De geschiedenis van de historie revisited

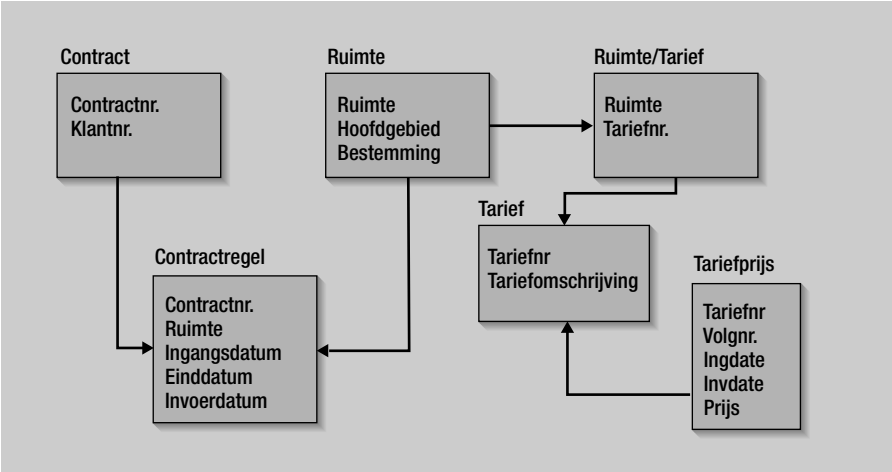
Dweilen in de toplaag

Harm van der Lek

In DB/M 5 stelt Karien Verhagen een aantal vragen naar aanleiding van mijn artikelen over toestandgeoriënteerde sterschema's.
Een mooie aanleiding om er nog een keer op terug te komen, vooral omdat blijkt dat tenminste één lezer dol is op dit soort schrijfsels.

Haar eerste vraag (waar staat de toplaag) is eigenlijk een vraag naar de architectuur van de DWH-omgeving. Aangezien mijn artikel voornamelijk ging over modellering en ETL-processen, zijn dit soort aspecten niet of nauwelijks aan de orde geweest. Het antwoord hangt af van de opzet van de DWH-omgeving. Werkt men met een centraal datawarehouse waarop geen queries worden toegestaan (waar query-gemak geen ontwerpcriterium is geweest) in combinatie met daaruit afgeleide datamarts? Of is het onderscheid niet zo strikt? In ieder geval, de toplaag bevindt zich in het raadpleegbare deel van de DWH-omgeving, want men kan er immers makkelijk detailrapporten (die alle detailgegevens van bijvoorbeeld enkele polissen bevatten) mee maken. Want wat was ook alweer 'de toplaag'? We hadden 'business-entiteiten' waarvan we de historie volledig willen volgen. Bijvoorbeeld: klanten, contracten, dekkingen. We deden dat via een toestandgeoriënteerde feitentabel. Als we even de feitentabel en de dimensietabellen daaromheen als één geheel zien, dan is de toplaag niets anders dan een tabel waarin de versies van deze objecten liggen opgeslagen. In feite geeft het ons huidige inzicht weer in de historie van de objecten (we noemen het 'de toplaag' ter onderscheiding van de 'volledige historietabel'). Van één object representeren de rijen naadloos aansluitende perioden waarin dit object niet aan mutaties onderhevig was.

TWK-mutaties kunnen af en toe tot de nodige hoofdpijn leiden



Figuur 1: Gegevensmodel vastgoedexploitatiesysteem.

Naast de genoemde detailrapporten is deze toplaag nog ergens anders voor nodig. Als er mutaties binnenkomen (komende uit een transactioneel systeem worden zij in de 'backroom' verwerkt), wordt deze nieuwe informatie gebruikt om, in combinatie met de informatie reeds in de toplaag aanwezig, het DWH bij te werken. In normale situaties betekent dat meestal dat de einddatum van het laatste record (31/12/9999) van het object wordt teruggezet naar één dag voor de nieuwe

Vragen en antwoorden

In de rubriek Datawarehouse schreef Harm van der Lek over toestandgeoriënteerde sterschema's (DB/M 6, 7 en 8 van 2001). In DB/M 5 reageerde Karien Verhagen met een aantal vragen. Haar eerste vraag betrof de status van de toplaag, de backroom, het centrale datawarehouse en de datamart. De tweede vraag ging over het dweilen van het centrale datawarehouse. Welke system key heeft een object dat transacties heeft gehad? Harm van der Lek formuleert nu zijn antwoorden.

