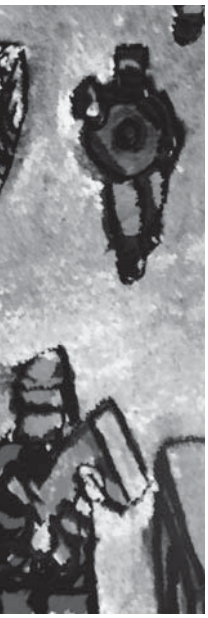


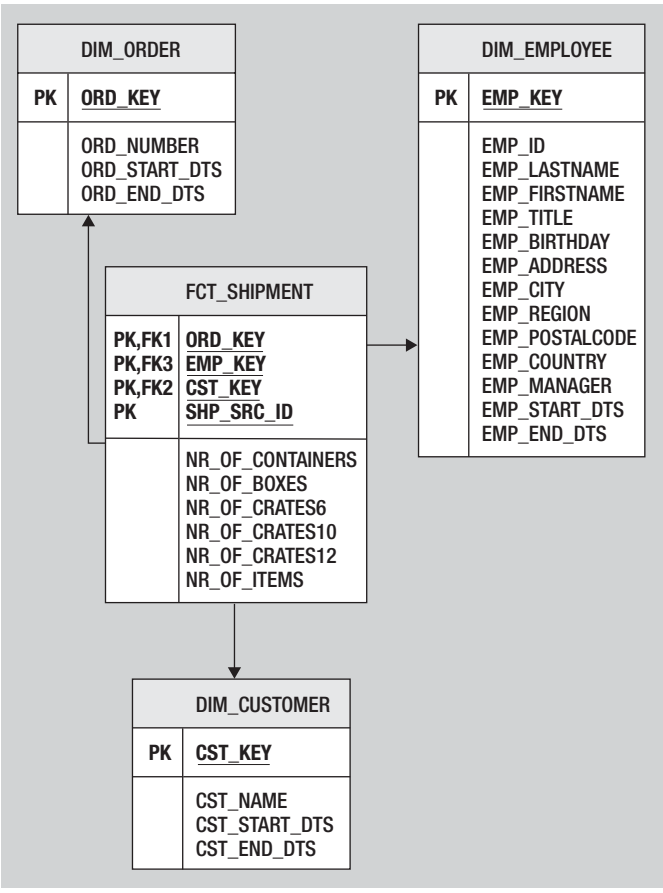
Consequente toepassing van één techniek bepaalt succes datawarehouse

Het datawarehouse modelleringsdilemma

Jan-Paul Fillié

In datawarehouse land bestaan twee kampen, die elkaar vooral op het punt van datamodellering bestrijden. Aan de ene kant heb je Bill Inmon, de grondlegger van de datawarehouse-theorie gebaseerd op het relationele model in de derde normaalvorm. Ralph Kimball daarentegen heeft teruggegrepen op het dimensionele model en deze verder ontwikkeld.

Op veel punten spreken deze wereldwijd bekende goeroe's elkaar tegen. Kimball heeft sinds de uitgave van zijn dimensioneel modelleerboek 'The Data Warehouse Toolkit' het meeste terrein gewonnen. De kloof tussen de beide kampen lijkt de laatste jaren steeds groter te worden. Voor veel ontwerpers van datawarehouses staat daarom de keuze voor de modellering bij voorbaat vast, namelijk óf relationeel óf dimensioneel. Dit maakt het wel overzichtelijk, maar het is maar de vraag of dit altijd tot de beste keuze voor de betreffende organisatie leidt. In 2003 heeft een nieuwe speler dit strijdperk betreden, namelijk Dan Linstedt met de modelleringstechniek Data Vault. Deze techniek is zowel een tussenvorm tussen dimensioneel en relationeel modelleren als een totaal nieuwe benadering.



Afbeelding 1: Voorbeeld van een dimensioneel schema.

