

Bilateraal

Kubussen, konijnen en Kimball

Paul van der Linden

Kubussen zijn erger dan konijnen. Je begint met een of twee en voordat je het weet heb je er 385. Het hoeft geen betoog dat dit niet te beheren is en ook dat het op geen enkele wijze meer past in een batch window. Het idee achter de kubus is om een beperkte dataset vooraf naar een aantal dimensies door te rekenen en klaar te zetten. Wie een kubus bevraagt selecteert derhalve bepaalde cellen uit de kubus. Er hoeft niet meer gerekend te worden, waardoor het snel werkt. Dat is ook het grote pluspunt van de kubus. Het grote nadeel is dat wat er niet in zit ook niet getoond kan worden. En dat het niet echt geschikt is om detailgegevens in op te nemen. De gedachte om vele kleine kubussen in te zetten komt dan al snel op. Dit zijn al de eerste redenen waarom kubussen zich voorspoedig voortplanten. Kleine kubussen kosten ook minder tijd om op te bouwen en te verversen en passen beter binnen het window. Zie hier de oorzaken van de kubusplaag.

Inmiddels is een ontwikkeling te zien naar virtuele kubussen die alleen in geheugen bestaan. Ook worden kubussen opgeslokt door de 'relationele' database. In beide gevallen zijn we gelukkig verlost van de kubusplaag. Wat alleen in geheugen bestaat of onderdeel is van de database, is wat betreft beheer en opbouw een prima oplossing. De technologie is er nu: laten we er gebruik van maken! In ieder geval is het een bezopen idee om de kubus als metafoor te gebruiken. De enige kubus waarvan mensen zich een beeld kunnen maken heeft precies drie dimensies. Leg nu maar eens uit dat het in dit geval gaat om een kubus die zeven dimensies heeft. Heb je daar al een beeld van? Ja? Goed zo. En wat maakt het ook allemaal uit. De gebruiker *gives a rat's ass* (om het maar eens in goed Nederlands te zeggen). Hij is terecht niet geïnteresseerd hoe de data zijn opgeslagen. Natuurlijk wil hij wel weten hoe hij de data kan analyseren, maar is een kubus dan de beste metafoor waar we mee op de proppen kunnen komen? Als het om de presentatie van de ontsluiting gaat zou je hopen dat gebruik wordt gemaakt van inzichten over informatieverwerking door het menselijk brein en de verschillende manieren waarop mensen tot inzichten en beslissingen komen. En laat me een geheim verklappen: een kubus staat dan niet boven aan de lijst.

Stel dat zo'n konijn voor je auto springt en je instinctief uitwijkt. Je ligt helemaal in de kreukels, krimpnd van de pijn. Hoe dankbaar zou je in die situatie zijn als iemand je pijn zou verlichten door je bijvoorbeeld morfine toe te dienen?

Waarschijnlijk zou je zeer dankbaar zijn. Maar als je een jaar later, al helemaal hersteld, als gevolg hiervan verslaafd bent geraakt denk je daar wellicht anders over. Als je in plaats van morfine nu stermodel invult wordt duidelijk dat deze modelle-ringtechniek inmiddels ook in de categorie kubuskonijn valt. Vroeger functioneel, maar inmiddels een pest. Let me explain. Er was eens een relationeel model. Niemand snapte het en leveranciers implementeerden het ook niet in hun database-producten (o.k., ik chargeer een beetje). Er waren mensen die relationele datamodellen probeerden af te stemmen met de business. Dat lukte niet echt, vanwege de complexe verschijningsvorm ervan. Enter Ralph Kimball, die een veel intuïtiever stermodel uit de hoge hoed tovert. Gevolg: het stermodel wordt gehanteerd om de relaties tussen de data vast te leggen, maar ook worden de data zo opgeslagen. Het aantal joins tussen tabellen neemt zodoende af hetgeen tijdwinst en snellere presentatie betekent. Maar sneller en intuïtiever kan inmiddels ook anders – en beter. Wie datamodellen door de business wil laten controleren kan beter gebruik maken van een methode als FCO-IM (of een willekeurige andere NIAM-variant). Je controleert gewone Nederlandse zinnen. Als die goed zijn rolt er een valide model uit. Bij het presenteren van informatie zou ik uitgaan van de mens als een informatie-verwerkend organisme (u kent hem als: gebruiker). Alle inzichten die we de laatste jaren op dat vlak hebben verzameld (lees René Kahn, Malcolm Gladwell etcetera) zijn voor onze sector van het grootste belang. Meer snelheid kan verkregen worden door het inlezen van data in geheugen en door gebruik te maken van snellere en inmiddels veel goedkopere hardware. Stermodellen waren tien jaar geleden dus een goed idee. Maar inmiddels is het een verslaving die betere manieren van besluitvorming en presentatie tegenhoudt. Ik zie Ralph Kimball dan ook als een sneaky versie van Doctor Drugs.

BI en datawarehousing zijn een boeiend onderdeel van IT. Maar af en toe moeten we schoonmaak houden om inmiddels achterhaalde inzichten en aanpakken op te ruimen en in te wisselen voor nieuwe. Over die nieuwe inzichten en aanpakken kunnen we het in de volgende column hebben.

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).