Nieuwe variant op sterschema

Status-georiënteerde feitentabellen

Gertjan Vlug

Voor het aanbieden van de juiste informatie aan business gebruikers is het inrichten van datawarehouses en datamarts onontbeerlijk. Deze bevatten feitentabellen en dimensies, beter bekend als sterschema's, die nodig zijn om de informatie in begrijpelijke vorm aan te bieden. Dit artikel introduceert een nieuwe variant van zo'n sterschema: status-georiënteerde feitentabellen.

Datamarts zijn extracties van het datawarehouse, ontworpen voor een specifiek doel: een groep mensen, een afdeling, een locatie, enzovoort, met bepaalde informatie-eisen. Datamarts zijn doorgaans ontworpen in een sterschema, zodat de informatie eenvoudig te begrijpen is door de zakelijke gebruikers. Qua structuur en betekenis. Deze sterren bestaan uit een of meer feitentabellen en dimensietabellen daar omheen. Je zou kunnen afstuderen op dit onderwerp en er enkele jaren aan besteden om dit goed te bespreken, maar in dit artikel beschouwen we dit als een bekend *feit*.

De meest bekende smaken van de feitentabellen, zoals beschreven door Ralph Kimball, zijn Transactionele Feiten en Periodieke Snapshots. Deze twee typen hebben betrekking op feitelijke gebeurtenissen, periodieke verslagen, en de meest recente status van een object. Deze soorten zijn algemeen bekend en inmiddels goed begrepen. We zullen ze hier daarom slechts kort beschrijven.

Transactionele Feiten

Een transactie is een zakelijke gebeurtenis en de eigenschappen hiervan zullen gewoonlijk nooit meer veranderen. Het klassieke voorbeeld is de verkoop van een bepaald product op een bepaald moment in een supermarkt. Dit betekent dat de rijen in zo'n feitentabel gewoon worden ingevoegd in het datawarehouse wanneer er nieuwe transacties bij zijn gekomen; deze worden achteraf nooit meer herzien. Deze transactie heeft dus zelf geen geschiedenis. Natuurlijk verwijst dit feiten-record naar de dimensies en sommige hiervan, de langzaam veranderende dimensies, zullen de geschiedenis van de dimensie-objecten bevatten (zoals klanten en producten). Elke rij in een dergelijke dimensie vertegenwoordigt in feite een versie van een dimensie-object. Het idee is dat de transactie-rij naar de versie wijst, zoals die gold ten tijde van de transactie.

Periodieke snapshots

Het is zinvol op te merken dat, voor zover het berekende meetwaarden – zoals de omzet in de supermarkt bijvoorbeeld – betreft, periodieke snapshots alleen aggregaten van de gedetailleerde transactionele feitentabel zijn. Bijvoorbeeld, als men maandelijks een periodiek snapshot wil, worden alle rijen van een bepaalde maand samengevoegd, dus waar de transactiedatum in die maand valt. Dit kan een fysieke afgeleide tabel zijn of aggregaten die niet zichtbaar zijn voor de gebruiker, omdat automatische aggregatienavigatie plaats vindt. Meer significant gebruik van periodieke snapshots van feitentabellen komt naar voren als er semi-additieve feiten bij betrokken zijn, zoals bijvoorbeeld het saldo op een spaarrekening. Men wil dan de waarden van de balans weten zoals deze waren op, bijvoorbeeld, het einde van een periode.

Een dergelijk maandelijks periodiek snapshot geeft dus inzicht in de geschiedenis van de saldi zoals ze waren op het einde van elke maand. Door deze manier van opslaan gaat echter veel detailinformatie verloren. Voor een bepaalde rekening zijn alleen de waarden van het einde van de maand beschikbaar en er is geen gedetailleerde informatie beschikbaar over hoe vaak en wanneer deze waarde is veranderd gedurende de maand. Bovendien kunnen andere attributen, eventueel in dezelfde dimensie, ook veranderen tijdens de maand.

