Goede techniek om DWH tot standaardapplicatie te maken

Datawarehouse-modelleren met Data Vault

Maarten Ketelaars

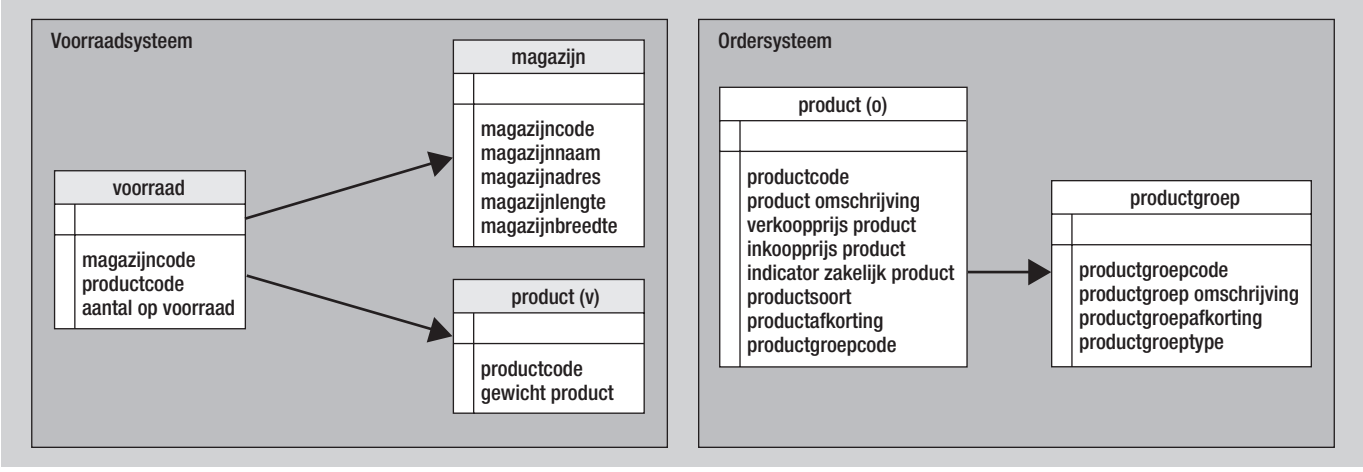
In 2002 introduceerde Dan Linstedt zijn nieuwe 'revolutionaire' manier van datawarehouse-modelleren: Data Vault. Naast de benadering van Bill Inmon en het dimensioneel modelleren van Ralph Kimball is Data Vault een derde aanpak. Wat houdt Data Vault in en hoe is dit te positioneren ten opzichte van de benaderingen van Inmon en Kimball? Is Data Vault echt iets anders of betreft het oude wijn in nieuwe zakken? In dit artikel wordt aangegeven wat Data Vault is, waar Data Vault te positioneren is ten opzichte van de twee andere benaderingen en wat de toegevoegde waarde van deze benadering is.

Dan Linstedt heeft sinds 2003 in een reeks van artikelen in The Data Administration Newsletter, een introductie gegeven over Data Vault. Data Vault wordt door Linstedt als volgt gepositioneerd: "Een detail-georiënteerde, historie-tracerende en uniek gelinkte verzameling genormaliseerde tabellen, die meerdere functionele business-domeinen ondersteunt. Het is een hybride aanpak, die het beste van 3NF en dimensioneel modelleren combineert. Het ontwerp is flexibel, schaalbaar, consistent en aanpasbaar aan de behoefte van een onderneming. Het is een datamodel, dat specifiek is ontworpen, om aan de eisen van een enterprise datawarehouse te voldoen."

