Big Data: hype of toekomst?

Tony Busker, Mortaza S. Bargh, Jan Kroon en Sunil Choenni

Inleiding

Door de ontwikkelingen op het gebied van o.a. informatietechnologie en de groeiende informatiebehoefte van organisaties is de verwachting dat er meer data zal worden geproduceerd en opgeslagen. De OECD (Organisation for Economic Co-operation and Development) schat dat er voor het einde van dit decennium vijftig miljard devices, zoals smart phones, middels mobiele netwerken met elkaar verbonden zullen zijn (OECD 2012; Kalidien & Choenni 2015). Schattingen wijzen uit dat er meer dan vijf miljard devices met elkaar verbonden zijn. Thans kan bijvoorbeeld een smart phone gebruikt worden om lokatiegegevens te genereren, tekstberichten te verwerken en foto's en video's te genereren.

De data die door en middels deze devices worden gegenereerd is zo omvangrijk en divers dat ze niet te beheren zijn met de gebruikelijke architecturen en systemen, zoals conventionele databasesystemen. Big data is een opkomend discipline die zich richt op het verwerken van grote hoeveelheden data van verschillende types in een snel tempo (real-time).

De verwachtingen rondom de toepassingsmogelijkheden van big data zijn hoog gespannen. Van koelkasten die aangesloten zijn op winkelketens en die zich automatisch aanvullen als ze leeg raken tot zelf rijdende auto's (Choenni et al. 2015). In deze bijdrage proberen we antwoord te geven op de vraag *of big data een hype is of de toekomst*.

Naar onze inschatting zijn er veel aanwijzingen om te veronderstellen dat big data geen hype is. Technisch gezien bevindt big data zich op het snijvlak van real-time, multimedia en conventionele large databases. Derhalve is het niet zo verwonderlijk dat bij het analyseren van big data gebruik wordt gemaakt van methoden en technieken uit deze vakgebieden. Uiteraard zullen deze methoden en technieken geïntegreerd en op maat gesneden moeten worden voor big data toepassingen.

Een tweede vraag waar we aandacht aan besteden in deze bijdrage is de vraag *of de hoog gespannen verwachtingen rondom de toepassingsmogelijkheden van big data al dan niet terecht zijn*. Op deze vraag is niet eenduidig een antwoord te geven. We zetten uiteen dat big data ook grote uitdagingen met zich meebrengen, in het bijzonder als het gaat om privacy en mogelijke misinterpretaties. Of potentiele big data toepassingen de werkelijkheid worden is mede afhankelijk van de mate waarin we adequaat kunnen reageren op deze uitdagingen.

We schatten in dat de verwachtingen rondom big data neerwaarts bijgesteld dienen te worden, maar dat het wel een discipline van betekenis wordt in de toekomst. Derhalve, hebben we aan de Hogeschool Rotterdam een 'Minor Big Data' opgezet die op termijn wordt omgedoopt tot 'Minor Data Science'. De minor beoogt de student een verzameling van methoden, technieken en tools aan te bieden zodat hij/zij is in staat om toegevoegde waarde te creëren aan bestaande kernprocessen van een organisatie door het gebruik van de grote hoeveelheden data die heden ten dagen gegenereerd worden. Voorts moet de student in staat zijn nieuwe toepassingen te creëren op basis van en met de data. In deze bijdrage *bespreken* we tevens de opbouw van de 'Minor Data Science'.

Mogelijkheden

Niet alles wat onder de vlag van Big Data wordt gepresenteerd is nieuw. Veel big data technieken zijn eerder toegepast bij Data Warehousing, Data Mining en Business Intelligence. Wat wel revolutionair is, is de schaal waarop deze technieken worden toegepast en de snelheid waarmee dat gebeurt.

De ontwikkeling van Big Data is begonnen bij grote internetbedrijven als Yahoo, Google en Amazon. Deze bedrijven zijn wereldwijde serviceproviders, die in principe elke muisklik van hun klanten (lees: browsersgedrag) kunnen opslaan en dat vaak ook doen. Amazon gebruikt deze click-data bijvoorbeeld om aanbevelingen aan hun klanten geven, als "Anderen die dit product kochten bekeken ook … Ook zouden ze iemand die net iets te lang aarzelt voordat op de 'Buy Now' knop gedrukt wordt een betere aanbieding kunnen doen, als "Dit is uw geluksdag! Als u het artikel nu koopt krijgt u 15% korting!.