Het beste datawarehouse voor de organisatie

Het grootste deel van de toegevoegde waarde van het datawarehouse is de vorm waarin de informatie is opgeslagen. Immers, de data zijn zo gemodelleerd dat de juiste informatie snel beschikbaar is. Dit in tegenstelling tot een operationeel systeem waarbij de doelstelling veel meer ligt bij het invoeren en opslaan van data. Althans, dat zou zo moeten zijn. In de praktijk komt het echter regelmatig voor dat het grootste deel van het datawarehouse volledig wordt afgeschermd van de gebruikers. Ten behoeve van een klein aantal specifieke gebruikers wordt een paar datamarts beschikbaar gesteld met geaggregeerde data. De doelgerichte modellering van het datawarehouse wordt dan verschoven naar deze datamarts. Dit is geen wenselijke situatie, omdat dit leidt tot veel inflexibiliteit en onvoldoende mogelijkheden om te kunnen voldoen aan nieuwe vragen van gebruikers. Dergelijke datawarehouses worden gekenmerkt door trajecten van meerdere jaren, alleen al om bijvoorbeeld een nieuwe bron aan te sluiten of rapportage te leveren voor een nieuwe afdeling.

De keuze voor een minder geschikte modelleringstechniek kan zorgen voor overbodige complexiteit en verlies van overzichtelijkheid. Dit komt vaak voor in relationele datawarehouses. Het omgekeerde komt ook voor wanneer complexe businessrelaties niet goed kunnen worden gemodelleerd in een te eenvoudig

model. Dit kan het gevolg zijn van een dimensioneel datawarehouse.
Er zijn veel datawarehouses waar wel gekozen is voor een datamodellering gericht op managementinformatie, maar waar dit niet consequent is doorgevoerd of zelfs voor een hybride modellering is gekozen. Binnen dergelijke datawarehouses moet iedere keer een keuze worden gemaakt als er een nieuw ontwerp moet worden gemaakt voor een uitbreiding van het datamodel. Ook is de kans op eindeloze discussies over de juiste keuze erg groot, zeker als er problemen optreden. Dit probleem maakt ook een geleidelijke overgang van het ene datamodel naar het andere zeer problematisch.

> De keuze voor een minder geschikte modelleringstechniek kan zorgen voor overbodige complexiteit

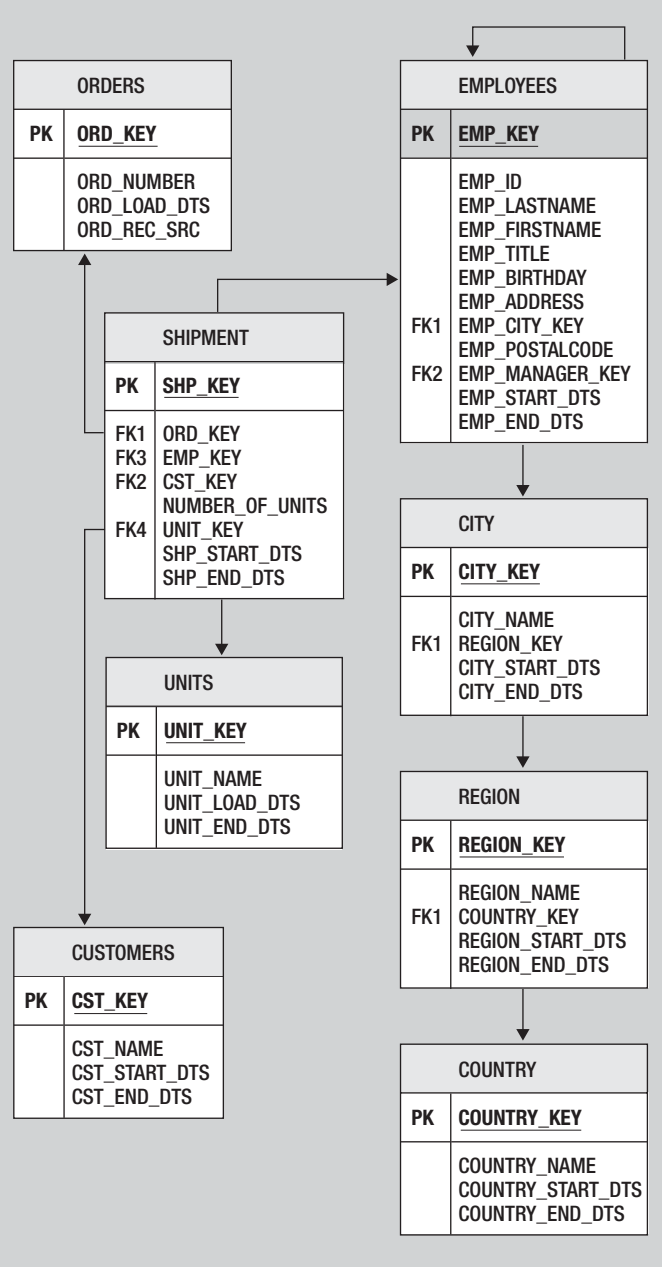
Een groot probleem voor datawarehouse-ontwerpers, die in een bestaande architectuur tot een ontwerp moeten komen, is dat de onbekendheid met een andere modelleringstechniek kan zorgen voor een suboptimale toepassing. Dit is vaak het geval bij dimensionele datawarehouses, waarbij een gebrekkige kennis van deze modelleringstechniek kan zorgen voor het verloren gaan van alle voordelen van de gekozen aanpak.

