Veranderende visie op data

In deze eerste aflevering van een serie over big data kijken we naar de techniek achter dataverwerking en databeheer. Plus de betekenis ervan voor het werk van informatieprofessionals.

Klaas Jan Mollema MSc.

NIEUWE SERIE: BIG DATA

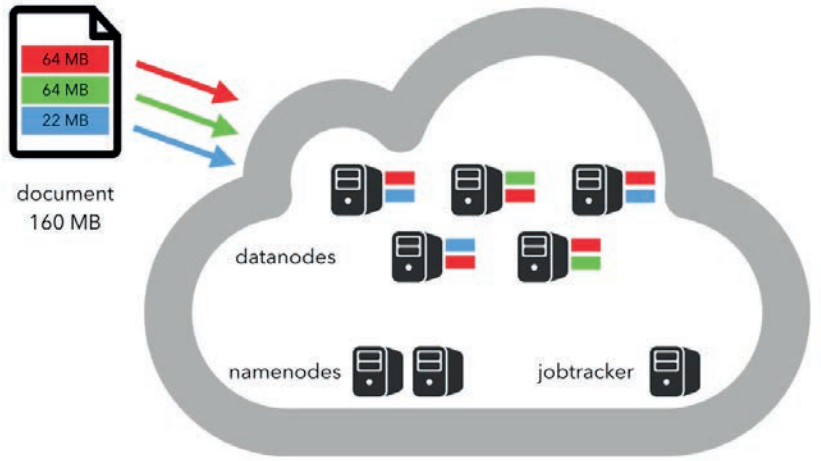
De ontwikkelingen rond dataverwerking en databeheer hebben elkaar de afgelopen jaren in rap tempo opgevolgd. Door zowel de toegenomen snelheid waarmee gegevens ontstaan als de steeds grotere hoeveelheid en verscheidenheid aan gegevens(stromen), voldeden klassieke (relationele) databasemanagementsystemen niet meer. Nieuwe vormen van bestandsopslag, gegevensverwerking, analyse en tools zagen het licht. Wat is de samenhang tussen die verschillende technologieën? Wat is waarvoor bedoeld? In deze serie wordt de lezer bijgepraat over terminologie en tooling rondom big data en big data-visualisatie.

Big data is geen verre toekomst meer, maar de huidige realiteit. Of je nu aan het werk bent of in de supermarkt rondloopt, je handelen wordt op basis van data geanalyseerd – waarna er vervolgens ten aanzien van jouw persoon conclusies worden getrokken. Ook elke smartphone analyseert het gebruik om te komen tot slimmere dienstverlening. Ja, zelfs de nieuwste generatie auto's onderzoekt aan de hand van sensoren wat de status van de auto-onderdelen is. Niet heel verwonderlijk dat het buzzword *big data* te pas en te onpas wordt gebruikt. Wanneer spreken we van big data en wanneer van 'veel data in een database'?

Veranderde visie op data

Technologie-onderzoeker Gartner definieert big data als informatie-eenheden die met een groot volume (*volume*), grote verwerkingssnelheid (*velocity*) en/of grote variëteit (*variety*) aan verschijningsvorm op ons afkomen. Dergelijke data vraagt innovatieve vormen van informatieverwerking, zodat de analyse ervan nieuwe inzichten kan bieden ten behoeve van beslissingen en procesautomatisering.[1] De drie v's (volume, velocity en variety) stellen andere eisen aan de capaciteit van systemen en de vaardigheden van de mensen die met die systemen werken. Dit vereist een andere visie op data.

Omdat de vorm en het volume van de data erg wisselend zijn, heb je al snel te maken met problemen rond fysieke opslag en serverperformance. Door vormen van databasearchitectuur te gebruiken die vrijer zijn dan de architecturen uit de jaren tachtig, negentig en tweeduizend worden hogere snelheden van dataverwerking en data-analyse behaald. Veel principes die informatici en informatiekundigen in de jaren tachtig en negentig leerden rond redundantie, normalisatie en relationeel databasemanagement hebben hiervoor wel moeten wijken. Zo ontstond een nieuwe visie op data: schaalbaarheid.



HADOOP FILE SYSTEM (HDFS)
Schematische weergave van HDFS waarbij een groot bestand wordt opgeknipt in kleinere bestanden. Vervolgens worden de kleine stukjes redundant opgeslagen om te garanderen dat de data niet verloren gaat en de performance van dataretrieval verbeterd wordt.