Zout op een terminologisch slakje

Veel mensen spreken over 'een relationeel model' als ze 'een genormaliseerd model' bedoelen. Maar je hebt niet meerdere 'relationele' modellen waaruit je er één zou kunnen kiezen. Je hebt 'het relationele model' en dat is het door Codd gedefinieerde model voor database management. Het heet zo, omdat Codd het baseerde op het wiskundige begrip 'relatie'. De beste manier om dit abstracte begrip te visualiseren is via een tabel. Dus eigenlijk geldt in dit verband: 'relatie' = 'tabel' en zouden we ook kunnen spreken over 'het tabellenmodel' in plaats van 'het relationele model'. Een datamodel conform het relationele model kan vervolgens al of niet genormaliseerd zijn.

transactiedatum en dat er vervolgens een rij wordt toegevoegd die de nieuwe situatie weerspiegelt. Belangrijk is dat deze nieuwe rij in het algemeen alleen maar kan worden gemaakt door de informatie die in de mutatie is bevat, te combineren met informatie uit de toplaag. Dit omdat de mutatie mogelijk niet een volledig beeld van het object geeft (betreft bijvoorbeeld alleen de adresgegevens). Samenvattend: de toplaag is nodig bij bepaalde backroom processen, maar moet wel aanwezig zijn in het voor gebruikers en/of applicaties zichtbare deel van de DWH-omgeving.

Je moet steeds goed kijken welke conventie gehanteerd is

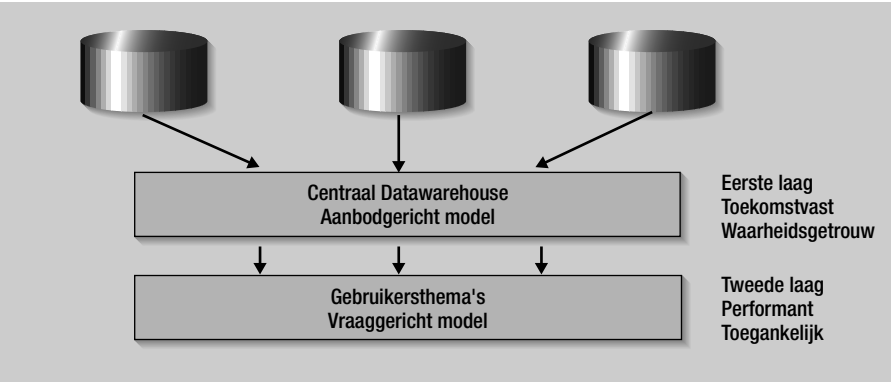
Het is duidelijk dat deze toplaag bij mutaties met terugwerkende kracht soms ernstig op de schop moet, omdat ons huidige inzicht in de historie behoorlijk moet worden bijgesteld. Gelukkig blijkt het voldoende om de rijen die uiteindelijk toegevoegd moeten worden ook (nu met zowel de mutatiedatum als de ingangsdatum) in de volledige historie van de historie weg te schrijven. Ik zal niet ontkennen dat deze TWK-mutaties af en toe tot de nodige hoofdpijn kunnen leiden, maar een probleem zoals Karien dat in haar tweede vraag schetst is er niet. Transacties slaan we altijd op met de gegevens van de bijbehorende objecten, zoals die waren ten tijde van de transactie en wel volgens de kennis die we hadden ten tijde van de transactie en dit verandert gelukkig nooit meer. Bij laat binnenkomende transacties (transactiedatum ligt ver voor de mutatiedatum) kunnen we juist dan dankbaar gebruik maken van de volledige historietabel. Er verdwijnt dan ook niks in het putje. Object 2 (DWH-key = 2) heeft dit nummertje en blijft dit altijd houden. Wat er bij het dweilen uit de toplaag verdwenen is, is een versie van object 2. Deze versie werd geïdentificeerd door het nummer 2 plus de ingangsdatum (had dus geen DWH-key) en naar deze versie werd niet verwezen.

NOTATIETECHNIEK

Ik heb nog enkele opmerkingen bij het artikel van Karien. Allereerst zie ik mijn vooroordeel tegen het gebruik van pijlen in datamodel-diagrammen bevestigd. Het probleem is (in tegenstelling tot de notatietechniek met behulp van kraaienpoten) dat je steeds goed moet kijken welke conventie gehanteerd is, terwijl met kraaienpoten meteen duidelijk is waar de 'veel'-kant van een één-op-veel relatie zit. In figuur 1 lijkt het duidelijk dat vanuit 'Contractregel' er twee verwijzingen naar respectievelijk 'Contract' en 'Ruimte' zijn. Maar van 'Ruimte/Tarief' en 'Tariefprijs' lijkt de verwijzing naar 'Tarief' te moeten lopen. Ik denk dus dat de pijlen hier verkeerd staan. Met deze interpretatie is de tabel 'Tariefprijs' inderdaad vergelijkbaar met de volledige historietabel uit mijn artikel. Karakteristiek namelijk is het voorkomen van zowel een ingangsdatum (INGDATE) als een mutatiedatum (INVDATE). Kennelijk heeft Karien dit zelf ook al eens toegepast. Het SQL fragment in haar 'Archeologie'-kader komt inderdaad overeen met een soortgelijk statement dat ik geef. Je hebt MAX nodig om, onder negeren van latere wetenschap, je kennis over een peildatum (PRIJSDATUM) op een bepaalde peildatum (RAPPORTDATUM) te achterhalen. Wat mij wel bevreemdt is dat haar statement ongelijkheden (#=) bevat en geen groter dan (>).