Selectiecriteria

De keuze voor een bepaalde modellering hangt af van een aantal factoren, waaronder: toepassing datawarehouse; soort gebruikers; organisatiegrootte; complexiteit processen; veranderingssnelheid omgeving; beschikbaarheid van 'out-of-the-box' datamodellen; verwachte datavolumes; de te ondersteunen bedrijfsprocessen.
De combinatie van factoren leidt tot één van de drie modelleringstechnieken. In volgorde van afweging zijn dat: het dimensionele model; het relationele model; Data Vault. Deze volgorde wordt bepaald door de het aantal (geslaagde) toepassingen van de modelleringstechniek in datawarehouses.

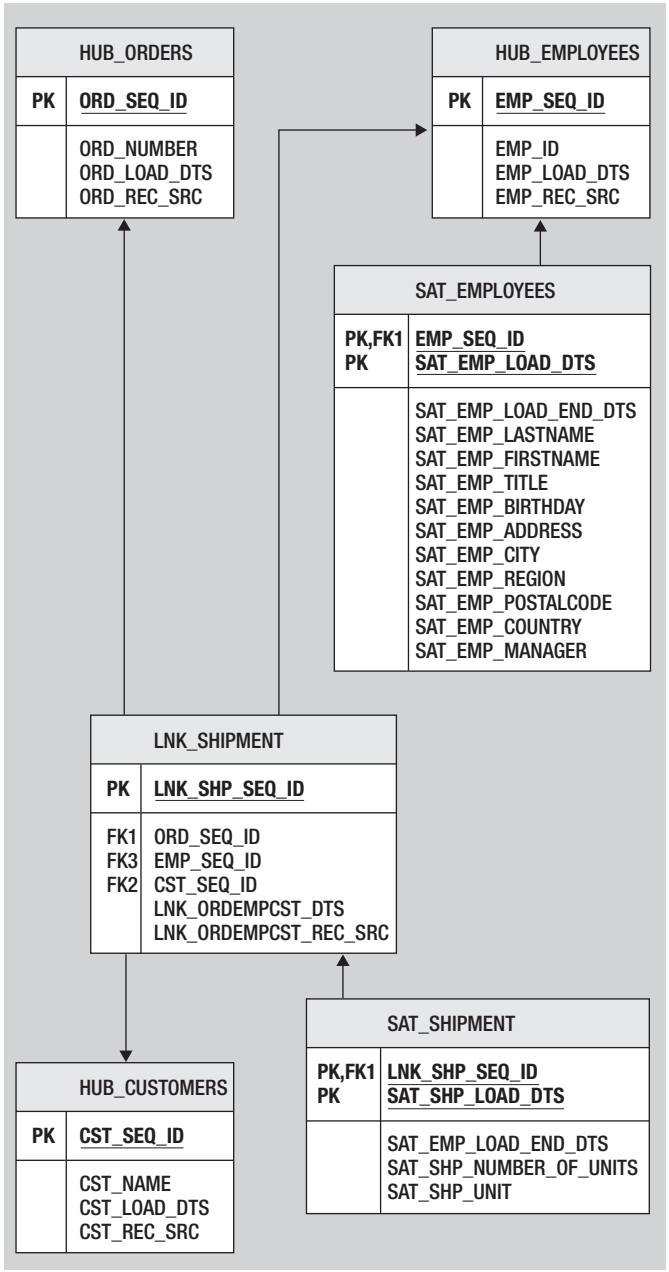
Het dimensionele model

Dimensionele modellering is speciaal door Ralph Kimball ontwikkeld voor toepassing in het datawarehouse. Zelfs Inmon stelt tegenwoordig dat bij het opzetten van datamarts de voorkeur bestaat voor het dimensionele model. Het dimensionele model wordt buiten het datawarehouse ook toegepast binnen OLAP-kubussen, waarmee gebruiksvriendelijke analyses kunnen worden uitgevoerd.
Het model sluit aan bij de beleving van de gebruiker door de opzet van verschillende kijkrichtingen (dimensies) waarmee gebeurtenissen (feiten) beschreven kunnen worden. Deze feiten



Afbeelding 2: Voorbeeld van een relationeel schema.

kunnen verder worden gekwantificeerd door één of meerdere meetwaarden. Binnen de dimensies kunnen weer hiërarchieën bestaan, zodat ook verschillende detailniveau's hierop kunnen worden aangegeven. Belangrijk kenmerk van het dimensionele model is de transparantie voor de eindgebruikers, vrijwel iedereen kan een dimensioneel datamodel lezen. Daarnaast is het combineren en aggregeren van detailinformatie gemakkelijk door de eenvoud van dit model. Door het gebruik van 'slowly changing dimensions' (langzaam veranderende dimensies) kan historie worden bewaard daar waar dit nodig is en op de wijze die aansluit bij de eisen die de business daaraan stelt. Daar waar mogelijk moet hergebruik gemaakt worden van dimensies, voor verschillende business-processen, door deze te conformeren.
