anders gekeken naar de drie grote stappen van het proces:

- **Big Data verzamelen** Dit gaat vooral over de verschillen met 'gewone databasesystemen' en de gevolgen die dit meebrengt bij het opzetten van een computersysteem voor het verwerken ervan.
- **Big Data onderzoeken** Dit bespreekt de belangrijkste technieken om Big Data te onderzoeken, geeft hun belangrijkste voor- en nadelen en beschrijft van elk van deze technieken voor welk soort onderzoek ze gebruikt worden.
- **Big Data laten spreken** Dit kan alleen als de sterke en zwakke punten bekend zijn, en dat met een goed zicht op de beperkingen. Geen enkele techniek geeft de absolute waarheid. Bovendien speelt

mee dat verkeerd gebruik of verkeerd vertalen van de resultaten van onderzoek van Big Data erg snel de grens van het ethisch toelaatbare kan overschrijden. Het inzicht dat Big Data ook hier 'op het scherp van de snede' werkt is van het grootste belang bij het vertalen van de uitkomsten.

Bij het bespreken van deze drie onderwerpen wordt zo veel mogelijk de techniek buiten beschouwing gelaten: dit is voor de specialisten (paragraaf 1.3). Belangrijker is het aangeven van de onderlinge relaties tussen de verschillende onderwerpen en de (constante) manier waarop zij elkaar beïnvloeden. Werken met Big Data betekent werken in een team waarbij iedere medewerker belangrijk is door zijn eigen taak en positie. Feitelijk gaat dit boek over de gemeenschappelijke basis voor de teams die werken met Big Data.

1.3 Voor wie?

Hoewel dit boek specifiek gaat over een onderwerp uit de wereld van de ICT is het niet uitsluitend geschreven voor deze doelgroep. Zoals hiervoor uiteengezet is Big Data te belangrijk, te omvangrijk en te veelomvattend om alleen binnen de wereld van de ICT te houden. Iedereen krijgt op kortere of langere termijn met de resultaten van Big Data te maken, niet alleen medewerkers van de ICT-afdeling. Wel verschilt per groep medewerkers het niveau en de diepgang van de kennis, ondanks het feit dat ze gezamenlijk één team vormen:

- **Gebruiker van Big Data** Deze persoon heeft algemene kennis nodig van de manier waarop Big Data onderzocht wordt en hoe de resultaten tot stand komen. Dit betekent kennis van de principes van verwerking, opslag, onderzoek en rapportage.
- **Analist van Big Data** Deze persoon heeft naast een algemene kennis van het hele proces een diepgaande kennis nodig van de technieken voor het onderzoeken van Big Data en de nauwkeurigheid en trefzekerheid van dergelijke onderzoeken.
- **Programmeur van Big Data** Deze persoon heeft naast algemene kennis van het hele proces een diepgaande kennis nodig betreffende

het doelmatig benaderen van grote hoeveelheden informatie en aangaande het verwerken van deze informatie

- **Ontwerper van Big Data** Deze persoon heeft naast algemene kennis van het hele proces een diepgaande kennis nodig van de opbouw van de systemen die het verzamelen, opslaan en beschikbaar stellen van grote hoeveelheden gegevens en informatie mogelijk maken.
- **Manager van Big Data** Deze persoon heeft naast algemene kennis van het hele proces een brede kennis nodig om een team dat vaak op de grens van het technisch haalbare zoekt naar bedrijfskundig verantwoorde oplossingen bij het gebruik van Big Data te begeleiden en te enthousiasmeren.

Iedereen die werkt met Big Data heeft, ongeacht zijn of haar positie binnen het proces, een goede basiskennis nodig. Juist die gemeenschappelijke basiskennis maakt een goede en zinvolle communicatie tussen de verschillende betrokkenen mogelijk en zorgt voor 'teamwork' (afbeelding 1.2).

Afbeelding 1.2: De optimale vorm van 'teamwork' op het rugbyveld, de scrum.