Het is duidelijk dat een periodiek snapshot vele rijen kan bevatten – soms te veel, zonder dat dit extra waarde oplevert. Stel je het geval van een verzekeringsmaatschappij voor. Een huis is verzekerd tegen brandschade en deze situatie verandert niet gedurende vele jaren. Toch wordt er iedere maand een nieuwe rij in de maandelijkse snapshot-feitentabel toegevoegd, niet meer dan het herhalen van dezelfde informatie dus. Het is om deze reden dat de status-georiënteerde feitentabel wordt geïntroduceerd.

Customer version key	Customer Code	From Date	To Date	City
5	CG067	01/01/1900	21/07/2009	New York
8	CG067	21/07/2009	31/12/9999	Boston

Afbeelding 1: Een langzaam veranderende dimensie.

Account Number	From Date	To Date	Customer version key	Balance
3200354	01/01/1900	10/07/2009	5	$ 100
3200354	10/07/2009	21/07/2009	5	$ 200
3200354	21/07/2009	31/12/9999	8	$ 200

Afbeelding 2: Status-georiënteerde feitentabel.

Status-georiënteerde feitentabel

Wanneer is het verstandig om een status-georiënteerde feitentabel te gebruiken? Daarvoor zijn drie criteria aan te geven:

- als een laag detailniveau van objecten relevant is voor het bedrijf;
- als deze veranderen op een onvoorspelbare manier;
- als de wijzigingen moeten worden vastgelegd op een zeer gedetailleerd niveau, met inbegrip van veranderingen in de bovenliggende objecten.

In een status-georiënteerde feitentabel geeft elke rij de status weer van de toestand van een object, gedurende een periode waarin deze toestand niet is veranderd. Merk op dat veranderingen op een hoger niveau dienen te worden beschouwd als veranderingen van het betreffende lagere object.
We geven een eenvoudig voorbeeld. Stel, een verzekeringsmaatschappij wil informatie over haar klanten: de stad waar ze wonen, hun saldo, enzovoort. Het laagste niveau van objecten bestaat uit spaarrekeningen.
We hebben hier al direct een langzaam veranderende dimensie (type 2 van Kimball): de klantdimensie. In afbeelding 1 zien we dat een klant blijkbaar verhuisd is van New York naar Boston op 21 juli 2009. Afbeelding 2 toont een voorbeeld van een status-georiënteerde feitentabel met betrekking tot de klantdimensie en de semi-additieve meetwaarde 'saldo'.

Op 10 juli 2009 is het saldo gewijzigd van 100 dollar in 200 dollar, dus er kwam een nieuwe toestand van deze klant (de tweede rij). De verhuizing naar Boston van deze klant moet ook correct worden weergegeven in de tabel, dus weer voegen we een nieuwe rij toe en herhalen hier het saldo, ook al is die niet veranderd. Dit is correct als we het totale saldo per stad op een bepaald moment in de tijd willen weten, bijvoorbeeld aan het eind van juli 2009. Het is duidelijk dat het saldo van deze CG067 klant bij Boston moet worden geteld.
Let op: de periodieke momentopnamen kunnen nu eenvoudig worden gedefinieerd als view over dit status-georiënteerde sterschema. Om dit te doen voor een maandelijks snapshot moeten we een maand tabel met (onder andere) de datum van de laatste dag van de maand hebben, zie afbeelding 3.

De view, die ons een gedetailleerd maandelijks periodiek snapshot oplevert, is als volgt:

```
CREATE VIEW MONTHLY_ACCOUNT_SNAPSHOT AS
SELECT Month.MonthNr
, SOFactTable.AccountNr
, SOFactTable.Customer_version_key
, SOFactTable.Balance
FROM SOFactTable
, Month
WHERE Month.LastDay >= SOFactTable.From_date
AND Month.LastDay < SOFactTable.To_Date
```

Merk op dat de twee genoemde voorwaarden iets dergelijks uitdrukken:

```
Month.LastDay BETWEEN2 SOFactTable.From_date AND
                                    SOFactTable.To_Date
```

die op hun beurt weer "neem de status, die geldig was op het einde van de maand" betekenen. In ons voorbeeld zal het resultaat zijn zoals getoond in afbeelding 4.
Deze aanpak is nuttig in situaties waar objecten niet heel vaak veranderen, terwijl het gedrag onvoorspelbaar is als ze wel veranderen. Voorbeelden daarvan zijn: contracten, verzekeringen, voorraad en nog veel meer.