Volgens Linstedt combineert de Data Vault aanpak de beste facetten van 3rd normaalvorm modelleren met dimensioneel modelleren. Maar maakt Data Vault deze bewering inderdaad waar? Om dit te bepalen wordt eerst ingegaan op de verschillende componenten, waaruit een Data Vault model is opgebouwd en wordt een voorbeeld uitgewerkt. Daarna wordt deze benadering vergeleken met de benadering van Inmon en het dimensioneel modelleren.
Tot slot wordt een aantal sterke en zwakke punten aangegeven. Dan zal duidelijk worden of het iets speciaals is, zoals Linstedt zegt, of dat het misschien wel letterlijk een gegevenskluis is: goed beveiligd, maar alleen voor de echte specialisten is er iets uit te halen.

Componenten

Er zijn drie componenten gedefinieerd in een Data Vault model: Hubs, Links en Satellieten.
Een ontwerp volgens de Data Vault-aanpak is geconcentreerd rondom de business-entiteiten, waarbij de 'Hub' de identificerende velden van een business-entiteit representeert. In de 'Links' worden de relaties tussen de Hubs gemodelleerd. Enerzijds is dit voor foreign key-relatie uit een bronsysteem. Anderzijds is dit voor transacties, geïdentificeerd door meerdere business-sleutels. In de 'Satellieten' worden alle relevante overige gegevens vastgelegd. Zowel een hub als link zal minimaal één, maar vaak meerdere satellieten om zich heen hebben.



Afbeelding 1: Brontabellen.

Hub-entiteiten ('Hubs')
Elke 'hub' vertegenwoordigt een business-entiteit. Deze vinden we terug als tabel, waarin de 'business-sleutel' (ook wel bekend als 'functionele sleutel', 'bronsleutel' of 'native key') aanwezig is. Dit kunnen zowel enkelvoudige als meervoudige sleutels zijn. Voorbeelden hiervan zijn productnummer, klantnummer en medewerkernummer; nummers die bij één of meerdere gebruikersgroepen bekend zijn en/of die men in een bron als unieke sleutel gebruikt.
Daarnaast wordt standaard per business-sleutel een surrogaat-sleutel gegenereerd. Deze surrogaatsleutel wordt in het gehele verdere model gebruikt. De hub-tabel is daarmee de enige plek, waar de business-sleutels opgeslagen zijn.

Link-entiteiten ('Links')
Elke 'link' vertegenwoordigt een relatie tussen twee of meer hubs. De link bestaat uit de surrogaatsleutels van de betreffende hubs. Er zijn twee belangrijke vormen:
- De 'relatie-link' representeert een relatie tussen twee hubs. Voorbeelden vanuit de bronsystemen zijn vreemde sleutels (foreign key's) en specifieke koppeltabellen. In het geval van foreign key-relaties betreft het altijd een link tussen twee hubs (zie link product/productgroep uit afbeelding 2). In het geval van specifieke koppeltabellen kunnen het ook meer dan twee hubs zijn.
- De 'transactie-link' respresenteert een specifieke transactie, zoals een aankooptransactie of een productietransactie. Een aankooptransactie wordt bijvoorbeeld gekenmerkt door een product, een klant, een locatie en een medewerker. Deze linktabel bevat van elke gerelateerde hub de betreffende surrogaatsleutel.

> Iedere datawarehouse-architect zou zich op de hoogte moeten stellen van de Data Vault-benadering

