Deeloplossingen voor verwerken van mutaties met terugwerkende kracht

Het laden van toestanden

Paul Hulst en Fons Rooijers

De Uitvoeringsinstelling Werknemers Verzekeringen (UWV) verzorgt de uitvoering van sociale verzekeringen voor werknemers en werkgevers, waaronder de Werkloosheidswet (WW) en de Wet op de Arbeidsongeschiktheidsverzekering (WAO). De behoefte aan informatie over de uitvoering van die sociale verzekeringen is zeer divers en loopt van informatie voor de strategiebepaling via rapportages over trendontwikkelingen tot fraudeonderzoek bij individuele gevallen. Om in de gegevens te voorzien is een datawarehouse ontworpen.

Objecten in dit datawarehouse zijn onder andere personen, werkgevers, dienstverbanden en claims. De informatiebehoefte over die objecten wordt gekarakteriseerd door een aantal zaken.
Allereerst veranderen kenmerken van de objecten in de tijd en al die verschillende waarden moeten bekeken kunnen worden, zodat de historie van een object kan worden gevolgd. Een voorbeeld van een informatievraag luidt; wat is de ontwikkeling van de claims in de afgelopen acht kwartalen en hoe ontwikkelde zich bijvoorbeeld de mate van arbeidsongeschiktheid van de claims?
Ten tweede verandert met iedere mutatie de kennis over de historie. Bij het vaststellen van een uitkering moeten de gegevens over de claim en de persoon beschikbaar zijn. Zo wordt bijvoorbeeld de burgerlijke staat meegenomen in de bepaling van de hoogte van een uitkering. Stel dat een persoon als ongehuwd in de administratie staat en deze krijgt daarom een uitkering van 80. Die persoon meldt vervolgens dat hij al sinds een jaar getrouwd is en krijgt daardoor een extra uitkering van 20. Aan de eerste uitkering moet als burgerlijke staat 'ongehuwd' gekoppeld worden en aan de tweede 'getrouwd'. Er is dus behoefte aan een duidelijk dossier, waarin vastgelegd is welke kennis en ontwikkelingen tot een bepaald besluit hebben geleid en op welk moment die bekend waren. Met andere woorden: er is behoefte aan 'geschiedenis van de historie'.[1]

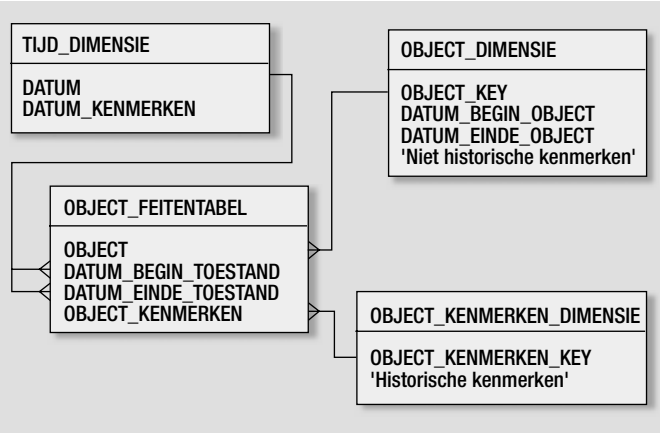
De gegevens van verschillende objecten moeten vaak gecombineerd worden. De koppeling tussen de verschillende objecten wordt in het datawarehouse gelegd en niet in de programma's die gebruikt worden om de gegevens te benaderen[2], zodat bevraging eenvoudig en snel is. Naast de claimkenmerken (zoals bijvoorbeeld de mate van arbeidsongeschiktheid) worden veelal ook de persoons- en werkgeversgegevens gevraagd, bijvoorbeeld het aantal claims per branche van de werkgever en geslacht en leeftijd van de persoon. Deze wensen hebben grote invloed op het ontwerp van het datawarehouse en het benodigde laadproces.

Toestandgeoriënteerd

Kimball stelt voor om de historie van kenmerken bij te houden door te 'stapelen' in de dimensie.[3] Een nadeel van deze werkwijze is dat dimensies die op zich al 'monsterdimensies' zijn, nog groter worden doordat ook nog eens de historie van het object moet worden bijgehouden. Zo kent het datawarehouse ongeveer 10 miljoen personen en 26 miljoen dienstverbanden en dat is nog slechts een deel van de objecten waarover gegevens worden vastgelegd.
In het datawarehouse is daarom gekozen voor een andere systematiek: een toestandgeoriënteerde feitentabel, zoals beschreven door Michael Schmitz.[4] Op ieder willekeurig moment verkeert het onderwerp in een bepaalde toestand, alle kenmerken die gevolgd worden hebben dan een specifieke waarde. Zo'n toestand begint op een bepaalde datum en eindigt ook op een bepaalde datum. Er is dus sprake van een periode (toestand) waarin die kenmerken niet wijzigen, die periode wordt in de toestandgeoriënteerde feitentabel vastgelegd. Iedere periode heeft een begin- en een einddatum en door ervoor te zorgen dat de begindatum van een periode gelijk is aan de einddatum van de vorige plus 1 dag, ontstaat een reeks die de complete historie van het onderwerp beschrijft.
In afbeelding 1 is de historie van een bepaald persoon geschetst die op datum D2+1 in het huwelijk treedt en op datum D3+1 verhuist.

