

# Inhoud in vogelvlucht

<b>Inleiding</b> .....	1
<b>Deel 1: Aan de slag met big data</b> .....	7
HOOFDSTUK 1: De fundamenteën van big data .....	9
HOOFDSTUK 2: Bigdatatypen verkennen .....	27
HOOFDSTUK 3: Oud en nieuw: distributed computing .....	39
<b>Deel 2: De technologische basis voor big data</b> .....	47
HOOFDSTUK 4: Meer over de technische onderdelen van big data .....	49
HOOFDSTUK 5: Virtualisatie en hoe dit distributed computing ondersteunt ..	65
HOOFDSTUK 6: De cloud en big data verkennen .....	77
<b>Deel 3: Bigdatabeheer</b> .....	91
HOOFDSTUK 7: Operationele databases .....	93
HOOFDSTUK 8: De grondbeginselen van MapReduce .....	111
HOOFDSTUK 9: De wereld van Hadoop verkennen .....	121
HOOFDSTUK 10: De fundering en het ecosysteem van Hadoop .....	131
HOOFDSTUK 11: Apparaten en bigdatawarehouses .....	141
<b>Deel 4: Analyse en big data</b> .....	151
HOOFDSTUK 12: Bigdata-analyse definiëren .....	153
HOOFDSTUK 13: Tekstanalyse en big data begrijpen .....	167
HOOFDSTUK 14: Aangepaste werkwijzen voor de analyse van big data .....	183
<b>Deel 5: Big data implementeren</b> .....	197
HOOFDSTUK 15: Gegevensbronnen integreren .....	199
HOOFDSTUK 16: Omgaan met streaminggegevens en complexe gebeurtenisverwerking .....	213
HOOFDSTUK 17: Big data operationeel maken .....	223
HOOFDSTUK 18: Big data toepassen in je organisatie .....	233
HOOFDSTUK 19: Veiligheid en toezicht voor bigdataomgevingen .....	247
<b>Deel 6: Bigdataoplossingen in de echte wereld</b> ..	259
HOOFDSTUK 20: Het belang van big data voor bedrijven .....	261
HOOFDSTUK 21: Gegevens in beweging analyseren .....	271
HOOFDSTUK 22: Bedrijfsprocessen verbeteren met bigdata-analyse .....	283
<b>Deel 7: Deel van de tientallen</b> .....	293
HOOFDSTUK 23: Tien beste werkwijzen voor big data .....	295
HOOFDSTUK 24: Tien uitstekende hulpbronnen voor big data .....	301
HOOFDSTUK 25: Tien dingen die je wel of niet moet doen met big data .....	307

<b>Woordenlijst</b> .....	311
<b>Index</b> .....	325

# Inleiding

---

Welkom bij *Big data voor Dummies*. Big data is snel bezig zich tot een van de belangrijkste technologische trends te ontwikkelen die de manier waarop bedrijven gegevens gebruiken om de klantervaring te verbeteren en hun zakelijke modellen te transformeren potentieel drastisch kan veranderen. Hoe gaat een bedrijf te werk om zo veel mogelijk voordeel uit gegevens te halen? Wat houdt het in om enorme hoeveelheden gegevens in kennis te transformeren? We bieden je in dit boek inzichten in hoe je met technologische ontwikkelingen in software, hardware en uitvoermodellen op nieuwe manieren gegevens kunt gebruiken.

Big data is geen afzonderlijke markt. Het is in plaats daarvan een combinatie van gegevensbeheertechnieken die in de loop van de tijd zijn ontwikkeld. Big data maakt het organisaties mogelijk om op het juiste moment en met de juiste snelheid enorme hoeveelheden gegevens op te slaan, te beheren en te manipuleren om de juiste inzichten te kunnen krijgen. De sleutel tot het begrijpen van big data is dat gegevens beheerd moeten worden om aan de zakelijke eisen te kunnen voldoen van een bepaalde oplossing. De meeste bedrijven zijn in een vroeg stadium in hun reis met big data. Veel bedrijven experimenteren met technieken die het hen mogelijk maken om enorme hoeveelheden gegevens te verzamelen en te bepalen of er verborgen patronen in deze gegevens voorkomen die een vroege waarschuwing kunnen vormen voor een belangrijke verandering. Sommige gegevens zullen misschien aangeven dat het aankoopgedrag van klanten aan het veranderen is of dat er nieuwe elementen in het spel zijn die aangesproken moeten worden voordat het te laat is.