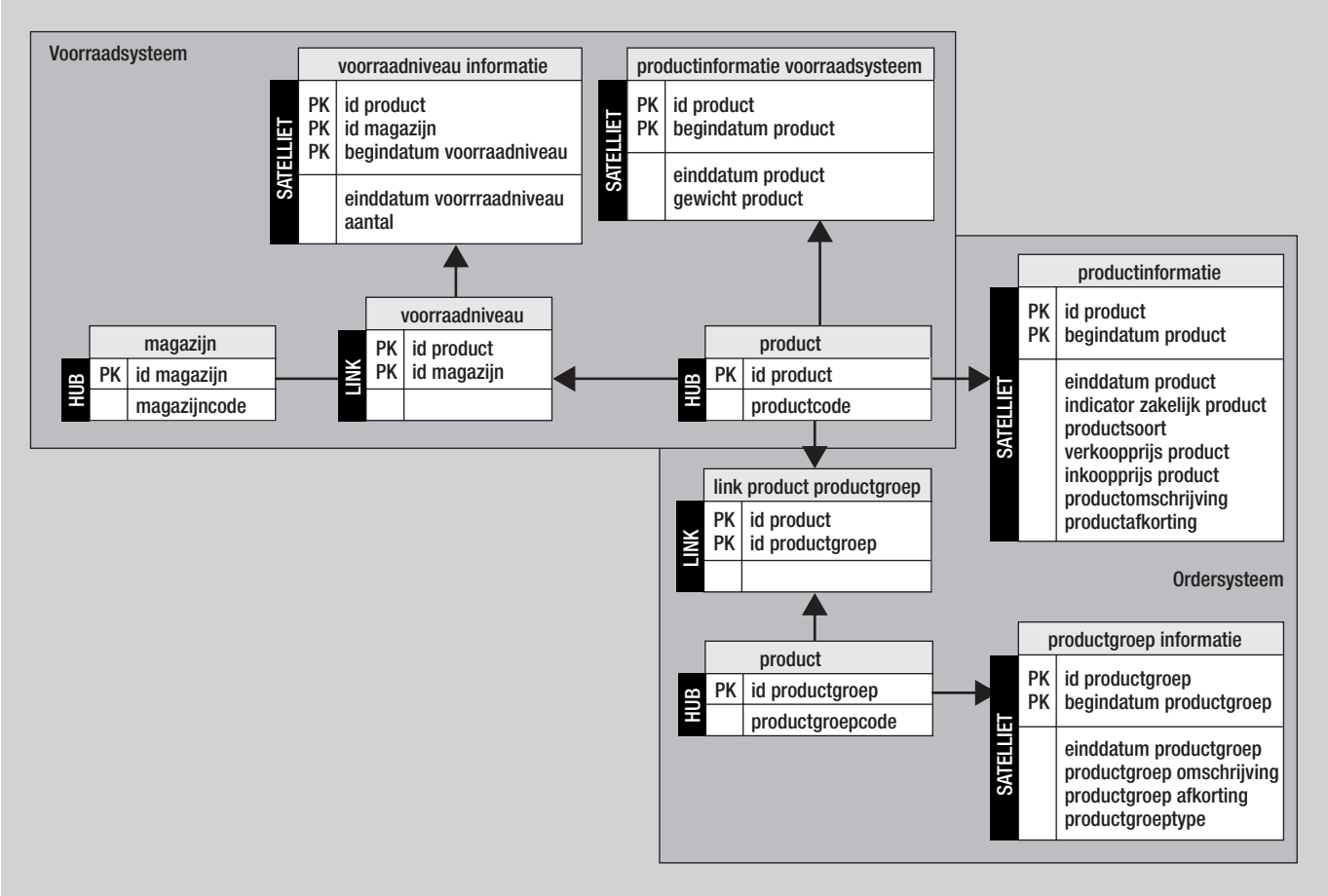
Satelliet-entiteiten ('Satellieten')
De 'satellieten' zijn de tabellen, waarin alle zogenaamde context-gegevens worden opgeslagen. Alle gegevens, die in de tijd kunnen wijzigen, komen terecht in een satelliet. Daarop is deze structuur berekend. Invulling gebeurt door een datumtijd- of datumveld. De surrogaatsleutels met dit datumtijd- of datumveld zijn de sleutel van de satelliet.
Er zijn twee soorten contextgegevens, die kunnen worden opgeslagen: de contextgegevens van een hub en de context-gegevens van een link.
Contextgegevens van een hub: hierin staan alle relevante gegevens van de met de hub corresponderende business-entiteit opgeslagen, met uitzondering van business-sleutel en verwijs-sleutels, die al in hub- en linktabellen zijn ondergebracht.