1.4 Hoe?

Om de beginselen van het gebruik van Big Data uit te leggen begint dit boek met een inleiding over Big Data zelf en de relatie met 'gewone data' (hoofdstuk 2). In de volgende hoofdstukken (3 en 4) wordt de totale procesgang bij het verwerken van Big Data uitgelegd en aangegeven wat de problemen en oplossingen zijn bij het opslaan en verwerken van (zeer) grote hoeveelheden informatie. De belangrijkste technieken voor het verwerken van de gegevens zelf komen in de hoofdstukken daarna aan bod. Achtereenvolgens zijn dit de beslisboom (hoofdstuk 5), neurale netwerken (hoofdstuk 6), clusteren (hoofdstuk 7), lineaire regressie (hoofdstuk 8), naaste buur (hoofdstuk 9) en afleiden van regels (hoofdstuk 10).

De volgende twee hoofdstukken gaan over het omgaan met de uitkomsten van de analyses van Big Data. Hoofdstuk 11 concentreert zich op de zin en de onzin van grootschalige analyses van heel veel informatie en hoe de kwaliteit hiervan bewaakt en geborgd kan worden. Hoofdstuk 12 kijkt ten slotte naar de ethische kant. Het laatste hoofdstuk, het nawoord (hoofdstuk 13) vat de belangrijkste zaken van opzet en aanpak samen.

De bijlagen geven aanvullende informatie bij de verschillende hoofdstukken. Bijlage A bevat een praktijkvoorbeeld van het gebruik van Big Data en bijlage B geeft enige wiskundige onderbouwing van gebruikte analysemethoden. In bijlage C wordt een korte handleiding gegeven hoe met verschillende algemeen toegankelijke Big Data-toolkits geoefend kan worden in het toepassen van de verschillende technieken en methoden. Een overzicht van de betreffende toolkits is in bijlage D te vinden. Tot slot leest u meer over Big Data-certificering in bijlage E.

2 Big Data

2.1 Inleiding

Wil een boek over Big Data zinvol en bruikbaar zijn, dan is het belangrijk dat duidelijk is wanneer er sprake is van data en wanneer van Big Data. Met de huidige computertechniek is het geen probleem om grote hoeveelheden gegevens (data) van verschillende bronnen op te slaan in één groot systeem. Vervolgens kunnen de gebruikers deze gegevens op verschillende manieren benaderen, op verschillende manieren rangschikken of de gegevens toepassen bij het ontwikkelen van nieuwe producten. Om dit soort acties mogelijk te maken worden de gegevens in een database opgeslagen. Een dergelijke database kan bijzonder veel gegevens bevatten, bijvoorbeeld gegevens van alle belastingplichtigen of de gegevens (brieven, röntgenfoto's, rapporten, laboratoriumuitslagen) van alle patiënten van een groot ziekenhuis. Toch wordt er in dit soort gevallen niet gesproken over *Big Data* maar gewoon over een *big database*.

De eerste vraag is dan ook waar de scheiding ligt tussen Big Data en een big database. De tweede vraag die direct in het verlengde ligt betreft de technieken die gebruikt kunnen worden bij het hanteren van Big Data, een proces dat *datamining* wordt genoemd. Zijn dezelfde technieken op dezelfde manier toepasbaar als bij een database of is een andere aanpak nodig?

Een antwoord op de eerste vraag wordt in dit hoofdstuk gegeven. Hiervoor wordt eerst naar de geschiedenis van gegevensopslag gekeken (paragraaf 2.2). Daarna wordt een meer formele beschrijving van het begrip Big Data ontwikkeld (paragraaf 2.3) vanuit het meer algemene begrip data. Dan volgt paragraaf 2.4 waarin gekeken wordt naar de relaties met de 'gewone' database en databasetechnieken. Ten slotte geeft de laatste paragraaf een kort overzicht van mogelijke toepassingen.