Er zullen veel nieuwe kansen opduiken terwijl bedrijven beginnen nieuwe soorten bigdataoplossingen te evalueren. Fabricagebedrijven zullen bijvoorbeeld misschien van machinesensoren afkomstige gegevens kunnen analyseren om te bepalen hoe processen moeten worden aangepast, voordat er een catastrofale gebeurtenis optreedt. Kleinhandelaren zullen gegevens in het oog kunnen houden waarmee ze prijzen voor aan klanten gerelateerd producten onmiddellijk kunnen aanpassen terwijl een transactie wordt uitgevoerd. Bigdataoplossingen kunnen in de gezondheidszorg worden gebruikt om de oorzaak van een ziekte te bepalen en een arts opties te bieden voor de behandeling.

Big data is echter geen geïsoleerde oplossing. Het implementeren van een bigdataoplossing vereist dat er een infrastructuur bestaat die de schaalbaarheid, de distributie en het beheer van de betreffende gegevens ondersteunt. Het is daarom belangrijk zowel een zakelijke als een technische strategie te implementeren om gebruik te kunnen maken van deze belangrijke technologische trend.

Er zijn veel belangrijke redenen waarom we denken dat het belangrijk is dat je bigdatatechnieken begrijpt en weet op welke manieren bedrijven opkomende technieken zoals Hadoop, MapReduce en nieuwe database-engines gebruiken om de waarde van hun gegevens te transformeren. We hebben dit boek geschreven om een beeld te geven van wat big data is en hoe het de manier verandert waarop organisaties meer uit gegevens kunnen halen dan in het verleden mogelijk was. We denken dat dit boek je de context zal geven om weloverwogen beslissingen te nemen.

## Over dit boek

Big data is voor veel mensen nog iets nieuws. We moeten daarom begrijpen wat de technische en zakelijke eisen zijn en hier onderzoek naar doen. Allerlei verschillende mensen moeten iets weten over big data. Sommigen zullen zich in de technische details willen verdiepen, terwijl anderen willen weten wat de economische implicaties zijn als je bigdatatechnieken gebruikt. Andere managers moeten genoeg informatie hebben om te kunnen begrijpen welke invloed big data op zakelijke beslissingen kan hebben. Het implementeren van een bigdataomgeving vereist zowel een architectonische als een zakelijke aanpak; en een heleboel planning.

Met welk doel je dit boek ook leest, we bespreken de volgende onderwerpen om je te helpen begrijpen wat big data is en welke invloed het op jouw bedrijf kan hebben:

- » Wat is de architectuur voor big data? Hoe kun je enorme hoeveelheden gegevens beheren zonder grote ontwrichtingen in je datacenter te veroorzaken?
- » Wanneer moet je de resultaten van je bigdata-analyse in je datawarehouse integreren?
- » Welke implicaties hebben beveiliging en toezicht op het gebruik van big data? Hoe kun je jouw bedrijf veilig houden?
- » Welke waarde hebben verschillende gegevenstechnieken en wanneer moet je overwegen ze als een onderdeel van je bigdatastrategie te gebruiken?

- » Welke soorten gegevensbronnen kun je gebruiken voor het analyseren van big data? Hoe kun je verschillende soorten analyses toepassen op zakelijke problemen?

## Dwaze veronderstellingen

Hoezeer we ook alles voor iedereen duidelijk willen maken, we moesten tijdens het schrijven van dit boek kiezen wie volgens ons het meeste geïnteresseerd zouden zijn in *Big data voor Dummies*. Dit is wie je volgens ons bent:

- » **Je bent slim.** Je bent zeker niet dom, maar het onderwerp van big data geeft je geen plezierig gevoel. Het is je niet helemaal duidelijk waar het om gaat en je zult misschien proberen van onderwerp te veranderen als je om een definitie wordt gevraagd.
- » **Je bent een zakenpersoon die weinig of niets met technologie te maken wil hebben.** Maar je leeft in de 21e eeuw en kunt er dan ook niet aan ontsnappen. Je hoort mensen zeggen 'het draait allemaal om big data' en je denkt dat je maar beter kunt uitzoeken waar ze het over hebben.
- » **Je bent een IT-persoon die een heleboel weet over technologie.** Je bent echter niet bekend met big data. Iedereen zegt er iets anders over. Je wilt nu toch echt weten wat het hele beeld is.

Wie je ook bent: welkom. Wij zijn hier om je te helpen.

## Hoe dit boek is opgebouwd

We hebben het boek in zeven delen verdeeld om het lezen gemakkelijk te maken. Voel je vrij om te bladeren.

### Deel 1: Aan de slag met big data

We leggen in dit deel de basisideeën uit die je nodig hebt om big data in zijn geheel te begrijpen, zowel vanuit een technisch als een zakelijk gezichtspunt. We introduceren ook de belangrijke concepten en onderdelen, zodat je niet zult onderdoen in welk zinvol gesprek over big data dan ook.