Dat wil zeggen dat bijvoorbeeld de dimensie 'klant' hetzelfde



Afbeelding 3: Voorbeeld van een Data Vault schema.

betekent over alle afdelingen van de organisatie heen. Het dimensionele model is het beste toepasbaar in een kleine tot middelgrote organisatie, waar de complexiteit niet al te groot is. Door de relatief grote investering in deze modelleringswijze sluit het goed aan op niet-snelwijzigende bedrijfsprocessen.

Toch gaat het regelmatig fout in de toepassing van deze modellering. Vaak komt dat door een onvolledige toepassing. Een veel voorkomend voorbeeld van een onvolledige toepassing is een datawarehouse, waarin geen keuze is gemaakt waar historie van te bewaren. Dus alles wordt opgeslagen met volledige historie. Dit zorgt voor dimensies (en soms ook feiten) die weliswaar zijn opgezet als langzaam veranderend, maar die in de praktijk toch echt snelgroeiend zijn. Het valt te voorspellen dat dit grote gevolgen heeft voor de benodigde opslagcapaciteit en voor de algehele performance van het datawarehouse.
Een ander fenomeen is voortijdige aggregatie van informatie, een keuze die vaak wordt gemaakt omdat meer detail toch niet van belang is voor de gebruiker. Probleem is dat dit zorgt voor een niet flexibel datawarehouse. Als namelijk wel een vraag komt waarvoor dit detail nodig is, dan moet het datawarehouse opnieuw worden opgebouwd. Vanaf het kleinste detailniveau kunnen wel altijd alle vragen worden beantwoord. Eén van de indicaties dat deze valkuil over het hoofd is gezien, is het bestaan van meerdere datumdimensies binnen één feit.