Schaalbare fysieke opslag en performance

Bij big data zijn opslag en performance de eerste zaken waar je mee te maken krijgt. Voerden we in het verleden data vooral handmatig in systemen in, door de komst van sensortechnieken en internet gaan we steeds vaker digital born data opslaan en verwerken. Die grote hoeveelheden data zorgen ervoor dat een systeem op een bepaald moment 'vol zit' en 'eruit klapt'. Excel bijvoorbeeld kan maximaal 1.048.576 rijen aan. En we hebben allemaal de ervaring dat de performance van een systeem direct onderuit gaat als een bestand te groot of te complex wordt.

Omdat Google direct bij haar ontstaan voor de grote uitdaging stond om de almaar groeiende datastromen te indexeren, ontwikkelde zij nieuwe oplossingen voor dataverwerking en data-opslag.[2] Klassieke dataverwerking tools (ETL, Extract, Transform, Load) kunnen immers dergelijke datastromen niet aan. Google ontwikkelde daartoe MapReduce: een geheel nieuwe vorm van dataprocessing. MapReduce is te vergelijken met een groep mensen die een deel van de data 'voorsorteert' op basis van een vooraf afgesproken criterium, om de 'stapeltjes' data vervolgens samen te voegen en per stapel ergens op te slaan. Op deze manier kunnen grote aantallen servers alle data tegelijk verwerken en voorgesorteerd opslaan.

Voor de opslag van al die data ontwikkelde Google het zogeheten Hadoop Distributed File System (HDFS). In tegenstelling tot gewone computers slaat dit systeem bestanden niet op één plek – met een eventuele back-up – op, maar deelt het systeem de bestanden op in kleine stukjes die vervolgens op vier of meer servers dubbel worden bewaard. Door vervolgens elk van deze opgeslagen kopieën beschikbaar te stellen voor gebruik wordt de performance van het opslagsysteem enorm vergroot. Opvragen van data en het analyseren ervan gebeurt nu op vier of meer servers in plaats van op een enkele.

Doordat data nu op zoveel plaatsen is opgeslagen en vandaaruit wordt aangeboden, is het gebruik van kostbare, geavanceerdere servers niet langer nodig. Men kan volstaan met servers die beschikken over middelmatige rekenkracht en minder goede harde schijven. Gaat een schijf kapot, dan wordt deze uit de werkende server getrokken en vervangen door een nieuwe. Die schijf wordt vervolgens door de rest van HDFS 'bijgepraat' en is vrij snel volledig in dienst, zonder dat het totale systeem of de gebruiker er last van gehad heeft.

Schaalbare vormen van databasearchitectuur

Als data is opgeslagen in het gedistribueerde bestandssysteem, kan deze worden ingeladen in een database-omgeving. Omdat de hoeveelheid en de vorm van data (aantal rijen; aantal kolommen) vooraf onbekend is en je schaalbaar wilt blijven werken, wordt data niet in tabellen met rijen opgeslagen, maar in tabellen die eindeloos uitbreidbaar zijn ('columnair').

Vooraf wordt dus geen vaste data-architectuur gemaakt en wordt niet genormaliseerd (zoals bij relationele databases gebruikelijk was). Pas in de query-fase worden objecten met elkaar gecombineerd. Dit vraagt van data-analisten om slimme queries op te stellen die de gegevens op de juiste wijze combineren tot een antwoord. Tevens is het belang van de metadata en de documentatie van de data toegenomen, om te kunnen garanderen dat de juiste gegevens met elkaar worden gecombineerd

en er geen verkeerde inzichten ontstaan. Om queries los te laten op in HDFS opgeslagen data, zijn verschillende tools ontwikkeld.

Schaalbare dataverwerking

Opslag, performance en structuur zijn nu geregeld. Een volgend probleem dat dient te worden opgelost is de snelheid waarmee data het systeem binnenkomt. Deze datastreams kunnen op twee manieren verwerkt worden. *Continuous processing* is het direct verwerken van data. Dat kost veel computerperformance en wordt bijvoorbeeld gebruikt voor het realtime weergeven van de status van een performance indicator bij een organisatie. *Real time querying* laadt data in kleinere hoeveelheden in om het klaar te zetten voor queries; dit kost door de stapsgewijze aanpak minder performance van systemen. Hoe je met de snelheid van de data omgaat is afhankelijk van de (realtime) urgentie van de vragen die je wilt stellen aan het systeem en de betrouwbaarheidsfactor van het antwoord. Wanneer je twitterberichten analyseert is het missen van één bericht niet zo heel erg, maar wanneer je banktransacties analyseert moet je wel op het systeem kunnen vertrouwen.