Meestal wordt slechts beweerd dat een genormaliseerd model 'flexibeler' is

Karien beschrijft een toepassing in de operationele sfeer. Mijn artikelen gingen uitsluitend over DWH-toepassingen en in dat kader was mijn claim dat er een zeer



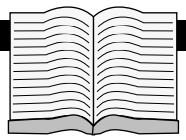
FIGUUR 2: HET TWEE-LAGENMODEL.

breed toepasbaar en elegant formalisme voor de ETL-problematiek bestaat.

Interessant is haar suggestie om de termen aanbod- en vraaggericht toe te passen op respectievelijk het centrale DWH enerzijds en zoiets als de datamarts anderzijds. Haar standpunt is dat het centrale DWH moet worden gebouwd volgens een genormaliseerd model. Hier heeft ze wellicht een punt, maar ze moet de heren Kimball en Inmon dan wel verschillend aanspreken, want juist hier verschillen deze goeroes van mening. Meestal wordt slechts beweerd dat een genormaliseerd model 'flexibeler' is, waarmee wordt bedoeld dat het makkelijker is aan te passen aan veranderde inzichten en nieuwe wensen. Dat is echter maar gedeeltelijk waar. Veranderde inzichten kunnen bijvoorbeeld ook inhouden dat een bepaalde beperkingsregel vervalt. Als dat dan een functionele afhankelijkheid is op basis waarvan we hebben genormaliseerd, dan zouden we die normalisatie weer moeten terugdraaien, terwijl we in een reeds gedenormaliseerd model misschien helemaal niets hoeven te doen. De reden waarom ik denk dat het soms een goed idee kan zijn om in een centraal DWH de historie van de entiteiten van het genormaliseerde tijdloze model expliciet op te slaan heeft te maken met, alweer, de TWK problematiek: juist bij de verwerking van 'late-arriving' mutaties is het prettig als de historie van de omringende zaken expliciet aanwezig is (desnoods in de staging area) in plaats van impliciet verborgen in gedenormaliseerde sterschema's.

Dit laatste klinkt allemaal erg abstract en moet nodig met behulp van voorbeelden worden verduidelijkt. Ik heb me dan ook voorgenomen om dat in een volgend artikel een keer te doen. ●

Dr. H. van der Lek (vdlek@vdlek.nl) is onafhankelijk consultant en docent.



Congressen, beurzen e.d.

13/11: De nationale GSE Conferentie 2002
Motel de Witte Bergen, Eemnes.
Org./inf.: http://gsenl.gse.org

20/11: Dutch DB2 User Group (DDUG)
Nederlandse conferentie voor DB2-gebruikers. Computer Associates (CA), Nieuwegein. Org./inf.: www.kbce.nl

22/11: Jubileumcongres Dutch Ingres Usergroup
Met Michael Stonebraker, Rick van der Lans, Govert Schilling. Planetarium Artis, Amsterdam, 9.30 uur. € 25.
Org./inf.: tjerk.post@erd.com (voorzitter IUGN), m.luyendijk1@chello.nl, marion.zumbrink@ca.com

Alle vermelde bedragen zijn excl. BTW.

10/12: Oracle Financial Services Industry Forum
Forum voor de financiële dienstverlening. Media Plaza, Utrecht.
Org./inf.: www.forum-fd.nl

Cursussen, seminars e.d.

11-12/11: Ontwerpen van de nieuwe generatie datawarehouses
Masterclass met Rick van der Lans. Leiden, Holiday Inn, 9.30-17.00 uur. Kosten: € 1250 (€ 1175 voor DB/M-abonnees). Org.: Array Publications, info: www.array.nl, (036) 5409111.

13-14/11: Fysiek database-ontwerp
Seminar met Rick van der Lans. Gent (B), Holiday Inn Gent Expo, 9.00-17.00 uur. Kosten: € 980. Org./inf.: I.T. Works, www.itworks.be/, (00) 32 9 2415613.

20-21/11: Praktisch modelleren met UML
Seminar met Sander Hoogendoorn. Gent (B), Holiday Inn Gent Expo, 9.00-17.00 uur. Kosten: € 980. Org./inf.: I.T. Works, www.itworks.be/, (00) 32 9 2415613.

27-28/11: Trends in databases en Business Intelligence
Cursus. Apeldoorn, 9.00-17.00 uur. Kosten: € 1.195. Org./inf.: Ir. J. van den Brink, Stichting PATO, tel.: (070) 3644957, e-mail: info@pato.nl

21, 22 en 28/1: Dimensionaal modelleren
Cursus met Harm van der Lek. Amsterdam ZO, Planetarium Gaasperplas. Org./inf.: VanderLek Advies BV, www.vdlek.nl, (035) 6216928.