Dit soort functionaliteiten is niet te realiseren met traditionele SQL (Structured Query Language)-databases. SQL-databases zijn ontworpen om transacties te verwerken. Ze zijn wel betrouwbaar, maar niet snel genoeg. Verder gaat het om dermate grote hoeveelheden data, dat het de capaciteit van SQL-databases vele malen overschrijdt.

Yahoo, Google en Amazon zijn daarom gaan experimenteren met massief parallelle verwerking van de click-data. De internetbedrijven beschikten al over een enorme computer capaciteit. Ze hebben datacenters vol standaard, goedkope Intel servers die via een snel lokaal netwerk verbonden zijn. De computers draaien software die in staat is om bij uitval van een server een andere computer in te schakelen die het werk naadloos overneemt.

Een manier om veel data snel te verwerken is door de klus te verdelen in een groot aantal kleine klusjes, die onafhankelijk van elkaar kunnen worden uitgevoerd en tussenresultaten opleveren, die snel kunnen worden gecombineerd tot een eindresultaat. Het bleek dat de software die dergelijke klussen kan uitvoeren eigenlijk steeds hetzelfde is: er moet data over computers verdeeld worden, als er een server uitvalt moet dat de boel niet in de war brengen, als tussenresultaten bekend zijn moeten ze opgehaald worden. Eigenlijk waren maar twee stukjes van de klus echt verschillend voor verschillende toepassingen: de verdelingen van de grote klus in kleine klusjes ('Map' genaamd), en het verwerken van de tussenresultaten tot een eindresultaat ('Reduce' genaamd). Beide operatoren zijn afgeleid van functies zoals toegepast binnen het functioneel programmeren. De map operator behelst veelal de uitvoering van een functie op elk element in een lijst en de reduce operator is een functie die een eindresultaat berekent voor een lijst van elementen.

In 2006 heeft Yahoo haar Map-Reduce software open source gemaakt, en is deze door een community van ontwikkelaars verder ontwikkeld tot Hadoop. Nu kan iedereen deze software gratis gebruiken (al zal niet iedereen over een computernetwerk beschikken waar Hadoop lekker op draait).

Inmiddels hebben veel meer bedrijven en instellingen te maken met een vloedgolf aan data. Elk bedrijf heeft een klantenadministratie, een website en vaak een mobile app. Klantcontacten worden vastgelegd in Customer Relationship Management systemen, op Twitter, Facebook en Instagram kunnen klanten 24 uur per dag, 7 dagen per week berichten plaatsen over de organisatie of haar producten.

Big data wordt gekenmerkt door vier V's: Volume, Velocity, Variety en Veracity. De eerste twee hebben we besproken: veel data en de wens om die snel te verwerken. De derde, Variety, duidt op de grote variatie aan datatypes (de Haan et al. 2011; van den Braak et al. 2012). De vierde, Veracity, duidt erop dat het waarheidsgehalte van de gegevens vaak minder dan 100% is.

In traditionele Data Warehouse toepassingen was er nauwelijks sprake van Variety: alle gegevens waren netjes in tabellen geordend, en pasten dus prima in SQL-Databases. Bij big data is dit niet meer het geval. Gegevens kunnen Tweets of Facebook posts zijn, met toegevoegde foto's of geluidsbestanden. Er wordt daarom gebruik gemaakt van nieuwe soorten gegevensopslag NO-SQL genaamd, een afkorting van Not Only SQL. De social graph van een klant (vrienden en vrienden van vrienden) kan bijvoorbeeld beter in een Graph-Database worden opgeslagen om snel doorzocht te kunnen worden. Verder zijn er document-databases, Key-Value-Stores en nog veel meer op de Variety van de gegevens toegesneden databases.

Bij traditionele Data Warehouse toepassingen was Veracity ook geen probleem. De gegevens waren door de organisatie zelf verzameld, geordend en opgeslagen in SQL-Databases. Uitgaan werd dat ze betrouwbaar waren. Bij big data is da anders. Sommige data sets zijn ingekocht en het waarheidsgehalte van Social Media berichten is ook discutabel. Bij big data is het dus verstandig om betrouwbaarheid van gegevens expliciet mee te nemen.