Voor het beantwoorden van de tweede vraag, welke technieken zijn beschikbaar voor datamining, is meer ruimte nodig. Zo gaat het volgende hoofdstuk (hoofdstuk 3) eerst nader in op de methodiek van het

verzamelen, opslaan en indelen van data. Vervolgens gaat hoofdstuk 4 in algemene zin in op de technieken en methoden voor het analyseren van de gegevens. Daarna volgen zeven hoofdstukken waarin de verschillende methoden voor het onderzoeken van deze data worden beschreven. In hoofdstuk 12 komen de *do's* and *dont's* van het analyseren van data aan de orde. Het laatste hoofdstuk (*Uitleiding*, hoofdstuk 13) kijkt terug en relativeert want ook hier is niet alles in beton gegoten en ook hier bestaan vele wegen die naar Rome leiden.

2.2 Geschiedenis

Met de uitvinding van het schrift vond de mensheid niet alleen de bureaucratie uit maar ook de grootschalige opslag van gegevens. Zo is een groot deel van onze kennis van het Nieuw-Assyrische rijk (negende tot zevende eeuw voor onze jaartelling) gebaseerd op gegevens uit het Koninklijke archief van Ashurbanipal (685 BCE tot 627 BCE, afbeelding 2.1). Op zijn hoogtepunt heeft deze bibliotheek meer dan 30.000 documenten en voor het beheer waren geen computers met een data-

Afbeelding 2.1: Een beeld van Ashurbanipal, de laatste grote heerser van het Nieuw-Assyrische rijk.

base management system (DBMS) beschikbaar. Door de verschillende teksten slim in te delen was het mogelijk om elk gewenst document te vinden. Feitelijk pasten de Assyrische bibliothecarissen technieken toe die tegenwoordig gemeengoed zijn bij het opzetten en vullen van een database.

> Data is de verzamelnaam voor gegevens en dan in het bijzonder voor gegevens die we in een computersysteem opslaan. Voorbeelden van data zijn een naam, een adres, het rekeningnummer van een bankrekening, het saldo op die bankrekening en een foto van de rekeninghouder. Data worden in een computer opgeslagen in één of meer bytes, waarbij een byte staat voor één teken, letter of getal. De grootte van een database wordt tegenwoordig in GB (gigabyte = 10^9 bytes) maar vaak ook al in TB (terabyte = 10^{12} bytes) gegeven. De echt grote databases worden opgegeven in PB (petabyte = 10^{15} bytes) en het is de verwachting dat de EB (exabyte = 10^{18} bytes) binnenkort ons leven binnenwandelt.
> Deze data worden gebruikt door computerprogramma's die transacties uitvoeren op basis van deze data. Een aantal transacties dat bij elkaar hoort vormt samen een proces.

Met het verstrijken van de eeuwen worden de bibliotheken steeds groter en groter maar nooit is er sprake van Big Data. Elke bibliotheek blijft toegankelijk via een kaartenbak en eenvoudige zoekstrategieën. Ook andere grote gegevensverzamelingen bleven met eenvoudige technieken toegankelijk, ondanks de vaak enorme hoeveelheden. De belangrijkste reden hiervoor was de tussenkomst van de mens bij het invoeren van nieuwe gegevens. Gelijktijdig met het invoeren werden de nieuwe gegevens gerubriceerd en ingebed in de gebruikte systematiek van de gegevensopslag. Maar rond de laatste eeuwwisseling kwam er een kink in deze eeuwenoude kabel.

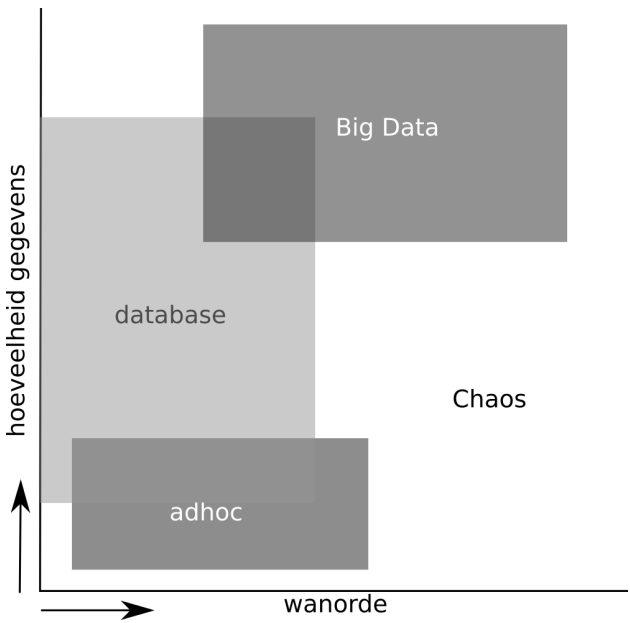
Aan het begin van deze eeuw werd het mogelijk om steeds grotere hoeveelheden gegevens automatisch en zonder de tussenkomst van mensen op te slaan. Een belangrijke stuwende kracht achter deze ontwikkeling waren de sterk verbeterde mogelijkheden van geautomatiseerde systemen. Zo maakte de groei van de opslagsystemen het mogelijk om veel