Kent u die analyses waarbij twee dimensies elkaar uitsluiten? Beginnende datawarehouse-ontwerpers willen nog wel eens iedere kolom in de grootste tabel van het bronsysteem gebruiken als dimensie, waardoor er veel te veel dimensies ontstaan. Veel van die dimensies zijn dan ook lineair met elkaar verbonden. De praktijk wijst uit dat volstaan kan worden met 5 tot 15 dimensies en dat als je goed zoekt veel verbanden kunnen worden aangetroffen tussen dimensies.
Bij veel organisaties heeft iedere afdeling zijn eigen datawarehouse, zodat er geen vergelijking mogelijk is tussen verschillende onderdelen van de organisatie. Zolang geen van deze afdelingen aan geconformeerde dimensies werkt, zal de situatie nooit wijzigen.

Het relationele model

Ruim voordat Kimball met het dimensionele model kwam, werd al relationeel modelleren gebruikt voor datawarehousing. De grondlegger van datawarehousing, Bill Inmon, heeft zijn voorkeur uitgesproken voor het relationele model, ook wel de derde normaalvorm genoemd. Dit is met name bedoeld voor de basis van het datawarehouse, de data op het laagste detailniveau.
Het relationele model is het meest divers van alle modelleringstechnieken. Alle mogelijke relaties tussen entiteiten kunnen worden beschreven. Het is niet voor niks dat het relationele model zoveel terrein heeft gewonnen in database-omgevingen.

> Van de replica-database met rapportagemogelijkheden is men teruggekomen

Daarnaast is het model zeer efficiënt in opslag waardoor ook grote volumes aan data goed kunnen worden verwerkt. Ook is het mogelijk om volledige historie te bewaren, waardoor altijd teruggegrepen kan worden op situaties in het verleden. Alle beschikbare standaardmodellen voor veelvoorkomende processen zijn relationeel in opzet. Iedereen die databasetechnieken als vak heeft gehad is bekend met dit model, waardoor er veel expertise te vinden is in de markt. Het relationele model is vooral geschikt

voor complexe processen en grote volumes en wordt daarom vaak toegepast in grote ondernemingen. Door de vaak hoge complexiteit is het minder geschikt voor snel wijzigende omgevingen. De grootste valkuil van het relationele model is door te dicht bij het bronmodel te blijven. Van de replica-database met rapportagemogelijkheden is men nu wel teruggekomen. Zonder hermodellering voor toegankelijkheid heeft het datawarehouse weinig toegevoegde waarde. Het is juist van belang om een geïntegreerde kijk op de business te bieden.

Data Vault

Als nieuwkomer in deze arena heeft Dan Linstedt nog veel te bewijzen. Zijn modelleringstechniek Data Vault wordt gepositioneerd tussen de twee voorgaande technieken in. Data Vault is net als dimensioneel modelleren speciaal ontwikkeld voor toepassing in het datawarehouse. Data Vault bestaat in essentie uit entiteiten die als *hub* worden aangeduid. De attributen van deze entiteit die kunnen veranderen worden aan de hub gekoppeld als *satellite*. 'Links' beschrijven vervolgens de relaties tussen verschillende hubs. Ook deze links kunnen weer satellites hebben voor bijvoorbeeld meetwaarden of andere attributen.
Door deze opzet is het mogelijk om het model naar alle kanten uit te breiden. De grote flexibliteit maakt een iteratieve ontwikkeling van het datawarehouse mogelijk, waardoor het zeer nuttig is in een snel wijzigende, innovatieve, omgeving. Data Vault is met name bruikbaar in een procesgerichte organisatie en is relatief onafhankelijk van de organisatiegrootte, doordat het tussen dimensioneel en relationeel in staat.
Het grootste probleem bij de toepassing van Data Vault is de onbekendheid van de techniek. Alleen ervaren ontwerpers die zich hebben verdiept in de materie zullen in staat zijn deze modellering met succes toe te passen.

