Warren Thornthwaite draagt Kimball's datawarehouse-boodschap uit

"Niet bouwen is een vaak vergeten optie"

Karien Verhagen

In april 2006 verzorgde de Ralph Kimball University voor Quest for Knowledge (Q4K) een cursus Datamodellering in Schiphol-Rijk. Daarmee wordt een lange traditie van data-modelleringscursussen door Kimball c.s., verenigd in de Ralph Kimball University (RKU), voortgezet.

Foto: Harry Otto

Warren Thornthwaite: 'Doornenruimer' ...

Datawarehouse-pionier Ralph Kimball is in de jaren negentig begonnen met de cursus Advanced Datawarehousing en de cursus Datawarehousing In Depth. Daarvan bestaan inmiddels vier varianten. Er is een cursus Datawarehouse Lifecycle in Depth, ETL Architecture in Depth, Dimensional Modeling in Depth en een Microsoft Datawarehouse in Depth. Ralph Kimball heeft daarvoor inmiddels assistentie gekregen van de co-auteurs van de 'Lifecycle Toolkit', Margy Ross en Warren Thornthwaite. Zij geven per jaar ieder zo'n 25 à 30 geplande cursussen, en daarnaast nog heel wat ad hoc en maatwerk cursussen. Database Magazine sprak met Thornthwaite tijdens zijn verblijf in Amsterdam.

What's in a name ?

Thornthwaite is geen alledaagse naam. Enig historisch onderzoek wijst uit dat *twaithe* oud-Engels is voor *opruimen*. Thornthwaite stamt dus waarschijnlijk af van een Engelse familie die beroepshalve de doornstruiken van de akkers verwijderde. Bestaat het vak van BI-expert ook niet uit het ruimen van obstakels?
Thornthwaite bevestigt dat de cursisten aan de lastige theorie uit het boek niet voldoende hebben. "Om te oogsten is het nodig de vele bedreigingen in een BI-traject te trotseren. Wij leren de aan-komende BI-experts om die bedreigingen te herkennen en te vermijden", zegt Thornthwaite. Daarbij kan hij bogen op een jarenlange ervaring – praktijkkennis is voor een docent in dit vak een absolute vereiste.
Datamodellering staat in de cursus nog steeds centraal. Daarnaast wordt er aandacht besteed aan andere vaardigheden, zoals

1. Het vergaren van de Business Requirements;
2. De prioriteit vaststellen van de informatiebehoefte;
3. Het bewaken en bewaren van de link met de business.

De Nederlandse BI-expert zou zo langzamerhand toch wel uitgeleerd moeten zijn. De cursussen vinden nog steeds gretig aftrek. Ook in de verkoop van *The Lifecycle Toolkit* is nog geen dalende trend te ontdekken, het aantal cursussen groeit en er zijn ook nog volop plannen voor nieuwe uitgaven.

Hoe ontstond de 'fit'?

Ralph Kimball, Margy Ross, Laura Reeves en Warren Thornthwaite hebben samen 'The Datawarehouse Lifecycle Toolkit' geschreven, de 'bijbel' voor elke BI-specialist.
Thornthwaite vertelt, dat zij elkaar leerden kennen in de jaren

tachtig bij een bedrijf met de naam Metaphor. "We bouwden daar datawarehouses *avant la lettre*. Eigenlijk waren dat rapportage-databases met daarop management-rapporten en analyse-omgevingen, die toen nog *decision support systems* heetten. Later werden dat Executive Information Systems (EIS) en nu heten dat dashboards of cockpits. De vorm is anders, de functionaliteit nog altijd hetzelfde.