Standaardiseren

Bij SNS bank is sinds 2003 ervaring opgedaan met Data Vault modelleren. Dit was de invulling van het datawarehouse als centrale integratielaag, waarvandaan datamarts worden afgeleid. Iedere datamart heeft daarbij een specifiek doel voor een specifieke groep gebruikers. Na de eerste oplevering werd al snel duidelijk dat deze manier van modelleren een aantal prettige voordelen had. Het was zeer goed mogelijk om een aantal stappen in het specificatie- en realisatietraject te standaardiseren. Dit leidde uiteindelijk tot een nieuw traject, waarbij zowel de Data Vault-modellen als de ETL worden gegenereerd. "Dit levert een enorme besparing op in de realisatie van het Corporate Datawarehouse en het aantal fouten blijft, door de hoge mate van codegeneratie en de volledig gestandaardiseerde aanpak, tot een minimum beperkt", stelt John Hendriks van SNS bank.
In dit kader biedt Data Vault een opstap naar een hoger niveau. Het is een goede techniek om een DWH tot een standaardapplicatie te maken, vergelijkbaar wat eerder ERP heeft gedaan voor de transactionele systemen. BIReady is een product dat hieraan invulling geeft. Het is een metadata-gedreven datawarehouse-generator. Data Vault is één van de principes die de makers heeft geïnspireerd tot het maken van dit product. Het is interessant om te kijken of dit initiatief tot een nieuwe standaard verwordt in DWH-land. Eerder heeft Kalido dit ook geprobeerd, maar het aantal implementaties is beperkt gebleven.

Typische voorbeelden zijn NAW-gegevens van een klant, productkarakteristieken of kenmerken van een verkoopkantoor. De satelliet is in dit geval sterk gerelateerd aan het type 2 'slowly changing dimension', zoals dat genoemd wordt bij het dimensioneel modelleren.
Contextinformatie van een link: van een transactie kunnen alle transactiespecifieke gegevens in de satelliet opgeslagen worden, zoals transactiedatum, aantal, prijs, korting etcetera. In dit geval moeten tenminste alle surrogaatsleutels van de link en de transactiedatum worden opgeslagen. De satelliet is in dit geval sterk gerelateerd aan de feitentabel in het dimensioneel modelleren.

Praktijkvoorbeeld
Aan de hand van een praktijkvoorbeeld van een logistiek bedrijf wordt een eenvoudig Data Vault-model uitgewerkt. Het bedrijf heeft een ordersysteem, waarin tevens de productadministratie wordt bijgehouden. Alleen deze productadministratie is voor dit voorbeeld relevant. Daarnaast is er een logistiek systeem, waar voorraden per product per magazijn worden bijgehouden, met een eigen beperkte opslag van producteigenschappen.
Mogelijke vraagstellingen, die vanuit logistiek oogpunt gesteld kunnen worden:
- wat is de waarde van alle producten in een magazijn;
- wat is het voorraadniveau per product en productgroep?



Afbeelding 2: Data Vault-model.

Om deze vragen te kunnen beantwoorden, moeten minimaal de tabellen voorraadniveau en product uit het voorraadsysteem en de tabellen product en productgroep uit het ordersysteem gevuld worden. Dit levert samen het Data Vault-model uit afbeelding 2. Daarnaast wordt dan een datamart gerealiseerd voor bovengenoemde vraagstelling, die de basisinformatie uit het Data Vault-model haalt.
De manier waarop de tabellen worden gevuld uit het bronsysteem staat in het overzicht in afbeelding 3. De belangrijkste kenmerken:

- alleen de hub-tabellen worden door meerdere brontabellen gevuld;
- alleen de hub-tabel product wordt door tabellen uit meerdere systemen gevuld.

