Bilateraal

Informatiekwaliteit (AAB)

Paul van der Linden

In 2003 werd door de University of California in Berkeley een onderzoek gedaan naar de hoeveelheid informatie die jaarlijks wordt gecreëerd. Uit onderzoek bleek dat in 1999 twee miljoen Terabyte aan nieuwe informatie werd geproduceerd (print, film, magnetische en optische opslag). In 2002 ging het zelfs om vijf miljoen Terabyte (dat is vijf Exabyte) aan nieuwe data. Hierbij ging het uitsluitend om oorspronkelijke data en niet om kopieën. Telefoon, radio, TV en internet zijn samen goed voor bijna 18 Exabyte aan nieuwe data. Het omgaan met deze gigantisch groeiende informatiebrij is een serieus probleem. Het wereldwijde web werd in genoemd onderzoek geschat op 170 Terabyte aan informatie (2003). In die zin kan het worden gezien als de grootste database waarover we beschikken. Bijzonder handig en voor velen onmisbaar voor het snel bij elkaar zoeken van benodigde informatie. Denk bijvoorbeeld aan CBS.nl, Wikipedia of de online versie van de Encyclopedia Brittanica. Maar ook aan websites als Nu.nl, Hyves, Youtube en de schier oneindige verzameling van blogs. Datakwaliteitsproblemen worden in DBM regelmatig beschreven. Het informatiekwaliteitsprobleem is hierbij vergeleken een wees. In de kern gaat het bij informatiekwaliteit om de betrouwbaarheid van de informatie waarmee we werken. Internet is een goede case om dit duidelijk te maken. In het onderstaande zijn drie willekeurige berichten opgenomen die van het internet zijn gehaald. Voor het gemak heb ik ze A, B en C genoemd. Eén van deze berichten klopt niet, de andere twee wel. Aan u de opdracht om te bepalen welk bericht niet correct is.

A. Business Objects, leverancier van Business Intelligence-oplossingen (BI), heeft overeenstemming bereikt over de overname van Cartesis S.A. Cartesis is leverancier van enterprise performance management (EPM) oplossingen en heeft wereldwijd meer dan 1300 klanten. Met de overname is een bedrag gemoeid van $ 225 miljoen in contanten. De afronding van de acquisitie wordt verwacht binnen 90 dagen.

B. De artsen zeiden dat hij nog een half jaar tot een jaar had te leven en adviseerden Brandrick de hem resterende tijd optimaal te benutten. Hij nam ontslag, ging elke avond uit eten, betaalde zijn hypotheek niet meer, gaf zijn kleding aan een liefdadigheidswinkel en regelde zijn eigen begrafenis. Zijn vreugde was in eerste instantie groot toen de artsen na een jaar constateerden dat hij toch geen kanker had, maar een ongevaarlijke ontsteking. Daarna realiseerde hij zich echter dat hij zijn huis moet verkopen. Brandrick heeft bovendien therapie nodig om de niet-fatale periode te verwerken.

C. SAS laadde 1,25 Terabyte aan transactiedata (de verkoopgegevens over drie jaar van een gemiddelde, grote internationale retailer) in een gebruiksklaar datamodel met een sterschema. Tijdens de uitvoering werden slowly changing dimension tables gevormd en data geschoond, getransformeerd en verwerkt zoals bij de datamart van een klant. De tijdsduur van deze klus bedroeg, inclusief alle processen, twee uur en 36 minuten, waarmee een nieuwe benchmark werd geïntroduceerd voor complexe en omvangrijke taken.

Oké. Genoeg tijd gehad om uw keuze te maken? Het onjuiste bericht is het eerste bericht. Het ging namelijk niet om $ 225 miljoen, maar om € 225 miljoen. Overigens is dit persbericht inmiddels gerectificeerd. Waar het om gaat is dat je vaak niet in staat bent om in te schatten of een bericht correct is of niet. Mijn inschatting is dat de meerderheid bericht B of C als onjuist heeft ingeschat.

Wat kunnen we hiervan leren? Het is niet altijd duidelijk welke informatie je mag geloven. In praktijk gaan we er vaak vanuit dat de informatie juist is. Slechts als bewezen wordt dat dit niet het geval is stappen we hiervan af. Positief bekeken zou je dat als een handelswijze in de traditie van Karl Popper kunnen noemen. Populair gezegd: zo gaat dat nou eenmaal. Je neemt iets voor waar aan totdat het tegendeel wordt bewezen. Dat leidt tot een aanpassing en een nieuwe 'waarheid'. Aangezien er op internet geen centrale autoriteit is (gelukkig maar!) is zo'n inzicht echter niet door te voeren naar alle delen van het web. Als we erachter komen dat bericht A onjuist is, is het praktisch gezien onmogelijk om alle kopieën en verschijningsvormen van bericht A aan te passen of te elimineren. Oude en nieuwe varianten blijven dus gebroederlijk naast elkaar bestaan, leidend tot maximale verwarring. Zie hier het informatieprobleem. We beschikken over steeds meer informatie waar we in principe elk moment van de dag (en nacht) bij kunnen komen. Wat we niet weten is wat de kwaliteit van die informatie is. Daarmee kennen we dus ook niet de bruikbaarheid ervan. Voor de hand ligt dat we in de toekomst een indicatie krijgen van de informatiekwaliteit. Informatie van de kwaliteit triple AAA kost dan het meeste. De laagste informatiekwaliteit is dan gratis, maar niet gecontroleerd. In de tussentijd blijft het behelpen. En hoe weet u dat dit bericht juist is?

Paul van der Linden (Paul.PFH.vanderLinden@AtosOrigin.com) is senior consultant Data Warehousing/BI bij Atos Origin en geeft leiding aan Data Warehousing Cost & Lifecycle Management (CLM).