MonthNr	LastDay
200906	30/06/2009
200907	31/07/2009

Afbeelding 3: Maandtabel.

MonthNr	Account Number	Customer version key	Balance
200906	3200354	5	$ 100
200907	3200354	8	$ 200

Afbeelding 4: Periodiek snapshot view.

Voorbeeld twee

Een ander voorbeeld waar status-georiënteerde feitentabellen kunnen worden gebruikt, is een zogenaamde *monster-dimensie*. Stel dat dezelfde verzekeringsmaatschappij over een enorm klantenbestand beschikt zodat de klantdimensie van een transactionele feitentabel zeer groot kan zijn, vooral als het gaat om een Kimball type 2 dimensie. Waarschijnlijk zal men mini-dimensies voor een aantal analytische eigenschappen gebruiken, zodat het aantal rijen in de langzaam veranderende klantdimensie niet te groot wordt, vanwege allerlei onafhankelijke wijzigingen in de talrijke klantattributen. De klantdimensie zelf zal waarschijnlijk dimensies bevatten waarvan de attributen zelden veranderen. Misschien is het zelfs een Kimball type 1 dimensie.
Maar veronderstel dat, afgezien van een analyse van de transacties, het analyseren van de complete geschiedenis van de klanten een vereiste is; dan is de gebruikte dimensie hier niet langer geschikt voor. Er kan dan gebruik worden gemaakt van een status-georiënteerde feitentabel voor de complete geschiedenis van elke klant. In een dergelijke tabel bevindt zich een rij voor elke klantversie tijdens een periode waarin er niets verandert. Misschien heeft deze tabel helemaal geen feiten, omdat de enige zinvolle – semi-additieve! – meetwaarde 'het aantal klanten' is. En dat is altijd 1 op dit laagste niveau.

Merk op dat deze klantdimensie en alle afgeleide mini-dimensies van de transactietabel *conformed dimensies* van deze 'klantgeschiedenis feitentabel' kunnen en zullen zijn.
Deze feitentabel zal net zoveel rijen als een complete type 2 dimensie hebben. De laatste (als een dimensietabel) kan onaanvaardbaar groot worden, terwijl de status-georiënteerde feitentabel niet al te groot zal zijn.

Conclusie

Bovenstaand is een nieuw type feitentabel geïntroduceerd, naast de drie reeds bekende smaken, een zogenaamde status-georiënteerde feitentabel. In een dergelijke feitentabel staat elke rij voor de toestand van een object gedurende een bepaalde periode van tijd waarin deze toestand niet verandert. Het is een soort van type 2 benadering van een feitentabel; de methodiek lijkt op wat bekend staat als de *Twin Timestamp Approach*.

Gertjan Vlug is Managing Director van BIReady.

Dit artikel is gebaseerd op de ideeën en teksten van Harm van der Lek, en de functionaliteit zoals deze in de BIReady-software beschikbaar is.

Update

Lezersenquête

Graag wil de redactie weten welke onderwerpen u als lezer bezighouden. Om die reden hebben we een enquête op onze website (www.dbm.nl) gezet, waar u kunt aangeven welke onderwerpen u nu en in de nabije toekomst bezighouden.
Het biedt u de gelegenheid ons te laten weten welke onderwerpen u graag in uw vakblad terug wilt zien; het geeft ons de gelegenheid aan uw wensen te voldoen.
De enquête staat links op de website.
Invullen kost nauwelijks een minuut.
Uw medewerking wordt hoogst gewaardeerd!