Een aantal opvallende aspecten van het model:

- Er is geen verplichte laadvolgorde. Als een nieuw product wordt opgevoerd in beide systemen, maakt het niet uit welk systeem het eerst het datawarehouse vult;
- Indien de producttabellen handmatig onderhouden worden in beide systemen, is dat voor het model geen probleem. Je kunt snel identificeren welke producten in beide systemen voorkomen, welke alleen in het voorraadsysteem en welke alleen in het ordersysteem;
- Uit het bronsysteem voorraad is de tabel magazijn nog niet ontsloten. In het Data Vault-model is al wel rekening gehouden met de hub. Op het moment dat de tabel magazijn wordt ontsloten, wordt het datawarehouse-model uitgebreid, maar hoeft er niets aan de bestaande programmatuur te veranderen.

Vanuit dit model is het eenvoudig om de benodigde datamart te definiëren. Voor de werkelijke invulling kan gekozen worden voor een fysieke datamart of een view op het Data Vault-model. Tevens is een basis gecreëerd om gelijk- of andersoortige datamarts te bedienen.

Incrementele aanpak

De hubs en links samen vormen het geraamte van het model. Zij vormen de meest 'kale' basis van het geheel. Aan de hand van de hubs en de links, kun je door het hele model manoeuvreren. De satellieten geven verdere opvulling rondom het geraamte en contextuele invulling aan het geheel, enerzijds als basis voor de meetwaarden in feitentabellen, anderzijds als attribuutkenmerken van dimensies.
Met deze aanpak is een enterprise datawarehouse te creëren, als basis voor een divers aanbod van informatie binnen een organisatie:

- Er kunnen eenvoudig datamarts ontworpen en gebouwd

worden. Individuele keuzes kunnen per datamart ingevuld worden. Voor de hand liggende keuzen zijn dimensionele datamarts. Hierbij kunnen de verschillende datamarts gebruik maken van gemeenschappelijke product- en klantdimensies. Maar het gebruik is niet gelimiteerd tot dimensionele datamarts. Ook aan andere toepassingen, zoals datamining-applicaties, kunnen gewenste gegevens op verzoek geleverd worden;

- Analisten met SQL-kennis kunnen data-analyses doen, zonder dat kennis en toegang tot een specifiek bronsysteem noodzakelijk zijn;
- Het kan dienen als Operational Data Store. De satellieten met het hoogste datumveld vertegenwoordigen de actuele situatie;
- Het is een goede basis om voor langere periode de historie van brongegevens bij te houden ten behoeve van toezichthouders of wettelijke verplichtingen.

Elke 'hub' vertegenwoordigt een business-entiteit

De Data Vault-benadering leent zich goed voor een incrementele aanpak. De hubs zijn de *linking pin's* in het gehele model. Op het moment dat bij een transactie een business-entiteit wordt onderkend, die nog niet gemodelleerd is, kan toch alvast de Hub in het model opgenomen worden. Bij toevoeging van de business-entiteit met de contextgegevens, hoef je aan de eerder gerealiseerde programmatuur niets te veranderen.
Het maken van ETL is één van de belangrijke tijdrovende processen van het datawarehouse. Bij de Data Vault-aanpak heb je slechts drie verschillende componenten. In de hubs en de links worden alleen nieuwe records toegevoegd. In alle satellieten wordt verder op een standaard manier historie bijgehouden. Dit biedt mogelijkheden om alle ETL-processen op een vergelijkbare manier te maken. Het is zelfs zo ver door te voeren, dat de meerderheid van het werk volgens standaard werkprocedures in te vullen is. Tot slot is de onderlinge afhankelijkheid van verschillende processen tot een minimum beperkt. Daarmee kunnen ook veel processen parallel draaien. De enige afhankelijkheid zit tussen processen die een gemeenschappelijke hub vullen.
De huidige serie artikelen over Data Vault, gepubliceerd op The Data Administration Newsletter, is helaas nog niet compleet. Voorbeelden van gewenste aanvullingen zijn geavanceerde join-strategieën voor het uitlezen van een Data Vault-model en een overzicht van best practices.