Activiteiten van big data scientists bestaan uit:

1. Verzamelen van ruwe gegevens,
2. Schonen, combineren, filteren en in bruikbaar formaat zetten van gegevens,
3. Verwerken van gegevens (vaak op Hadoop cluster met Map-Reduce software),
4. Presenteren van resultaten, visualiseren van gegenereerde data sets.

Het doel is dat de managers of bestuurders die deze resultaten lezen en/of de visualisaties bekijken tot nieuwe inzichten komen, waardoor ze betere besluiten kunnen nemen en de concurrentie net een stapje voor blijven.

Uitdagingen

Naast de mogelijkheden brengt Big Data ook een aantal uitdagingen met zich mee. Twee van de uitdagingen die we in deze bijdrage willen bespreken zijn het *waarborgen van privacy* en het *voorkomen van misinterpretaties* (Choenni et al. 2015; van den Braak et al. 2013) Privacy is een veelomvattend en complex begrip. Privacy poogt de persoonlijke levenssfeer van mensen te beschermen en is derhalve subjectief en contextgevoelig. De aard van privacyvraagstukken is afhankelijk van het domein waar men in opereert. In het domein van openbare veiligheid hebben privacyvraagstukken veelal te maken met de onthulling van de (administratieve) identiteit van een persoon, terwijl in de gezondheidszorg de identiteit veelal geen probleem is, maar gaat het om de handelingen van een persoon. In de context van de openbare veiligheid is het niet gewenst dat de politie allerlei reisbewegingen zou kunnen relateren aan de naam en het adres van een persoon, terwijl in de context van een ziekenhuis deze gegevens van een patiënt bekend zijn. Echter, het is ongewenst dat anderen meekijken met behaalde handelingen die een patiënt verricht, bijv. als hij/zij aan het douchen is.

De wet bescherming privacy gegevens is in het leven geroepen om zorgvuldige opslag en verwerking van persoonsgegevens te bewerkstellingen. Onder persoonsgegevens wordt data verstaan waaruit de identiteit van een persoon kan worden afgeleid. Voorbeelden van dergelijke persoonsgegevens zijn login id, BSN nummers, kentekennummers enz. Daarnaast is het opslaan en verwerken van bepaalde kenmerken, zoals seksuele geaardheid, levensovertuiging enz., aan strenge eisen onderworpen. Doordat met Big Data diverse bronnen worden ontsloten en gerelateerd, wordt de kans op onthullingen van deze kenmerken steeds groter.

Stel dat een school de gemiddelde examencijfers, uitgesplitst naar jongens en meisjes publiceert. Als bij een bepaalde vak maar twee meisjes aan het examen hebben deelgenomen en een van de meisjes op social media een bericht plaats dat zij een negen heeft gehaald voor het examen, kan het cijfer van het andere meisje afgeleid worden.

Een ander voorbeeld is het combineren van data met de reisbewegingen die iemand maakt. Met veel "smart devices" kan bijgehouden worden op welke positie iemand zich bevindt. Door deze data te combineren met Google Maps kan precies bepaald worden welke clubs/instanties iemand bezoekt. Als bekend is wat voor activiteiten binnen een club plaatsvinden, kan met een grote waarschijnlijkheid bepaald worden wat voor levensstijl een persoon er op nahoudt, voor andere voorbeelden zie (Choenni et al. 2015; Stottelaar et al. 2014).

Om de gewenste privacy van mensen te waarborgen dient wellicht veel beter het gebruik en verwerking van data geregeld te worden. Dit moet niet alleen beperkt worden voor persoonsgegevens Thans wordt adequate toegang tot data – middels access control mechanismen – goed ondersteund door systemen. Mechanismen voor adequate gebruik van data (use control) wordt nauwelijks ondersteund.