## **Deel 2: De technologische basis voor big data**

Deel 2 is voor zowel technische als zakelijke professionals die de verschillende soorten onderdelen van big data en de achterliggende technische ideeën die big data ondersteunen moeten begrijpen. We helpen je in dit deel te begrijpen welke soort infrastructuur big data praktischer zal maken.

## **Deel 3: Big data beheren**

Deel 3 is voor zowel technische als zakelijke professionals, maar gaat veel dieper in op de details van verschillende databaseopties en opkomende technieken zoals MapReduce en Hadoop. Als je meer weet over deze onderliggende technieken begrijp je ook beter wat er achter deze belangrijke trend schuilt.

## **Deel 4: Analyse en big data**

Hoe analyseer je enorme hoeveelheden gegevens die een onderdeel van je bigdata-infrastructuur worden? We gaan in dit deel dieper in op de verschillende soorten analyses die nuttig zijn om echte betekenis aan je gegevens te ontleen. Dit deel helpt je na te denken over manieren waarop je big data kunt omzetten in acties voor je bedrijf.

## **Deel 5: Big data implementeren**

Dit deel gaat in op de details van wat het betekent om gegevens echt te beheren. Hierbij kun je denken aan onderwerpen als het operationeel maken van je gegevens en het beschermen van de veiligheid en privacy van deze gegevens. Je krijgt hier veel stof tot nadenken op dit zeer belangrijke gebied.

## **Deel 6: Bigdataoplossingen in de echte wereld**

Je leert in dit deel hoe bedrijven beginnen big data te gebruiken om hun zakelijke werkzaamheden te transformeren. Mocht je willen weten wat je in de toekomst allemaal met gegevens kunt doen, dan is dit deel voor jou.

## Deel 7: Het deel van de tientallen

Mocht je de schatkist van *Voor Dummies* nog niet hebben ontdekt, dan zul je vast en zeker niet bekend zijn met het deel van de tientallen. De uitgever dwingt *Voor Dummies*-auteurs in dit deel nuttige brokjes informatie te leveren die gemakkelijk toegankelijk zijn in lijsten van tien (of daaromtrent) verhelderende onderdelen. We begonnen deze hoofdstukken al tegenstribbelend, maar we zijn uiteindelijk heel blij dat ze er zijn. En we denken dat jij dat ook zult zijn, nadat je de beste werkwijzen voor big data en voorbeelden van wat je wel en juist niet moet doen hebt gelezen. Dit staat allemaal in het deel van de tientallen.

### Woordenlijst

We hebben een woordenlijst opgenomen van termen die vaak gebruikt worden als mensen big data bespreken. Hoewel we er naar streven om termen te definiëren waar we ze in dit boek introduceren, denken we dat je de woordenlijst een nuttig hulpmiddel zult vinden.

## Pictogrammen in dit boek



PAS OP

Let op. Dit kan je veel moeite besparen.



BELANGRIJK

Het zal je misschien spijten als je dit weetje mocht vergeten.



TIP

We markeren uitzonderlijk nuttige punten waar je aandacht aan moet besteden met dit pictogram.



TECHNISCHE  
INFO

Je vindt hier weetjes voor de meer technisch aangelegden.

# Hoe verder?

---

We hebben je een overzicht gegeven van big data en alle belangrijke onderdelen ervan geïntroduceerd. Lees de eerste vier hoofdstukken om een idee te krijgen van waar big data over gaat en welke technieken er bestaan om implementaties mogelijk te maken. De volgende twee hoofdstukken introduceren enkele dingen die je moet weten over de onderliggende infrastructuur. De volgende acht hoofdstukken gaan gedetailleerder in op de verschillende soorten gegevensstructuren die fundamenteel zijn voor big data.

Je kunt het boek van omslag tot omslag lezen, maar voor het geval dat dit niet jouw stijl is, hebben we geprobeerd de *Voor Dummies*-stijl aan te houden waarbij hoofdstukken op zichzelf staan. Je kunt dus rechtstreeks naar de onderwerpen gaan die je het interessantst lijken. Waar je ook begint, wij wensen je het beste.

Veel van deze hoofdstukken zouden uitgebreid kunnen worden tot hele, op zichzelf staande boeken. Big data en het landschap van de opkomende technieken zijn een belangrijk richtpunt voor ons bij Hurwitz & Associates en we nodigen je uit onze (Engelstalige) website te bezoeken en onze blogs en inzichten te lezen op [www.hurwitz.com](http://www.hurwitz.com).