CIF en Dimensioneel Modelleren

De Corporate Information Factory (CIF) is geïntroduceerd door Bill Inmon. Het is een hub & spoke-aanpak, bestaande uit meerdere componenten. Centraal staat het datawarehouse. Dat is de plek, waar integratie plaatsvindt en historie wordt bewaard. Het heeft niet als doel om direct antwoord te kunnen geven op business-vragen. Hiervoor zijn datamarts gepositioneerd. De manier waarop invulling moet worden gegeven aan het datawarehouse wordt grotendeels in het midden gelaten. Data Vault is een modelleertechniek, die kan worden gebruikt om invulling te geven aan de datawarehouse-component van het CIF. Het voldoet aan alle eisen, die het CIF hiervoor stelt.

Het dimensioneel datawarehouse volgens de benadering van Ralph Kimball is een gecombineerde methodiek/techniek, waarin het datawarehouse op een heel andere manier wordt gepositioneerd. Het datawarehouse is opgebouwd uit een veelvoud van dimensionele modellen. De feitentabellen staan daarbij centraal. De dimensies zijn veelal geconformeerd en worden door verschillende modellen gedeeld.
Indien deze methode 100 procent zuiver wordt gehanteerd, is Data Vault geen alternatief. In de praktijk wordt vaak afgeweken van de zuivere theorie:

Complexiteit bij meerdere bronnen.
Met name datawarehouse-systemen, die vanuit veel bronnen worden gevuld, hebben vaak grote uitdagingen in de integratie- en ETL-sfeer. Dit heeft weer consequenties voor de onderhoudbaarheid. Wijzigingen in de modellering hebben vaak grote consequenties voor de ETL-processen. Aangezien direct de eindgebruikerapplicaties worden geraakt zijn dit ingrijpende en tijdrovende trajecten.

Mate van detail.
Daarnaast is de mate van detail (granulariteit) in een dimensioneel datawarehouse afhankelijk van de vraagstelling. Dit betekent dat niet per se het laagste niveau van detail van de bron wordt gehanteerd. Dit heeft twee grote nadelen:

- Bij wijzigingen van requirements, met meer detail zal veel verandering plaatsvinden;
- Bij discussie over correctheid zijn geen detailgegevens meer beschikbaar, waardoor validatie zeer complex wordt, omdat je weer terug moet naar de bron.

	voorraadsysteem		**ordersysteem**	
	voorraad	product	product	product-groep
magazijn	X			
voorraadniveau-informatie	X			
voorraadniveau	X			
productinformatie voorraad-systeem		X		
product	X	X	X	
productinformatie			X	
link product/productgroep			X	
productgroep			X	X
productgroepinformatie				X

Afbeelding 3: Tabel.

Tegenwoordig is het bij dimensioneel modelleren een algemeen geaccepteerd uitgangspunt om altijd het laagste niveau van detail ergens in het model op te slaan.

Sleutels.
Volgens de theorie wordt elke keer een nieuwe sleutel uitgegeven bij iedere nieuwe versie van een dimensie. Dit stuit in de praktijk op een aantal nadelen:
- Het is alleen mogelijk om analyses te maken, tegen de dan geldende dimensiewaarden;
- Het is lastig om historie van dimensies te gebruiken;
- Je kunt dit alleen goed doen, als alle brontabellen, die deze tabel voeden beschikbaar zijn. Dit kan timing-problemen veroorzaken.

Als alternatief voor een aparte identificatie per versie wordt deze problematiek in de praktijk afgehandeld met een combinatie van identificatie en datum als unieke sleutel per versie. Per unieke business-entiteit wordt dan een sleutel gegenereerd. Dit lijkt sterk op de satelliet uit Data Vault.

Om aan bovenstaande uitdagingen van dimensioneel modelleren het hoofd te kunnen bieden, wordt in de praktijk vaak een laag gerealiseerd, waarin voorbereidingen worden getroffen voor de integratie. Afhankelijk van de mate van integratie kun je dat in het ene geval een staging-area en in het andere geval een datawarehouse noemen. Deze tussenlaag bevat de laagste mate van detail, zodat ook een andere vraagstelling beantwoord kan worden.
Indien een organisatie kiest voor een tussenlaag is Data Vault een mogelijke keuze. Het is niet moeilijk om van een dimensioneel model een Data Vault-model te maken. Omgekeerd is het ook zeer eenvoudig om vanuit het Data Vault-model het dimensionele model te vullen. Het dimensionele model kan dan als exploratie-laag gebruikt blijven, en het Data Vault-model, als tussenlaag of integratielaag.