Een andere uitdaging die Big Data met zich meebrengt is het voorkomen van misinterpretaties. Data Analytics technieken zijn in staat om geavanceerde analyses uit te voeren op gekoppelde/gerelateerde gegevensverzamelingen/databases. Om de juiste duiding aan deze analyses toe te dichten is het vaak van belang om de beperkingen, de aard en de voorwaarden waaronder de dataverzameling tot stand is gekomen te kennen. Als dit niet het geval is, dan is de kans op misinterpretaties groot. Stel dat we een database hebben aangelegd over zwanen. Analyse van de database wijst uit dat alle zwanen wit zijn. Om een juiste duiding te geven aan dit resultaat is het van belang om te weten of de database was aangelegd om eigenschappen van witte zwanen op te slaan, of dat er alleen naar witte zwanen is gezocht enz. Immers als alleen naar witte zwanen is gezocht zou een conclusie als "(bijna) alle zwanen zijn wit op de planeet aarde" moeilijk te rechtvaardigen zijn.

De kans op misinterpretaties wordt in de context van Big Data groter omdat verschillende bronnen die verschillend van aard zijn met elkaar worden gekoppeld. Relevante informatie over deze bronnen zijn soms moeilijk te achterhalen omdat het veelal legacy databasesystemen betreffen. Voorts kan de tijdsdimensie voor sommige bestanden relevant zijn en voor andere bestanden minder. Bijvoorbeeld een database waarin de omloopstijd van planeten en de afstand van de planeten tot de aarde is opgeslagen is niet aan veranderingen in de tijd onderhevig. We kunnen uit deze database ten alle tijden de wet van Keppler afleiden zonder ons te bekommeren over de geldigheid van deze wet. Dit ligt bijvoorbeeld anders als we een database hebben over delicten die door mensen worden gepleegd op verschillende leeftijden. Stel dat we de verdelingsfunctie plotten van het aantal delicten en de leeftijd waarop men delicten pleegt (age-crime curve), dan is het goed voor te stellen dat zo een curve in de jaren 20 er anders uit zou kunnen zien dan in 2015. Immers de bevolkingsopbouw, waarden en normen zijn in de loop der tijd veranderd.

Voor legacy systemen geldt dat ze vaak data bevatten die in het verleden zijn verzameld en opgeslagen en derhalve zijn ze een representatie van hoe de werkelijkheid er toen uitzag. Door ze te betrekken in huidige analyses dient men zich dit te realiseren om misinterpretaties te voorkomen.

Minor op de Hogeschool Rotterdam

Big data neemt zo'n grote vlucht, dat er een tekort aan data scientists wordt voorspeld van ruim 5 maal zoveel vraag als aanbod op de arbeidsmarkt in 2020. Een goede reden om een specialisatie data scientist op te nemen in ons curriculum!

De minor Data Science is enkele jaren geleden gestart als de minor Rotterdam Open Data (Conradie et al. 2013; Conradie et al. 2013a). Na een jaar te hebben gedraaid bleek dat de nadruk niet zozeer op het verwerken van Open Data ligt, maar eerder op het verwerken van grote

hoeveelheden data (Big Data). Het doel van deze minor is de student een verzameling van methoden, technieken en tools aan te bieden zodat hij/zij in staat is om een toegevoegde waarde te creëren door het gebruik van grote hoeveelheden data. Een belangrijke actuele vraagstuk in de minor is: hoe ga je om met grote, complexe bestanden in diverse formaten afkomstig van verschillende organisaties?

Met deze kennis en vaardigheden kan een student een rol spelen in het vakgebied van de data scientist. Een vakgebied wat je zou kunnen onderverdelen in globaal drie gebieden:

- de data analyst: voor het analyseren van data (met behulp van statistiek), effectief presenteren en visualiseren van data,
- de machine learning engineer: voor het maken en bestuderen van algoritmen om van data te leren en voorspellingen te doen (vaak geschikt voor academici),
- de data engineer: minder statistiek, meer engineering om op basis van data een probleem in de praktijk op te lossen.

Van de data scientist wordt gevraagd om met een opdrachtgever een goede vraagstelling te formuleren en vervolgens hierop een antwoord krijgen, die vaak in de vorm van een (data)visualisatie is. Daarnaast moet hij/zij inzicht hebben in de mogelijkheden om (pre)processing van data te parallelliseren. Tenslotte moet hij/zij voldoende kennis van statistiek, van data mining, en machine learning algoritmen, inclusief de beschikbare tools (R, machine learning libraries, Spark), hebben.