meer gegevens zonder meer op te slaan, dus zonder enige selectie vooraf. Dit betekende dat er veel gegevens ongestructureerd werden opgeslagen en dat bij een deel van de systemen deze opslag een min of meer continu proces is. Het is in deze situatie dat de conventionele methoden voor gegevensbeheer niet meer bruikbaar waren en het is hier dat voor het eerst de term Big Data werd gebruikt.

2.3 Beschrijving

Op grond van de ontstaansgeschiedenis is het mogelijk om een aantal eigenschappen van Big Data af te leiden:

- **eigenschap 1** Grote hoeveelheden gegevens.
- **eigenschap 2** Geen controle op het compleet zijn van elk onderdeel.
- **eigenschap 3** Afwezig zijn van enige ordening.
- **eigenschap 4** De hoeveelheid gegevens kan op elk moment (sterk) veranderen.



Afbeelding 2.2: Onderlinge verhoudingen tussen Big Data en andere systemen voor het hanteren en opslaan van gegevens.

Vergeleken met de conventionele (gestructureerde) opslagsystemen valt direct op dat Big Data zich onderscheidt van andere systemen door de afwezigheid van enige ordening of structuur in combinatie met de grote hoeveelheid gegevens (afbeelding 2.2) waarvan niet gegarandeerd is dat elk onderdeel compleet is. Indien aan één van de voorwaarden is voldaan, dus de afwezigheid van ordening of zeer veel gegevens, dan is er geen sprake van Big Data, maar slechts van een ongeordende verzameling gegevens cq. zeer veel gegevens. Zo kunnen grote hoeveelheden gegevens met een zekere ordening uitstekend benaderd worden met de bestaande databasesystemen. Kleinere hoeveelheden ongeordende gegevens zijn uitstekend op een ad-hocbasis te benaderen. Worden dergelijke verzamelingen vaker benaderd, dan is het lonend om ze alsnog te ordenen en op te nemen in een bestandssysteem. Juist de combinatie van beide eigenschappen maakt dat we in een geheel afwijkende situatie zitten. Hieruit volgt dat een definitie van Big Data kan zijn:

> *Big Data* is een zo grote hoeveelheid ongestructureerde en (soms) niet-complete set gegevens dat verwerking met conventionele databasesystemen niet mogelijk is.

Een andere definitie legt juist de nadruk op de belangrijkste verschillen tussen 'gewone data' en 'Big Data' waarbij verwezen wordt naar de *drie V's* uit één van de eerste publicaties:

> *Big Data* zijn verzamelingen gegevens die zich kenmerken door de grote hoeveelheid (*volume*), de grote snelheid waarmee nieuwe gegevens ontstaan (*velocity*) en de grote onderlinge variatie (*variety*).

In dit boek wordt vanuit de combinatie van beide definities van het begrip Big Data gewerkt. Deze (combinatie van) definitie(s) wordt ook gebruikt bij het bespreken van de verschillende zoekstrategieën en bij het aanreiken van oplossingen bij gerezen problemen.