Sterke en zwakke punten

De belangrijkste voordelen van deze manier van data-modelleren:
- Het biedt de mogelijkheid om alle historie op een eenduidige manier in het model vast te leggen;
- Het is een uiterst flexibele vorm van modelleren. Je kunt heel gemakkelijk nieuwe stukken toevoegen;
- Vanuit het Data Vault-model is het niet meer moeilijk om een dimensioneel model te vullen. Het is ook mogelijk om datamarts te ontwerpen en te implementeren als views op het Data Vault model. In dat geval moet wel gekeken worden of aan de performance-eisen voldaan kan worden;
- Het Data Vault-model is tevens geschikt om andersoortige data-marts, zoals datamining-applicaties van gegevens te voorzien;
- Er zijn slechts drie componenten. Dit betekent ook dat er slechts een beperkt aantal verschillende ETL-processen zijn. Dit maakt het mogelijk om de ETL-processen te standaardiseren. Een stap verder is het genereren van ETL. In de praktijk zijn er al implementaties met gegenereerde ETL;
- Er is een beperkte afhankelijkheid tussen processen bij het laden van het datawarehouse. Alleen de hubs worden door meerdere processen geladen, zodat daar een afhankelijkheid tussen bestaat. Verder mag alles parallel gedraaid worden;
- Het is geschikt om real-time bij te laden.

Er is echter ook een aantal zwakkere punten te noemen:
- Het sluit niet direct aan op de Business Intelligence-producten. Deze producten verwachten een dimensioneel model;
- Het aantal tabellen kan gigantisch groeien. Er zullen meer tabellen zijn dan in een dimensioneel datawarehouse. Het voorbeeld uit afbeelding 2 heeft al 9 tabellen, terwijl een vergelijkbaar dimensioneel model 2 of 3 tabellen zou hebben. Performance tuning is daarom een aandachtspunt. Afhankelijk van de keuze van de database is dit een discussiepunt of niet;
- Het heeft pas een korte historie en daarmee een beperkt track record.

> De mate van detail in een dimensioneel datawarehouse is afhankelijk van de vraagstelling

Kijkend naar de twee traditionele benaderingen is er bij beide benaderingen bestaansrecht voor Data Vault. Bij de benadering van Inmon is Data Vault een goede optie als modelleertechniek voor het enterprise datawarehouse. En bij een dimensioneel datawarehouse zou het ingezet kunnen worden als hulpmiddel om de integratieproblematiek op te lossen.

Conclusies

Iedere datawarehouse-architect zou zich ten minste op de hoogte moeten stellen van de Data Vault-benadering. Net zoals de aanpak van Bill Inmon en het dimensioneel modelleren van Ralph Kimball bevat Data Vault concepten die een bredere kijk op het vakgebied van datawarehousing geven. Voor de invulling van datawarehouses, waar is gekozen voor een hub & spoke-architectuur, is het zeker als één van de alternatieven te overwegen. Op dit moment is het een nadeel dat er geen compleet literatuuroverzicht is en dat er een beperkt track record is. Hierdoor zal de architect zelf nog een groot aantal keuzes moeten maken om een slimme en bruikbare invulling te geven. Voor de ervaren architect zal dit niet een echt probleem zijn en Data Vault biedt dan ook goede mogelijkheden. Voor de minder ervaren architect zal dit echter de nodige risico's met zich mee kunnen brengen.

Maarten Ketelaars
Maarten Ketelaars (mketelaars@cibit.nl) is senior adviseur bij CIBIT Adviseurs|Opleiders.