De cloud

Met de komst van Hadoop had Google een enorme use-case te pakken: goedkope, flexibele dataopslag. Google wilde nog een stap verder gaan door ook de hardware goedkoop en flexibel te maken. Het bedrijf doneerde de technologieën MapReduce en het Hadoop Distributed File System (HDFS) aan de open source-community (Apache Hadoop) en zette de stap richting clouddiensten die bedrijven kunnen afnemen. Google ontwikkelde servers met goedkope hardware en een compleet nieuwe vorm van dataopslag. In enorme datacenters verspreid over de hele wereld worden alle taken uitgevoerd die nodig zijn voor de Googlediensten. De capaciteit die 'over is' verhuurt Google als cloud-service aan bedrijven: zij kunnen hier hun data opslaan en analyseren. Amazon en Microsoft bieden inmiddels ook dergelijke diensten aan (de voor- en nadelen van cloud-diensten komen later in deze serie aan bod).
Organisaties die werken met minder gevoelige data kunnen kiezen voor de cloud-variant. Organisaties waar privacy en veiligheid belangrijk zijn kunnen binnenshuis servers inrichten met de Hadoop-software.

Kracht van big data

Nu de drie v's getackeld zijn door een verbeterde efficiënte opslag, flexibele datastructuren en snelle verwerkingsmogelijkheden, wordt de kracht van big data-oplossingen zichtbaar. Niet alleen is het mogelijk om die enorme hoeveelheid data die beschikbaar is, op te slaan en met elkaar te combineren tot nieuwe inzichten over deze data, ook is het op basis van de historie mogelijk voorspellende analyses te doen met computeralgoritmes. Een bekend voorbeeld is Amazons Recommendation Algorithm. Maar ook je eigen mobieltje weet op basis van GPS waar je woont, werkt en slaapt om vervolgens te voorspellen waar je waarschijnlijk over een uur naartoe wilt en je alvast te informeren over de verkeersdrukte. Je auto weet op basis van data van andere autobezitters exact te voorspellen wanneer een auto-onderdeel defect zal raken.
De mogelijkheden van dergelijke algoritmes op basis van data zijn enorm. Dat geldt ook voor de waarde ervan. En niet alleen in economische zin: denk ook aan de waarde voor bijvoorbeeld medisch onderzoek en veiligheidssituaties. Het gaat hier om een spannende ontwikkeling die zowel technische en informatie-inhoudelijke alsook ethische aspecten raakt.

Rol informatieprofessionals en ICT'ers

Informatieprofessionals en ICT'ers dienen oog te hebben voor deze big data-ontwikkelingen die leiden tot revolutionaire vormen van automatisering en kunstmatige intelligentie. Klassieke taken van de informatiespecialist worden overgenomen door slimme algoritmes die op basis van big data slimmer zijn dan ooit. Blendle[3] bijvoorbeeld maakt geautomatiseerde knipselkranten met krantenartikelen op basis van diverse profieldata die je zelf hebt achtergelaten op het internet.
Gezien deze ontwikkelingen ligt de kracht van de informatieprofessional vooral in het vermogen kritisch en conceptueel data en informatie te duiden, alsmede ervaring met statistiek in te brengen om deze grote hoeveelheden data door middel van goede queries te kunnen bevragen.
ICT'ers zullen steeds vaker te maken krijgen met grote datavraagstukken en dienen zorg te dragen voor een veilige, stabiele en betrouwbare opslag en representatie. Waar veel ontwerpprincipes uit de jaren tachtig en negentig vervangen zijn door nieuwe, geldt het 'garbage in/garbage out'-principe nog steeds: als je troep in het systeem stopt, komt er troep uit. Datakwaliteit is een hot item en de trend van 2017. Voor informatieprofessionals is dit bekend terrein: zij zijn bij uitstek goed in het bewaken van de kwaliteit van informatie en het duiden ervan. <

Noten

1] Zie: www.gartner.com/it-glossary/big- data/.
2] Zie: www.mapr.com/blog/5-google-projects-changed-big-data-forever.
3] Zie: tinyurl.com/hr8jsqq.

Klaas Jan Mollema MSc. (www.zijlmo.nl/mo) is specialisatiecoördinator Business Data Management aan de opleiding Informatica van Hogeschool Leiden.