Op transparante wijze een Data Vault ontsluiten voor data marts

Open de kluis: Data Vault Unfold

Jan Willem van der Meij

Er is veel geschreven over manieren om een Data Vault te vullen. Maar uiteindelijk gaat het om het klaarzetten van een data mart die is ontworpen vanuit de informatiebehoefte en geoptimaliseerd voor uitvragen. In dit artikel wordt beschreven hoe met behulp van een generieke extractie uit de Data Vault de verschillende data marts incrementeel zijn bij te werken. Ik noem deze methode: Data Vault Unfold.

Met data mart bedoel ik een toestandgeoriënteerde ster met niet-historische conforme dimensies. De grote uitdaging is hier de omzetting van de historische tijdlijnen uit de Data Vault[1] naar de historie in de data mart. Transactiegeoriënteerde data marts laat ik nu buiten beschouwing. Voor meer gedetailleerde informatie over een toestand- of status-georiënteerde feitentabel, zie DB/M 8, december 2010, 'Status-georiënteerde feitentabellen' van Gertjan Vlug. Bij ASR is de methode met succes toegepast. Ik zal de methode beschijven aan de hand van een case waarin de belangrijkste elementen van de methode voorkomen. Een geladen Data Vault is het startpunt.

Randvoorwaarden Data Vault

De essentiële kenmerken van de Data Vault voor het gebruik van deze methode zijn:

1. Elke satelliet heeft alleen unload-datums als waarde voor de DATUM_VANAF. Elke DATUM_VANAF is ooit gevuld met een unload-datum. Dit is de datum die aangeeft wanneer de gegevens aan de bron onttrokken zijn. Naast de DATUM_VANAF heeft iedere satelliet een DATUM_TM. De unload-datums moeten apart worden opgeslagen;
2. Elke hub en link heeft standaard een satelliet met INDICATIE_GELDIGHEID. Met behulp van deze satelliet wordt historisch vastgehouden wanneer een hub- of linksleutel aanwezig is in het bronsysteem. De 1-op-1 relatie in de tijd in links wordt geborgd via deze INDICATIE_GELDIGHEID. In de terminologie van Harm van der Lek zijn dit de asymmetrische links.

Kenmerken data mart

De gewenste data marts zijn toestandgeoriënteerde sterren met niet-historische conforme dimensies. Kern hiervan is dat elke rij in het feit een toestand representeert waarbij alle kenmerken tussen de DATUM_VANAF en DATUM_TM gelijk zijn gebleven. De data mart wordt incrementeel bijgewerkt. Dit betekent dat er geen toestanden kunnen worden tussengevoegd, dus geen mutaties met terugwerkende kracht.

Transformaties

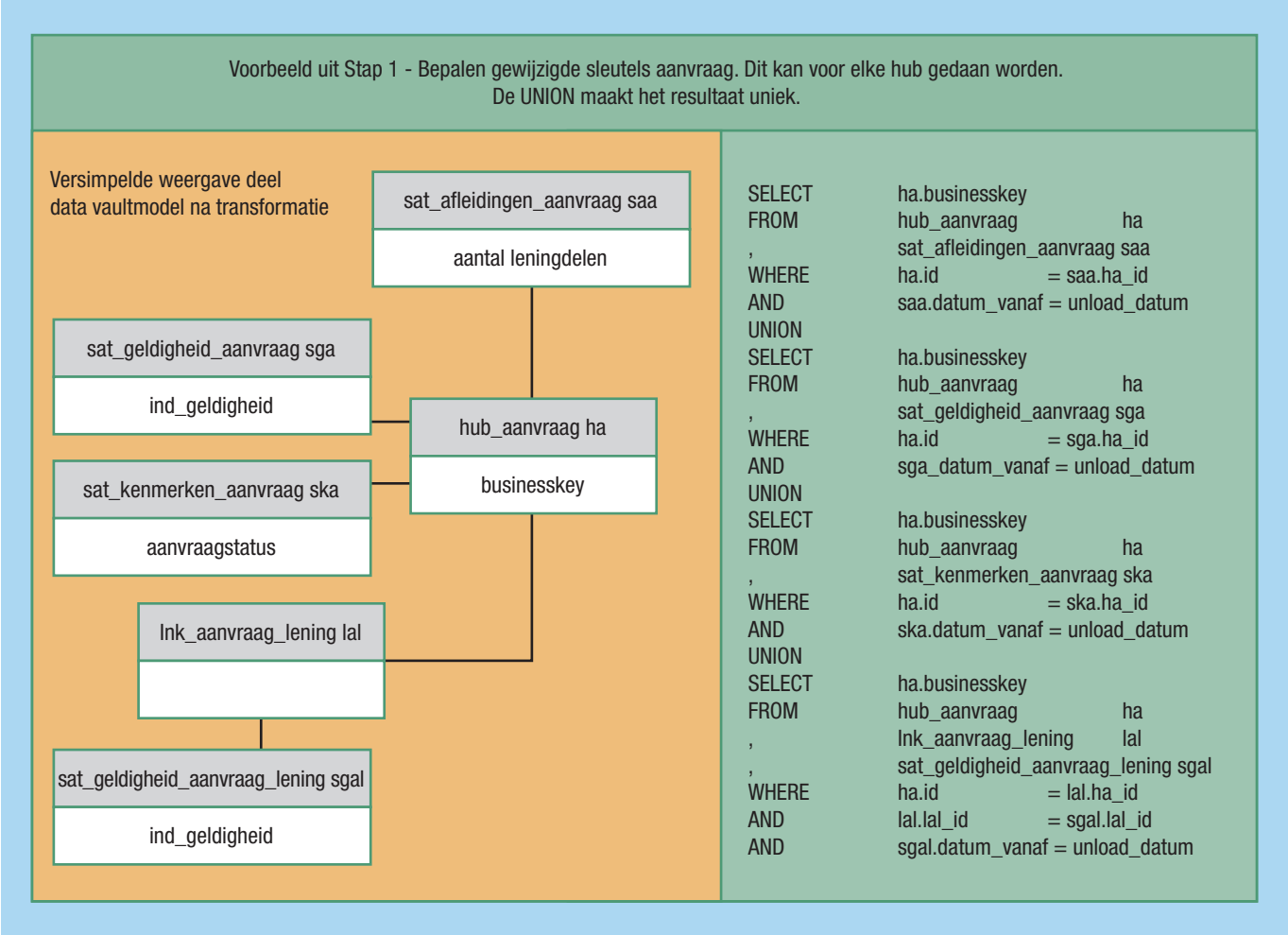
Om bepaalde informatievragen snel en correct te kunnen beantwoorden, is het nodig om de onbewerkte gegevens uit de Data Vault te transformeren, bijvoorbeeld een aggregatie of een formule. Deze getransformeerde gegevens moeten uiteindelijk opgenomen worden in de toestandgeoriënteerde feitentabel. Om de mogelijkheid te behouden om data marts helemaal opnieuw op te bouwen en om de getransformeerde gegevens ook voor andere toekomstige data marts te kunnen hergebruiken, zijn deze gegevens in aparte satellieten in de Data Vault opgeslagen.

> Getransformeerde gegevens moeten uiteindelijk opgenomen worden in de toestand-georiënteerde feitentabel

Transformatielogica is maatwerk, foutgevoelig en blijft zo netjes afgescheiden van de grotendeels generieke manier van bijwerken van data marts.

Case: vullen van de toestandgeoriënteerde ster 'Aanvraag Lening'.

Ik ga uit van de volgende situatie die een klein deel van de Data Vault representeert. Klanten kunnen meerdere aanvragen doen om een lening af te sluiten, bijvoorbeeld voor een hypotheek.



Afbeelding 1: Stap 1, bepalen van de gewijzigde sleutels.

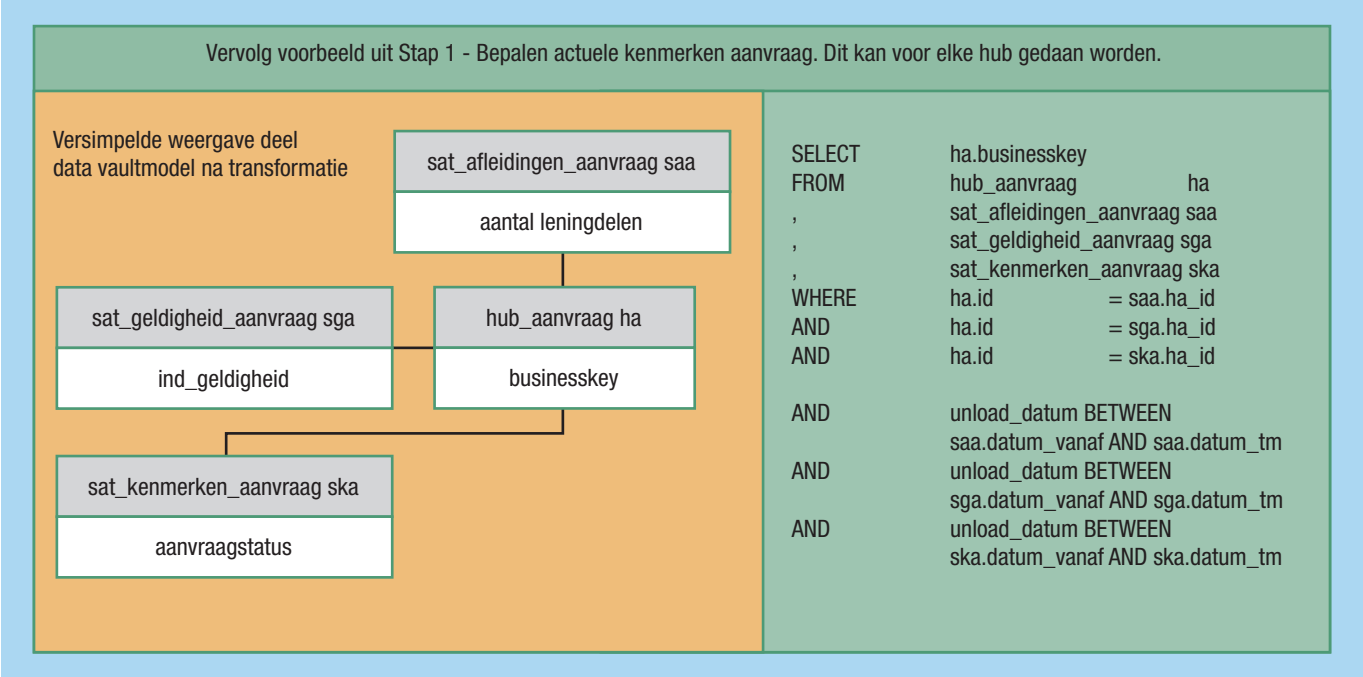
Hypotheken zijn samengestelde producten waardoor een lening bestaat uit meerdere leningdelen. Per leningdeel gelden andere voorwaarden, rentevaste periodes en bijbehorende rentepercentages. Per aanvraag wil de business graag weten uit hoeveel leningdelen de lening is opgebouwd. Dit moet worden afgeleid en wordt apart opgeslagen in een satelliet bij de hub-aanvraag. Het oorspronkelijke leningbedrag en de status van de aanvraag zijn ook essentiële kenmerken.

Iedere hypotheek wordt afgesloten bij een notaris. Een lening kan op één moment in de tijd maar één notaris hebben. Bij overdracht aan een andere notaris, wordt de relatie naar de originele notaris afgesloten met behulp van de IND_GELDIGHEID in de satelliet bij de link tussen lening en notaris. Alleen de naam van het notariskantoor van belang.

Stap 1: Bepalen van de gewijzigde sleutels en koppelen aan actuele kenmerken.
Dit mechanisme staat los van de data marts. Het kan gebruikt worden voor het bijwerken van alle data marts, zowel feit- als dimensietabellen. Hierin ligt de kern van de Data Vault Unfold-methode.

Elke DATUM_VANAF in de Data Vault is ooit gevuld door een unload-datum. Per hub met omliggende linktabellen kan worden bepaald welke hub-sleutels gewijzigd zijn door te kijken of er in de satellieten inclusief de transformaties records voorkomen waarbij de DATUM_VANAF gelijk is aan de laatste unload-datum. Doordat elke hub en link een satelliet heeft waarin de IND_GELDIGHEID wordt bijgehouden, komen ook de nieuwe hub- en linksleutels mee.
Bij iedere gewijzigde sleutel worden uit alle omliggende satellieten alle geldige kenmerken op unload-datum opgehaald. De volgende twee selecties moeten worden gekoppeld op business key. Om redenen van performance worden deze gegevens opgeslagen in een staginglaag.[2]

Stap 2: Bijwerken dimensie Notaris.
De wijzigingen rondom de hub notaris staan in de staginglaag. De dimensie kan van hieruit gemakkelijk worden bijgewerkt. Deze gegevens zijn niet-historisch opgeslagen.
Standaard voeg ik nog twee records toe: surrogate key -1 met business key 'Onbekend' en surrogate key -2 met business key 'Niet van toepassing'. Op deze manier kan elke nieuwe toestand altijd gekoppeld worden aan de dimensie Notaris.



Afbeelding 2: Stap 1, koppelen aan actuele kenmerken.

Stap 3: Bijwerken feit Aanvraag Lening.
Deze stap vergt het meeste denkwerk. Ik zal het helder uiteen zetten.
a. Bepaal per data mart de relevante wijzigingen in de Data Vault.
De data mart volgt de historie van de business key van aanvraag. De geaggregeerde leningdeelgegevens zijn via een transformatie gekoppeld aan de hub aanvraag. De stagingtabel van leningdeel is hierdoor niet meer van belang. Ik heb kenmerken van aanvraag en lening nodig en dus ook beide stagingtabellen.
Van notaris is alleen een wijziging in de relatie tussen lening en notaris van belang. Dit is opgeslagen in de link lening notaris. Gewijzigde business keys van lening en de actuele kenmerken zijn nodig en dat staat in de stagingtabel van lening. De stagingtabel van notaris is hiermee niet meer van belang.

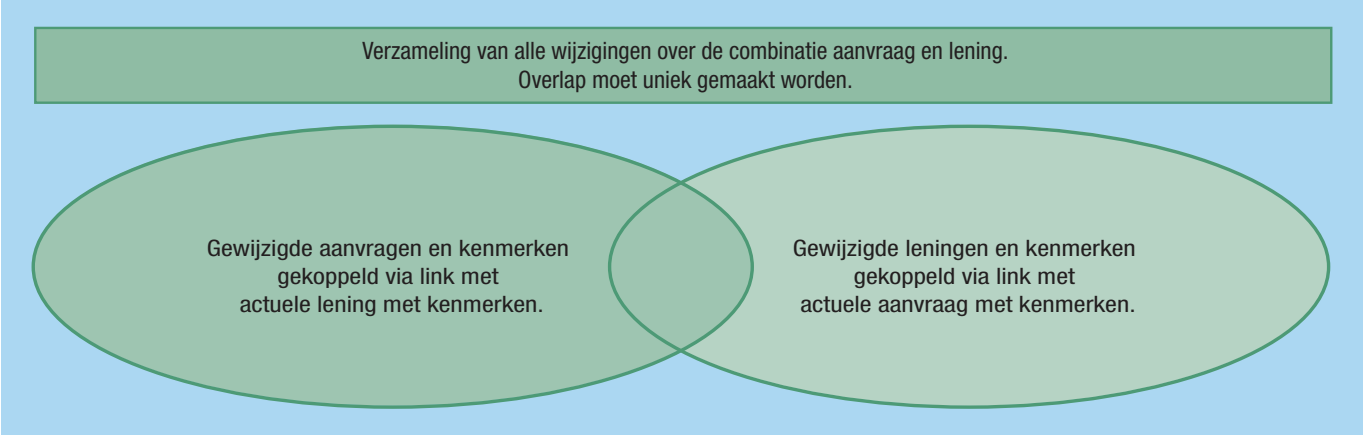
b. Ophalen alle actuele kenmerken bij alle relevante gewijzigde business keys.
Het feit bestaat uit een combinatie van gegevens rondom de hubs aanvraag en lening. De gewijzigde sleutels uit beide hubs zijn bepaald. Van elke gewijzigde sleutel staan de actuele eigen kenmerken in de stagingtabel (zie stap 1).

ASR

ASR is een van de grootste verzekeraars in Nederland. Met de merken ASR Verzekeringen, ASR Pensioenen, Ardanta, De Amersfoortse, Europeesche Verzekeringen en Ditzo biedt ASR een breed assortiment van financiële producten op het terrein van schade-, levens- en inkomensverzekeringen, collectieve en individuele pensioenen, zorgverzekeringen, reis en recreatie- en uitvaartverzekeringen. Daarnaast is ASR Nederland als belegger onder meer actief in vastgoed met ASR Vastgoed Vermogensbeheer en ASR Vastgoed Ontwikkeling. Bij ASR werken momenteel ruim 4.000 medewerkers op verschillende locaties verspreid over het land. Het hoofdkantoor van ASR staat in Utrecht.
Ruim twee jaar geleden is ASR gestart met het gebruik van Data Vault-modellering bij grote nieuwe projecten. Er zijn tegelijkertijd stappen gezet richting een meer geïntegreerd EDW. De officiële toolset bestaat op dit moment uit IBM InfoSphere Datastage (ETL), Cognos en Oracle RDBMS.

> Het eindresultaat moet op de business key van aanvraag uniek gemaakt worden

Omdat ik nu alle actuele kenmerken nodig heb, zal ik de actuele kenmerken van de andere hubs moeten koppelen aan de stagingtabel. In dit voorbeeld dus het koppelen van de stagingtabel van aanvraag aan de actuele kenmerken van lening, maar ook het koppelen van de stagingtabel van lening aan de actuele kenmerken van aanvraag! Voor deze koppeling maak ik gebruik van de actuele waarde van de IND_GELDIGHEID bij de link aanvraag lening. Het eindresultaat moet op de business key van aanvraag uniek gemaakt worden.
Ik weet nu de actuele waarde van bijna alle kenmerken die van



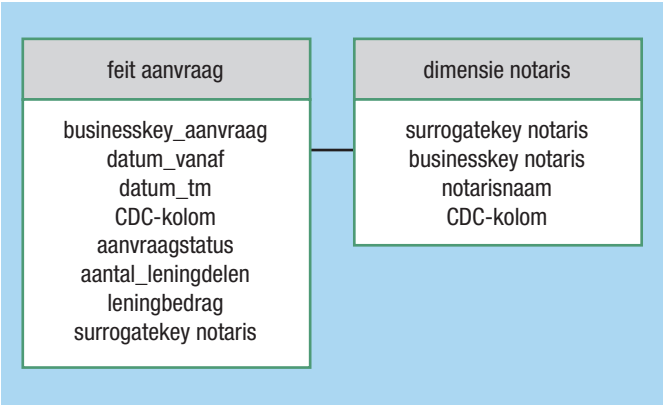
Afbeelding 3: Koppelen stagingtabellen.

belang kunnen zijn bij een data mart op basis van de hubs aanvraag en lening. Wat ik nog nodig heb is de actuele en geldige business key van notaris die uiteindelijk zal resulteren in een verwijzing naar de dimensie notaris.

> Vlak voor de INSERT's moeten de surrogate keys voor de dimensies worden opgehaald

c. Toevoegen actuele en geldige business keys voor koppeling naar dimensies.

Voor het ophalen van business keys voor koppelingen naar dimensies is meer maatwerk nodig. De complexiteit hangt samen met de afstand in het datamodel van de Data Vault tussen in dit geval de hub aanvraag en de hub notaris. Met behulp van de links tussen aanvraag en lening en tussen lening en notaris, wordt de actuele en geldige business key van notaris opgehaald. Als er geen sleutel gevonden wordt, dan gebruik ik de business key 'Onbekend'.



Afbeelding 4: Laden data mart.

d. Laden data mart.

Op basis van een CDC-mechanisme moet worden bepaald welke toestanden moeten worden toegevoegd. Voor de verwijzing naar de hub notaris moet de business key worden gebruikt. Alleen de business keys zijn constant en leven in zowel de Data Vault als de data mart. Surrogate keys kunnen en mogen hiervoor niet gebruikt worden. Vlak voor de INSERT's moeten de surrogate keys voor de dimensies worden opgehaald.

Als laatste moeten alle oude toestanden die nog openstaan worden afgesloten op de dag voorafgaand aan de DATUM_VANAF van de nieuwe toestand, zie afbeelding 4. Ik ben nu klaar. De data mart is bijgewerkt. Morgen weer opnieuw.

Afsluitend

Met bovenstaande methode is het mogelijk om op transparante wijze een Data Vault te ontsluiten voor data marts. Pas op het laatste moment wordt het laadproces beïnvloed door de daadwerkelijke structuur van de specifieke data mart die moet worden bijgewerkt. De flexibiliteit die je creëert door zoveel mogelijk gegevens uit de bron mee te nemen naar de Data Vault wordt zo ver mogelijk in het laadproces doorgevoerd. Eventuele transformaties worden hierin ook meegenomen.

Ik hoop dat de Data Vault Unfold-methode kan helpen bij een lastig, weinig besproken maar zeer essentieel onderdeel: de ontsluiting van een Data Vault.

Noten

1. *Het is mij duidelijk dat Data Vault een manier is om data te modelleren. Voor het gemak zal ik spreken over 'de Data Vault' waarmee ik de omgeving bedoel waarin de data zijn opgeslagen volgens de Data Vault-methode.*
2. *De kenmerken bij de linktabellen kunnen niet standaard worden meegenomen vanwege de veel-op-veel relaties. Impliciet kunnen deze gegevens meekomen in afleidingen bij de hub. Het kan ook zijn dat de ster zich bevindt op het niveau van een link. In dat geval heeft de link een eigen stagingtabel nodig.*

Drs. H.J.W. van der Meij (jan.willem.van.der.meij@asr.nl) is werkzaam als Datawarehouse Architect bij ASR.