2.4 Relaties

Een logische vraag is in hoeverre Big Data en databasesystemen ondanks hun duidelijk andere wortels en andere manier van benaderen aan elkaar gerelateerd zijn. Een gebruiker gaat in beide systemen op zoek naar informatie en die informatie wordt op één of andere manier met zoekopdrachten verzameld. Juist op dat moment spelen de bepalende verschillen tussen beide systemen een grote rol:

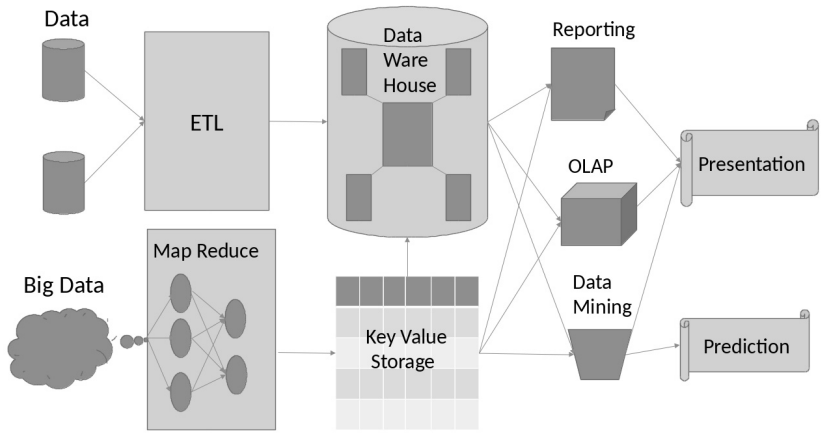
- **Big Data** Ongeordend en deels incompleet, dus een zoekopdracht werkt altijd op de hele verzameling en heeft (soms) aanvullende zoektermen nodig.
- **Databasesysteem** Geordend en compleet, dus de meest voorkomende zoekopdrachten werken alleen op indexbestanden.

Gevolg van deze andere strategie is dat het zoeken in Big Data over het algemeen meer tijd vraagt. Dit verschil in zoektijd wordt overigens niet alleen veroorzaakt door de afwezigheid van ordening in Big Data maar ook door de grootte van de gegevensverzameling. Daarnaast is het maar de vraag of in beide gegevensverzamelingen met dezelfde zoekstrategieën gewerkt kan worden.

De praktijk leert echter dat er naast zuivere 'databasesystemen' en zuivere 'Big Data' ook nog gegevensverzamelingen bestaan met kenmerken van beide systemen. Dit betekent dat een deel van de gegevens inderdaad een structuur kent maar dat die structurering ontbreekt bij een ander deel. Wordt een gemengd systeem geschematiseerd (afbeelding 2.3) dan is ook direct duidelijk dat verschillende gereedschappen (tools) niet voor alle soorten informatie toepasbaar zijn.

Een voorbeeld van een dergelijk gemengd systeem is een ziekenhuisinformatiesysteem waarin allerlei medische informatie van iemand wordt opgeslagen. Voor een deel gaat het hier om een databasesysteem met indexering, bijvoorbeeld door gebruik te maken van zoektermen als geboortedatum, adres en behandelend arts. Dit soort gegevens is eenvoudig te controleren op volledigheid en juistheid. Dat ligt anders bij laboratoriumgegevens en beeldinformatie. Op het eerste gezicht is ook

hier een indexering mogelijk op externe eigenschappen, zoals type apparaat of bepaling en dergelijke.



Afbeelding 2.3: Schema van een gemengd informatiesysteem waarin databases worden gecombineerd met Big Data met de belangrijkste gereedschappen (tools).

Het wordt anders als de waarden van een laboratoriumbepaling of de gebruikte meetmethode meegenomen worden. Datzelfde geldt als informatie uit medische beelden zoals de dikte van bepaalde botten onderdeel wordt van een bepaalde zoekactie. De gevraagde informatie is mogelijk aanwezig, maar zeker is het niet en bovendien is niet duidelijk waar deze informatie precies te vinden is en hoe volledig de informatie is. Bovendien speelt zeker bij medische gegevens ook de tijd een belangrijke rol. Zo kunnen bloedwaarden in de tijd veranderen, soms zelfs over het verloop van een enkele dag. Het is dan ook de vraag op welke manier deze 'gemengde systemen' benaderd moeten worden. Wordt de ordening vergeten en wordt de verzameling slechts als 'Big Data' gezien of hangt het antwoord af van de zoekacties?