Kernconcepten binnen de minor hebben betrekking op:

- Het verzamelen van data (privacy, transparantie, accountability, identiteit en identificatie, keuzevrijheid en efficiëntie en effectiviteit).
- Het opslaan van data in data spaces, Big Data infrastructuren zoals Hadoop (incl. complexiteit en benodigde activiteiten zoals opschonen en uniformeren van data).
- Het gebruiken van methoden en/of technieken voor visualisatie van informatie wat voor de doelgroep nieuwe inzichten oplevert bij complexe vraagstukken.

De minor is opgebouwd uit kennis-gestuurde cursussen en praktijk-gestuurde cursussen en ziet er als volgt uit:

- Onderwijsperiode 1: Statistiek met R (4 cp); Big Data, privacy, security analytics (3 cp); en Project (8 cp)
- Onderwijsperiode 2: Data Mining, Machine learning (4 cp); Business, Data visualisatie (4cp); en Project (7 cp)

Tijdens het project, dat over twee onderwijsperioden loopt, krijgen projectgroepen de beschikking over grote hoeveelheden data uit de praktijk. De teams formuleren zelf in samenwerking met een potentiele opdrachtgever een case. In de projecten komen alle aspecten zoals het exploreren, het converteren en visualiseren van data aan bod.

Conclusies

Samenvattend kunnen we concluderen dat big data geen hype is maar nu al voor veel bedrijven een reële optie is om concurrentievoordeel te behalen. Heel veel bedrijven, zeker bedrijven die op Internet actief zijn zoals Twitter en Spotify, verzamelen grote hoeveelheden gegevens waaraan een hoge waarde wordt toegekend. Lang niet al die bedrijven weten al hoe ze die waarde gaan realiseren. Er zal daarom een groeiende vraag naar Data Scientists ontstaan, om Actionable Insights te destilleren uit big data

Het programma dat Hogeschool Rotterdam in de ‘Minor Data Science’ aanbiedt, is een eerste stap op weg naar een specialisatie ‘Data Scientist’. Op deze manier leiden we Informatica

studenten op om behalve met computers, software en mensen ook met grote hoeveelheden data om te kunnen gaan.

Bronnen

R. Choenni, M.S. Bargh, C. Roepan, R. Meijer (2015), Privacy and security in smart data collection by citizens, in smarter as the new urban agenda: A Comprehensive View of the 21st Century City, Gil Garcia, J. Pardo, T., Nam, T., (eds), Springer, New York, USA.

P. Conradie, J. Lemmens, T. Busker, R. Choenni (2013), De SunnyApp – open data ontsluiten via het dataspaceprincipe, Proc. NIOC 2013, Nationaal Informatica Onderwijs Congres: Meeting the Future, Arnhem, D. Smeets (red.), Hogeschool van Arnhem en Nijmegen, pp. 408-412.

P. Conradie, J. Lemmens, N. Stembert, R. Choenni, I. Mulder (2013a), De toekomst is open: Rotterdam open data in onderzoek en praktijk, Kenniscentrum Creating 010, Hogeschool Rotterdam, Rotterdam, 2013, ISBN 978-90-820924-0-0.

G. de Haan, R. Choenni, I. Mulder, S. Kalidien, P. van Waart (2011), Bringing the research lab into everyday life: Exploiting sensitive environments to acquire data for social research, Oxford Handbook of Emergent Technologies in Social Research, Hesse-Bibber (Ed.), Oxford Publications, Oxford University Press, pp. 522-541.

S. Kalidien, S. Choenni (2015), Exploiting data for supporting developing countries, Newsletter IFIP W.G. 9.4 & Centre for Electronic Governance IIM Ahmedabad, Information Technology in Developing Countries 25(1) February 2015.

OECD (2012), OECD Internet economy outlook 2012, OECD Publishing. http://dx.doi.org/10.1787/9789264086463-en

B. Stottelaar, J. Senden, L. Montoya (2014). Online social sports networks as crime facilitators. Crime science, 3 (8). ISSN 2193-7680.

S. van den Braak, R. Choenni, S. Verwer (2013), Combining and analyzing judicial databases, in: Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases, Chapter 10, Springer Verlag, Berlin, pp. 191-208.

S. van den Braak, R. Choenni, R. Meijer, A. Zuiderwijk (2012), Trusted third parties for secure and privacy-preserving data integration and sharing in the public sector, Proc. DG.O 2012, 13th Annual Int. Conf. on Digital Government Research, Maryland, US, June 4-7, ACM Press, pp. 135-144.