De technische complicaties van near real-time datawarehousing worden onderschat

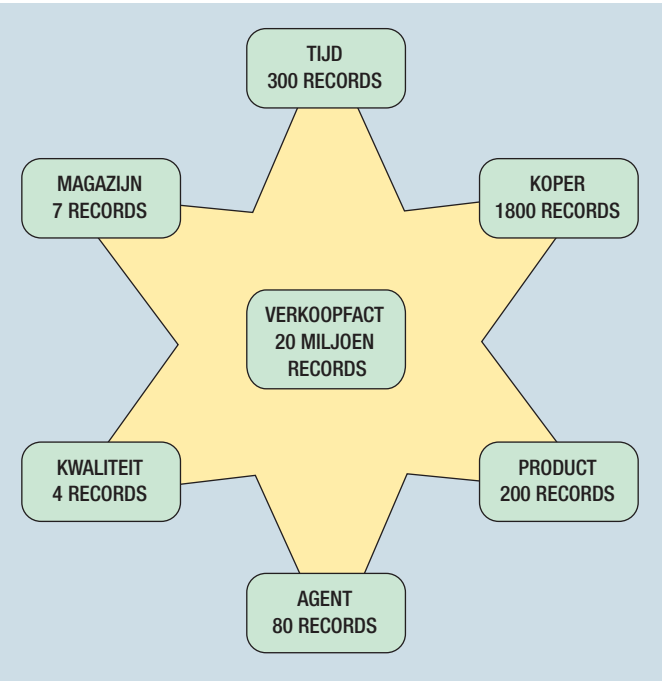
Bij Metaphor werd het stermodel of *starscheme* geboren. Het is een volkomen natuurlijke opslagstructuur voor analyses op geconsolideerde gegevens. De dimensies met de selectiecriteria rond een vette fact table met de numerieke waarden werden toentertijd op vele plaatsen *ontdekt*", verhaalt Thornthwaite. Ralph Kimball heeft die ontdekking commercieel geëxploiteerd in het product Redbrick, een database-engine gebaseerd op ster-schema's. Redbrick is nu eigendom van IBM.
Ook de starjoin index was zo'n *ontdekte* praktische oplossing. Thornthwaite: "Bij het creëren van analyse-omgevingen blijkt dat een RDBMS toch in de eerste plaats gemaakt is om transactionele systemen te ondersteunen. Een query die de verkopen van product X in de maanden maart tot en met mei moest selecteren, koos daarom vaak het verkeerde toegangspad. De query selecteerde eerst één productrecord en vervolgens zes miljoen verkoop-records. De join met de tijdtabel leverde van die selectie van zes miljoen vervolgens 20.000 records op die voldeden aan het criterium maart, april of mei!" In afbeelding 1 is dit voorbeeld uitgewerkt.

```
Select    productnaam,
          maandnaam,
          aantal,
          bedrag
From      verkoopfact,
          tijd,
          product
Where     productnaam = 'X' and product.
          productkey = verkoopfact.productkey
And       tijd.maandnummer <= 5 and tijd.
          maandnummer >= 3 and verkoopfact.
          tijdkey = tijd.tijdkey
```

Thornthwaite vervolgt: "Zelfs indexen hielpen niet, omdat bijvoorbeeld de query optimizers vervolgens toch gewoon besloten door het verkoop-fact heen naar de tijdtabel te gaan. Dat was destijds de reden om een apart soort index te maken. De starjoin index selecteerde eerst het product en de 92 tijdrecords en verenigde met die selectie pas als laatste de grote fact-tabel om 20.000 records op te halen, in plaats van zes miljoen. De starjoin index werkt altijd van buiten naar binnen en benadert als laatste de fact-tabel.
Als je met elkaar het sterschema hebt ontdekt, *herken* je feiten en dimensies. Dat is de essentie van dimensioneel modelleren. Dat onderscheid is echter lastig uit te leggen, want facts en dimensies lijken soms veel op elkaar. Zo is het ontstaan van een historische *occurence* in een *slowly changing dimension* tabel eigenlijk ook een fact. Een klant die verhuist, veroorzaakt een nieuwe historische klantversie. Maar die verhuizing kan ook een fact zijn als het voor de bedrijfsvoering maar belangrijk genoeg is. Voor een verhuisbedrijf is dat zelfs het primaire fact", zo stelt Thornthwaite. "Een ander voorbeeld is de wijziging van een verzekeringspolis. Voor een schadeclaim is 'polis' een dimensie. Voor het evalueren van de bedrijfsvoering van een intermediair ligt dat anders: het aanmaken of schrappen van een polis is voor een agent een transactie, een fact. In een 'conformed dimensions' model kun je dus de polis zowel als fact (polis ontstaat, wijzigt of wordt geschrapt) als ook als dimensie (kenmerk van schadeclaim) tegenkomen."

De RKU heeft zijn ziel verkocht?

Thornthwaite vindt niet dat hij met het uitbrengen van de Microsoft Lifecycle Toolkit zijn ziel heeft verkocht. De oplossingen die 'The Lifecycle Toolkit' biedt zijn redelijk uitputtend, maar een oplossing met een specifiek alles dekkend BI-tool helpt de lezer een stuk verder op weg met de implementatie van de BI-infra-structuur. "Dat moet dan wel een tool zijn dat de totale lifecycle



Afbeelding 1: Sterschema met 20 miljoen records.

Foto: Harry Otto

"Om te oogsten is het nodig de vele bedreigingen in een BI-traject te trotseren."

De doornstruiken

In 2002 gaf Thornthwaite een interview met Freek Kamst voor Database Magazine. Daarin schetste hij de drie belangrijke valkuilen voor BI-projecten :

1. Het bouwen van een datawarehouse zonder de eisen en wensen te beschouwen vanuit de business;
2. Het ontbreken van een sponsor;
3. Het bouwen van een datawarehouse zonder gebruikersfocus.

"Daar is eigenlijk weinig verandering in gekomen. Nog steeds komt het voor dat de technologie achter het stuur zit", zegt Thornthwaite desgevraagd. "Een voorbeeld is near real-time datawarehousing. BI-consultants zouden wel wat terughoudender mogen zijn als ze zich de technische complicaties in hun volle omvang zouden realiseren. Er is in een near real-time datawarehouse bijvoorbeeld geen tijd om de vaak uitgebreide T (Transform) van het ETL-traject te doen. Meestal leidt dat tot een fysieke scheiding van de near real-time laag en de historische (datawarehouse) laag. Dat kan vervolgens weer inconsistenties geven met de referentiedata, omdat het datawarehouse de actuele referentiegegevens pas later – in de batchload – krijgt.
De koppeling tussen beide lagen gebeurt bovendien op grond van de logische referentiesleutel uit het operationeel systeem en niet op basis van de *surrogate key*. Die surrogate key ontstaat immers pas in het datawarehouse. Voordeel van een near real-time datawarehouse is natuurlijk wel dat de batchload vervolgens een beroep kan doen op de near real-time tussenlaag. Het ETL-proces hoeft voor de batchload het operationeel systeem niet meer lastig te vallen." De technische complicaties van near real-time datawarehousing worden volgens Thornthwaite onderschat en de baten overdreven. Hij vindt die ontwikkelingen dan ook gevaarlijk en vooral *vendor driven*.

en het datawarehouse productieproces dekt. Microsoft is zo'n tool, Oracle en IBM zijn ook zulke tools. 'The Oracle Lifecycle Toolkit' zit dan ook in de planning en de 'IBM Lifecycle Toolkit' volgt. Ook de leveranciers van Business Objects en Cognos kunnen uitzien naar hun eigen Lifecycle uitgave", kondigt hij aan.
De originele 'Datawarehouse Lifecycle Toolkit' beleeft intussen zijn zoveelste druk en is toch enigszins gedateerd. Gevraagd naar zijn bijdrage, wijst Thornthwaite met name op de infrastructurele onderdelen: de ETL-techniek, de architectuur en BI Applications. "Anno 2006 zou ik in het boek andere accenten aanbrengen. De oplossingen zijn nu anders. De portal-techniek zou nu veel meer aandacht krijgen. De kracht en de prijs van de tegenwoordige hardware en software maakt bijvoorbeeld ook de aggregaten vrijwel overbodig. In de oude versie wordt 64-bit computing nog niet vermeld. De techniek maakt andere oplossingen mogelijk, maar mag niet de enige reden zijn om een oplossing aan te bieden. Dat is nog steeds een bekende valkuil. Daarnaast zijn er andere risico's."

> De starjoin index werkt altijd van buiten naar binnen en benadert als laatste de fact-tabel

Andere complicaties in een BI-traject kunnen zich voordoen wanneer de bouwers van een datawarehouse worden geconfronteerd met de cultuur van een bedrijf. Kort geleden werd Thornthwaite advies gevraagd voor de inrichting van een BI-architectuur. Na enige gesprekken was het hem duidelijk hoe het bedrijf bestuurd werd. "Voorstellen, hoe kundig ook onderbouwd met kwalitatief goed cijfermatig materiaal, werden door de CEO genegeerd.
De CEO nam in feite in zijn eentje alle besluiten. Hij negeerde daarbij alles wat hem door de cijfers en de onderbouwing daarvan werd aangereikt. Het geplande BI-traject moest nu juist die cijfers

volgens uniforme definities in chique rapportages oprollen. Tenzij de CEO zijn werkwijze aanpast heeft dit BI-traject geen enkele zin. Dat moet dan ook aan de kaak worden gesteld. Het vervolg is moeilijk voorspelbaar en zeker niet de verantwoordelijkheid van de BI-consultant. Die moet de situatie op tijd signaleren en vervolgens een stap opzij doen. Vergeet niet: het niet bouwen is een vaak vergeten optie."

De lijnen naar de business kunnen ook bewaakt worden door een DWH team

De aard van en de hoeveelheid risico's stellen hoge eisen aan een BI-expert. "Dat uit zich ook in het salaris", constateert Thornthwaite. "Een BI-expert is van alle markten thuis en heeft een breed pakket aan vaardigheden: hij of zij moet een vaardig technicus zijn, kennishebber van de business en weten waar het in de onderhavige business om gaat. Bovendien moet het een vaardig prater zijn die het verwachtingspatroon bij alle betrokkenen helder krijgt. De BI-expert is dan ook een duur betaalde specialist."

En als het ontwerp gebouwd is ...

"Dan is er nog de zorg voor de continuïteit. Een datawarehouse is nooit af, het is een dynamische bedrijfsinfrastructuur. Die dynamiek is essentieel. Daar horen bijvoorbeeld ook Public Relations bij: potentiële gebruikers enthousiasmeren, pronken met de behaalde resultaten en de mooie applicaties verhogen het gebruik en daarmee het rendement. Technici zijn geneigd dat aspect te vergeten. De betrokkenheid van account managers en business-kenners moet de continuïteit waarborgen. Dat kan in een Business Intelligence Competency Center (BICC), maar zo'n team kan ook anders heten. De lijnen naar de business kunnen ook bewaakt worden door een DWH team of een (key) User forum. Er zijn ook bedrijven waar het informatie-management die rol vervult. De functie goed beleggen is belangrijker dan hoe zij genoemd wordt, want ... What's in a name?", besluit Warren Thornthwaite.

Drs. C.W.J. Verhagen is senior BI-consultant bij 4BIS Scholing en advies.

Update

Autonomy en Cognos bundelen krachten voor BI

Autonomy en Cognos hebben een integratie gerealiseerd waarmee, met behulp van Autonomy's retrievalsoftware, resultaten uit ongestructureerde informatie en gestructureerde data voor Business Intelligence-toepassingen samengevoegd worden. Deze integratie biedt gebruikers de mogelijkheid om klantcommunicatie te analyseren en vervolgens hanteerbaar te maken voor BI software.
Gebruikers van Cognos verzamelen dagelijks ongestructureerde informatie, zoals klantreacties, klachten, telefoongesprekken, e-mails, digitale documenten en andere bronnen die gewoonlijk niet in traditionele databases opgeslagen worden.
Autonomy is expert op het gebied van het automatisch beheren, verwerken en beschikbaar stellen van de immer groeiende hoeveelheid ongestructureerde data (zowel tekstuele documenten als beeld- en geluidsmateriaal). De integratie tussen Autonomy en Cognos 8 BI biedt onmiddellijk inzicht in en analyse van bedrijfskritische informatie.
Met Cognos Go! kunnen gebruikers concept-base zoekopdrachten uitvoeren om snel toegang te krijgen tot de meest relevante bedrijfsinformatie. Zakelijke beslissingen kunnen sneller en nauwkeurig genomen worden doordat het systeem de belangrijkste onderwerpen uit de verschillende informatiebronnen onttrekt en gebruikers alarmeert wanneer nieuwe informatie beschikbaar is.
Zie www.cognos.nl en www.autonomy.com

MicroStrategy brengt BI en Enterprise Information Integration samen

MicroStrategy, leverancier van Business Intelligence software kondigt aan dat IBM WebSphere Information Integrator door MicroStrategy is gecertificeerd om te opereren met MicroStrategy 8. Deze nieuwe integratiemogelijkheid brengt de nieuwste versies van de twee technologieën samen en kan worden gezien als een 'milestone' in de langlopende relatie tussen IBM en MicroStrategy.
De combinatie van MicroStrategy en IBM WebSphere Information Integrator technologie maakt het voor bedrijven mogelijk om naadloos zakelijke inzichten te distilleren uit verscheidene, heterogene databronnen. MicroStrategy's geïntegreerde rapportage, analyse en monitoring software staat organisaties toe om direct meerdere databases binnen te treden en helderheid te creëren in operationele situaties door gedetailleerde informatie te onthullen.
IBM WebSphere Information Integrator stelt gebruikers in staat toegang te verkrijgen tot diverse verspreide databronnen, gebruik makend van slechts één enkele SQL query. Dit levert klanten een geconsolideerde blik op de informatie uit separate bronnen, inclusief gedistribueerde en mainframe databases, bedrijfsbrede applicaties, XML-documenten, bestanden en content repository's.
Meer informatie: www.microstrategy.com