Kies de best passende datamodellering

Staat u voor de taak om een nieuw datawarehouse tot een succes te maken, kies dan de meest geschikte modellering op basis van de criteria die van toepassing zijn op uw organisatie. Vanaf dit punt kan de benodigde competentie verder worden ontwikkeld en kan het datawarehouse groeien. Met een verkeerde keuze moet vaak jaren lang doorgewerkt worden en moeten dezelfde obstakels in het ontwerp bij iedere uitbreiding weer worden overwonnen. Een succesvol datawarehouse wordt gekenmerkt door een consequente toepassing van één modelleringstechniek. Dit zorgt ervoor dat de keuzes voor het ontwerp beperkt worden en dat er veel hergebruik kan plaatsvinden van eerdere modelleerontwerpen.

Jan-Paul Fillié
Jan-Paul Fillié (janpaul.fillie@capgemini.com) is als consultant werkzaam bij de Business Intelligence practice van Capgemini.

Update

Consul begeeft zich op gebied van database auditing

Consul risk management, aanbieder van bedrijfsbrede IT-beveiligingssoftware komt met een meer uitgebreide versie van de InSight Suite. Deze suite heeft mogelijkheden voor database auditing voor Oracle, MS-SQL, DB2, UDB en Sybase platforms. Consul InSight biedt de verzekering dat privileged users, zoals databasebeheerders, super users en applicatiebeheerders hun data en hun systemen beheren overeenkomstig het beleid op het gebied van beveiliging en regelgeving. Vanuit een centraal dashboard kunnen de verantwoordelijken op het gebied van beveiliging- en compliance de rapporten en alarmeringen genereren die zij nodig hebben om gevoelige systemen en data op misbruik te monitoren, en om snel met succes interne of externe audits te doorstaan.
De Consul InSight Suite kan ook worden gebruikt voor applicaties, operating systems en beveiligingsapparaten.

Cognos en IBM sluiten wereldwijde strategische overeenkomst

Cognos en IBM maken bekend dat ze hun samenwerking verder uitbreiden. De nieuwe Strategic Alliance-overeenkomst houdt de gezamenlijke ontwikkeling, marketing en verkoop in van SOA (Service Oriented Architecture) gebaseerde oplossingen.
Onderdeel van de samenwerking is dat IBM meer advies gaat bieden ter ondersteuning van de activiteiten en oplossingen van Cognos. Cognos zal zijn BI-producten verder verbeteren en optimaliseren voor hardware, software en services van IBM. Cognos gaat ook IBM's WebSphere- en Information Management-technologieën leveren als onderdeel van de aanbevolen referentiearchitectuur voor Cognos 8 Special Edition. Dit is een complete BI-oplossing die functionaliteit biedt voor rapportage, analyse, scorecarding, dasboarding en event management.
Meer informatie op www.cognos.nl en www.ibm.com

Microsoft neemt BI-leverancier ProClarity over

Microsoft investeert al lang in de verbetering en verbreding van haar BI-aanbod, zoals blijkt uit de introductie in het afgelopen najaar van SQL Server 2005, Office Business Scorecard Manager 2005, en de uitgebreide BI-mogelijkhden in de Excel en SharePoint versies in Office 2007.
De overname van ProClarity die analyse- en virtualisatietechnologie en business-logica-gedreven 'guided' analyses levert, sluit daar geheel op aan.