**De geschiedenis van het gegevensbeheer****Waarom big data belangrijk is voor bedrijven****Big data toepassen op zakelijke effectiviteit****De fundamentele elementen van big data definiëren****De rol die big data in de toekomst zal spelen**

## Hoofdstuk 1

# De fundamentele van big data

**H**et beheren en analyseren van gegevens bood altijd al de grootste voordelen en de grootste uitdagingen voor organisaties in alle industrietakken van welke grootte dan ook. Bedrijven hebben er lang mee geworsteld om een pragmatische aanpak te vinden voor het bewaren van informatie over hun klanten, producten en diensten. Dat was redelijk rechttoe rechtaan en eenvoudig voor bedrijven met maar weinig klanten die allemaal op dezelfde manier hetzelfde product kochten. Maar bedrijven en de markten waar ze in participeren zijn langzamerhand steeds ingewikkelder geworden. Deze bedrijven hebben meer productlijnen toegevoegd en de manier waarop ze hun producten leveren gediversifieerd om te overleven of beter te kunnen concurreren. Maar niet alleen bedrijven worstelen met gegevens. Organisaties die aan onderzoek en ontwikkeling doen hebben er bijvoorbeeld mee geworsteld voldoende rekenkracht te krijgen. Deze hebben ze nodig om geavanceerde modellen te kunnen gebruiken of beeldmateriaal en andere bronnen van wetenschappelijke gegevens te kunnen verwerken.

We hebben zeker met veel ingewikkelde dingen te maken als het om gegevens gaat. Sommige gegevens zijn gestructureerd en worden opgeslagen in een traditionele relationele database, terwijl andere gegevens ongestructureerd zijn, waaronder documenten, klantenservicerecords en zelfs afbeeldingen en video. Bedrijven moeten ook rekening houden met

nieuwe bronnen van door machines zoals sensoren gegenereerde gegevens. Andere nieuwe informatiebronnen worden door mensen gegenereerd, zoals gegevens van social media en de door interacties met websites geproduceerde klikstroomgegevens. De beschikbaarheid en acceptatie van nieuwere, krachtigere mobiele apparaten gekoppeld aan alomtegenwoordige toegang tot globale netwerken zal ook tot het ontstaan van nieuwe gegevensbronnen leiden.

Hoewel elke gegevensbron onafhankelijk kan worden beheerd en doorzocht, bestaat tegenwoordig voor bedrijven de uitdaging uit hoe ze betekenis kunnen vinden in de maalstroom van al deze verschillende soorten gegevens. Het is onmogelijk om op traditionele manieren over het gegevensbeheer te denken als je met zoveel informatie in zoveel verschillende vormen moet omgaan. We hebben weliswaar altijd al veel gegevens gehad, maar het verschil is dat er tegenwoordig significant veel meer zijn, in allerlei soorten met verschillende tijdsgevoeligheid. Organisaties vinden ook meer manieren dan ooit tevoren om gebruik te maken van deze informatie. Je moet daarom anders over het beheren van gegevens denken. Dat is de kans en de uitdaging van big data. In dit hoofdstuk lees je meer over de evolutie van de overgang op big data, de context waarin dit gebeurt en wat deze evolutie voor jouw organisatie betekent.

## De evolutie van het gegevensbeheer

Het zou mooi zijn als we elke nieuwe innovatie in het gegevensbeheer als nieuw begin konden zien zonder verbinding met het verleden. De meeste nieuwe stadia of golven in het gegevensbeheer, of die nu revolutionair of incrementeel zijn, bouwen echter voort op eerdere ontwikkelingen. Hoewel het gegevensbeheer gewoonlijk wordt gezien vanuit de softwarebril, moet dit feitelijk vanuit een holistisch oogpunt worden bekeken. Het gegevensbeheer moet rekening houden met technische vooruitgang op het gebied van hardware, opslag, netwerken en rekenmodellen zoals virtualisatie en cloudcomputing. Het samenspel van opkomende technieken en de prijsdaling van opslagmogelijkheden tot rekenkracht hebben het gegevenslandschap getransformeerd en nieuwe kansen mogelijk gemaakt.

De samenkomst van al deze technische factoren transformeert de manier waarop we gegevens beheren en uitbaten. Big data is de nieuwste trend die door deze factoren ontstaat. Wat houdt big data nu dus eigenlijk in en waarom is het zo belangrijk? We bieden verderop in dit boek een uitgebreidere definitie. Om je op weg te helpen, big data wordt gedefinieerd als elke soort gegevensbron die minstens de volgende drie kenmerken heeft: