Data Vault concept past in DW2.0 oplossing

Dan Linstedt in Amerongen

Karien Verhagen

De Data Vault techniek staat in de belangstelling. Een cursus Data Vault, gegeven door Dan Linstedt zelf werd druk bezocht. Circa vijftig deelnemers uit vier landen bezochten Kasteel Amerongen om les te krijgen van de meester zelf.

De driedaagse cursus werd georganiseerd door DNV (voorheen CIBIT) en kon worden afgesloten met het examen *Gecertificeerd Data Vault Ontwikkelaar*. Het is zelfs al een zelfstandig naamwoord geworden. De Data Vault is populair. De volgende cursus door Dan Linstedt staat alweer ingepland voor 27 oktober. Ook in andere landen in Europa zoals Duitsland, Frankrijk en Zweden zal DNV cursussen gaan organiseren.
Database Magazine ging naar Amerongen om Dan Linstedt te ondervragen en de roots van de Data Vault bloot te leggen.

Waar komt de naam 'Data Vault' vandaan?
"Puur toeval. Een avondje namen verzinnen met een aantal kennissen en collega's leverde weinig op, tot iemand zei: 'Waarom noem je het niet een Data Vault , een mooie plek om gegevens op te bergen?' Achteraf ben ik niet zo blij met de associatie met een kluis die de naam Data Vault heeft, terwijl het model toch juist bedoeld is om informatie te ontsluiten.

> Introductie van een Data Vault EDWH kan het best gebeuren door een proof of concept

Ik zou het nu niet meer zo noemen. De oorspronkelijke naam was Common Foundational Warehouse Modeling Architecture. Daarmee konden we natuurlijk ook niet voor de dag komen."

Wie is Dan Linstedt?
"Ik zie mezelf graag als een integer persoon. Ik heb het geluk dat er zoveel grootheden baanbrekend werk hebben verricht. Mijn theorieën leunen zwaar op de kunde van mensen als Bill Inmon, Ralph Kimball, Claudia Imhoff en vele anderen. Daarnaast heeft mijn werk als system administrator mij de noodzakelijke praktijkervaring gegeven. Ik heb een graad in Computer Science en mijn achtergrond ligt in de systeemarchitectuur. Ik heb informatiesystemen gebouwd en ontworpen en vooral ook veel integratieproblemen opgelost."

Hoe verklaart u die plotselinge populariteit van een methode die in zijn huidige vorm toch al acht jaar bestaat?
"Toen het datawarehouse als tussenlaag een noodzaak bleek tussen de bronnen en de datamarts heeft Bill Inmon daarvoor het traditionele klassieke relationele model genomen en daarop optimalisaties aangebracht. Dr. Kimball bedacht de 'conformed dimensions' als aanpassing van het stermodel. Beide modellen zijn echter compromissen. Ze zijn niet bedacht vanuit de functie die een Enterprise Datawarehouse heeft. Dat is de Data Vault wel."

Wat mankeerde er dan aan die twee oplossingen? Wat waren de evidente tekortkomingen in de bestaande modellen?
"Ik ben begin jaren negentig begonnen met het splitsen van entiteiten in allerlei types. Voor elk soort informatie een entiteittype. In het begin waren dat er wel vijftig, in 1995 was het al teruggebracht naar tien en vanaf 1997 kwam ik uit op de drie huidige types, de hubs, de link-tabellen en de satelliettabellen. Dat idee is geïnspireerd op de gedachte van *Natural World Models*. Hoe werken structuren in de natuur en wat is hun functie in de omgeving? In 1999 ben ik de drie types in de Data Vault methode gaan propageren als oplossing voor een Enterprise Datawarehouse. In een EDWH heeft een identity (hub) een andere functie dan de attributen (satellieten) en de relaties (links). Een verkoop is bijvoorbeeld een link tussen een product en een klant en kan zonder die twee niet bestaan.
Een ander principe is dat het EDWH een volledige audit moge-

Foto: Ronald Damhof.

Dan Linstedt tijdens de signeersessie.

lijk moet maken. De facts worden opgeslagen als een exacte weergave van wat er in de transactionele systemen is gebeurd. Het EDWH geeft dus het System of Record authentiek weer en gooit niets weg. Je hoort steeds vaker dat een bedrijf er naar moet streven om *one version of the truth* vast te leggen, maar dat is een illusie. Er bestaat niet one version of the truth.
Elke business heeft zijn eigen waarheid. Het is aan de business te bepalen wat hun waarheid is. IT kan dat niet doen, het EDWH kan dat niet doen. Business rules controles vinden daarom plaats tussen het EDWH en de datamart. Elke interpretatie in de ETL maar ook een eenvoudige ontdubbelingsprocedure vanuit de staging area naar het datawarehouse doet de traceerbaarheid geweld aan."

System of Record (SOR) is een moeilijke term, wat wordt daar precies mee bedoeld?
"Het is inderdaad een lastige uitdrukking. Voor technische mensen is een goede definitie van SOR even moeilijk als de definitie van *klant* voor de business. Met SOR worden verschillende zaken bedoeld. Voor een operationeel systeem is dat het punt waar de gegevens ontstaan. Maar als er gegevens ontstaan in een Operational Datawarehouse (ODW) dan ligt het System of Record daar. Het SOR kan ook in de datamart liggen.
Het resultaat van alle business rules, de condities, de integratie en als gevolg daarvan de samenstelling van een nieuw gegeven in een rapportage is ook een SOR. Als je resultaten in een financiële rapportage toont aan de business dan worden daar ook gegevens gecreëerd. Een rapportage kan een financiële view zijn op de business, die gebruikt wordt voor allerlei doelen."

De Data Vault is dus ook een oplossing voor de belangrijke nieuwe eisen die gesteld worden aan compliance?
"Zeker, deze nieuwe eis aan het EDWH heeft recentelijk een enorm gewicht gekregen. Het is ook de reden dat het Inmon concept wel en het concept van Kimball niet past op de Data Vault. In een *conformed dimension datamart verzameling* ligt de nadruk op de toegankelijkheid voor de eindgebruiker, maar je verliest daardoor wel de transparantie naar de bronnen. Voor compliance en audit-mogelijkheden is die transparantie essentieel.
Formaatfouten kunnen vanuit de staging in het EDWH wel in aparte tabellen worden gestopt die later als *error datamarts* kunnen worden geraadpleegd. Maar fouten tegen de business rules worden er in het EDWH niet uitgefilterd. Dat gebeurt pas in het proces van EDWH richting datamart. Als er drie adressen van dezelfde klant zijn, worden ze alle drie met bronvermelding opgenomen."

In een Data Vault model wordt het EDWH log en groot en is het niet toegankelijk voor eindgebruikers. Het bevat geen optimalisaties om er snel datamarts uit te kunnen halen. Maakt dat het laden van de datamarts niet erg ingewikkeld en traag?
"Nee en ja. Eindgebruikers kunnen niet bij het EDWH maar je kunt wel optimalisaties hebben in de vorm van point-in-time tabellen, dat zijn snapshot-tabellen die de toestand op een bepaald moment vastleggen of bridge-tabellen, tabellen die de vele joins platslaan en ook berekende procesvariabelen kunnen bevatten. Verder is er plaats voor *user grouping sets*, gegevensgroeperingen die speciaal gemaakt zijn voor een gebruiker of groep van gebruikers.
Onder datamarts verstaan wij trouwens alle voor de eindgebruiker toegankelijke datasets, dus ook datamining sets."

Het is aan de business te bepalen wat hun waarheid is

Er zijn twee stromingen. Eén stroming gaat in de richting van steeds grotere BI-architecturen met steeds meer lagen in en naast de EDW database. Daarnaast worden tools als Qlikview populair die de noodzaak voor een datawarehouse bijna ontkennen en de data desnoods 'on the fly' uit de operationele systemen halen. De opslag wordt dan snel in bijvoorbeeld een in-memory database geplaatst. Dan zijn er dus eigenlijk geen tussenlagen meer. Hoe kijkt u daar tegenaan?
"Er is zeker een plaats in de markt voor Qlikview, maar naar mijn mening is het een datamart tool, geen DWH tool. Het kent grenzen; opslaggrenzen, grenzen aan het historisch en auditeerbaar inzicht dat je kunt bieden. Wanneer we voorzien dat klanten tegen die grenzen gaan oplopen proberen we ze op te voeden, want als ze later tegen die grenzen oplopen kan een reconstructie inderdaad kostbaar zijn. Qlikview blijft dan natuurlijk nog steeds bruikbaar als datamart tool.
Vaak is er inderdaad al een groot aantal datawarehouses. Dat zijn dan meestal *stovepipe datamarts*; losse oplossingen voor gescheiden groepen gebruikers. Om daar iets aan te doen moet de business 'pijn' voelen. Soms voelt men een noodzaak voor auditeerbaarheid, voor kostenreductie, voor een geïntegreerde

centrale omgeving. De introductie van een Data Vault EDWH kan dan het best gebeuren door een *proof of concept*. Je begint met een paar tabellen. Werkt het, wordt de pijn minder, dan ga je er mee door, werkt het niet dan laat je alles bij het oude."

Het scheiden van een identiteit van haar attributen en relaties klinkt goed. Het zou misschien ook wel passen in een Master Data Management oplossing. Waarom zou de Data Vault niet ook geschikt zijn voor een transactioneel systeem?
"We sluiten dat niet uit. Dat zal de toekomst uitwijzen. We dragen die mogelijkheid echter (nog) niet actief uit. Als het gebeurt zullen we ons daar niet tegen verzetten. Van de vijftig implementaties in de wereld van Data Vault ken ik er drie die betrekking hebben op een transactioneel systeem. Het werkt daar wel. Er zijn ook bewegingen om het Operationeel Datawarehouse te integreren met het transactioneel systeem. Toch zijn kenmerken als de auditeerbaarheid en de flexibiliteit vooral bepaald vanuit de problematiek van Enterprise Datawarehousing."

Genesee Academy heeft een ideaal

Als Dan Linstedt spreekt over ons dan bedoelt hij zichzelf, zijn partner Hans Hultgren, ook aanwezig in Amerongen en Kent Graziano, samen de schrijvers van 'The New Business Supermodel, de Business of Data Vault Modeling' en drie van de twaalf business partners van Genesee Academy. Het concept is niet gepatenteerd. Doel is het verkopen van consultancy en opleidingen. Genesee Academy wil de methode propageren, ontwikkelaars certificeren en een correcte, consistente invulling ondersteunen, zodat de voordelen van de Data Vault maximaal tot hun recht komen. Dat klinkt als zuivere wetenschap en het najagen van idealen, maar de academie verkoopt toch ook een engine die Data Vault modellen genereert, de RapidACE? Hans Hultgren ziet die engine niet als een product dat zo vaak mogelijk verkocht moet worden. Hij noemt het een DW2.0 Enabler, een product dat je ondersteunt bij de inrichting van DW2.0 van Bill Inmon. Het helpt je de DW2.0 architectuur beter en sneller in te richten. Het Data Vault concept past prima in de DW2.0 oplossing zoals Bill Inmon die voorstelt. Genesee Academy werkt ook samen met Inmon en Imhoff en probeert naast het EDWH ook invulling te geven aan goede metadata- en masterdata managementoplossingen. Ook het integreren van ongestructureerde data en real-time oplossingen zijn belangrijke onderdelen van het DW2.0 concept. Er zijn voldoende uitdagingen en de ambities strekken ver. Het doel is uiteindelijk om de internationale community te helpen met opleidingen, het verzamelen van best practices en het waken over de consistente invulling van het totale DW2.0 concept. Doel is dus toch wetenschap bedrijven zelfs als de partners er toevallig rijk van zouden worden. Liever fans dan omzet? "Het zit elkaar niet in de weg", geeft Hans Hultgren toe.

Karien Verhagen
Drs. K.Verhagen is senior BI consultant bij Getronics PinkRoccade en 4BIS Scholing en advies.

Update

Microsoft wil appliance DATAllegro overnemen

Microsoft meldt dat het van plan is DATAllegro, leverancier van datawarehouse appliances, over te nemen.
De overname is een uitbreiding op Microsoft's mission-critical dataplatform. Over de hoogte van het overnamebedrag is niets bekendgemaakt.
In tegenstelling tot andere leveranciers van datawarehouse-appliances die zich richten op omgevingen met 1 tot 25 TB aan data, is DATAllegro gespecialiseerd in grotere high-performance datawarehouses, tot honderden TB's in één systeem. Klanten van DATAllegro zitten vooral in de retail, telecommunicatie en fabricage.
Naast dat met zeer grote datavolumes om kan worden gegaan, is DATAllegro's technologie (waarop patent is aangevraagd) ontworpen voor complexe workloads zoals tijdens hoge pieken en gemengde query's. DATAllegro is een van de weinige datawarehouse appliances die gebouwd is op een non-proprietary hardwareplatform (Dell en Bull servers en EMC storage). Deze flexibele architectuur maakt het ideaal voor integratie met SQL Server.

Na afronding van de overname, zal Microsoft bijna alle DATAllegro-medewerkers inlijven, en blijft het hoofdkantoor in Aliso Viejo (Californië) voortbestaan als Center of Excellence voor datawarehousing. Uiteraard wordt de support aan bestaande klanten voortgezet.

SAS neemt software-bedrijf IDeaS over

Door de overname kan SAS naast zijn Business Intelligence-, analytische, en industriespecifieke oplossingen, nu ook revenue management-applicaties specifiek voor de reis- en hospitality-branche aanbieden. Deze oplossingen zijn bovendien complementair aan SAS' retail revenue optimization software. Klanten van IDeas profiteren op hun beurt van de wereldwijde expertise van SAS.

IDeaS wordt met zijn 200 medewerkers een volledige dochtermaatschappij van SAS. Het huidige management van IDeaS blijft hierbij gehandhaafd.
Wereldwijde klanten van IDeaS zijn onder meer Hilton, Hyatt International, Mandarin Oriental en Intercontinental Hotels Group.
Zowel SAS als IDeaS zijn private ondernemingen, financiële details rond de overname worden niet bekend gemaakt.