De kenmerken waarvan de verschillende waarden in de tijd gevolgd worden, noemt men historische kenmerken. Naast die historische kenmerken zijn er ook kenmerken waarvan alleen de

Afbeelding 1: Historie van een persoon die op datum D2+1 in het huwelijk treedt en op datum D3+1 verhuist.



Afbeelding 2: Basisopzet van de sterstructuren.

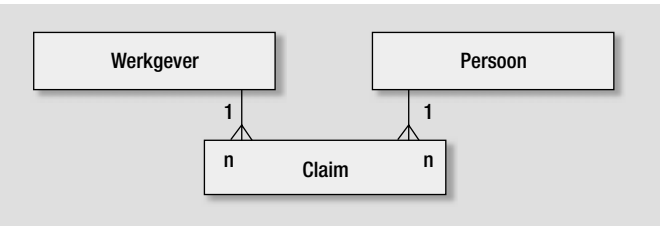
actuele waarde wordt vastgehouden, deze worden de niet-historische kenmerken genoemd. Voorbeelden van niet-historische kenmerken zijn geslacht en geboortedatum.
De historische kenmerken worden vastgelegd in een rij in de feitentabel. De feitentabel verwijst naar een dimensietabel waarin de logische sleutel is opgenomen van het betreffende onderwerp. In die dimensietabel worden ook de niet-historische kenmerken ondergebracht. Dit leidt tot sterstructuren zoals weergegeven in afbeelding 2.

Deze sterstructuren hebben de volgende basisopzet:
- Een feitentabel, waarin de toestanden met datum_begin_toestand en datum_einde_toestand worden vastgelegd. In het tijdvak tussen die twee data zijn de historische kenmerken ongewijzigd;
- Een objectdimensie, waarin de niet-historische kenmerken worden vastgelegd. In deze dimensie wordt tevens de periode vastgelegd waarin het object 'bestaat'. Voor een persoon wordt de periode bepaald door de geboortedatum en de overlijdensdatum. In die periode zijn de historische kenmerken geldig en mogen gebruikt worden voor rapportages en analyses. Vanuit de feitentabel wordt verwezen naar de primaire sleutel van deze dimensie. Deze dimensie wordt in het vervolg de *primaire dimensie* bij de feitentabel genoemd;
- Een objectkenmerken-dimensie, waarin de historisch vast te leggen kenmerken worden vastgelegd. In de feitentabel is een verwijzing opgenomen naar deze tabel. Die verwijzing komt in plaats van de kenmerken zelf. Dit is een technische constructie met als doel het beperken van de benodigde opslagruimte. Deze tabel wordt een minidimensie[5] genoemd;
- Een tijddimensie, die twee keer aan de feitentabel is gekoppeld, voor de datum_begin_toestand en voor de datum_einde_toestand van de periode. In de tijddimensie zijn allerlei kenmerken van een datum opgenomen, zoals de kalendermaand of administratieve periode waartoe die datum wordt gerekend. Andere vaak gebruikte kenmerken van een datum, zijn de indicaties of die datum de laatste werkdag of de laatste vrijdag van de kalendermaand was.

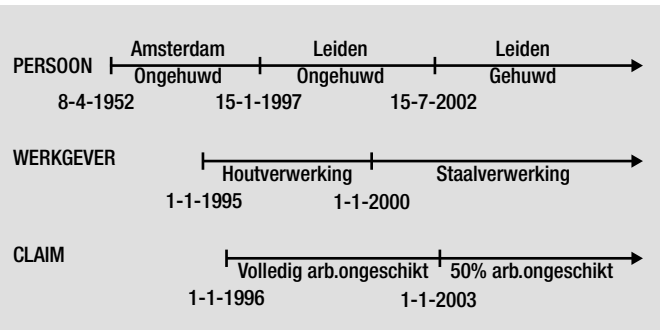
Combineren van gegevens

De meeste informatieverzoeken betreffen gegevens van meerdere objecten, bijvoorbeeld claims naar leeftijd en burgerlijke staat van de aanvrager en regio waarin de bijbehorende werkgever actief is. In het datawarehouse wordt de koppeling tussen de claim enerzijds en de persoon en de werkgever anderzijds al gelegd. Tevens worden alle kenmerken van de persoon en werkgever bij de kenmerken van de claim gevoegd. De historie van de persoon en werkgever wordt gecombineerd met de historie van de claim. Dit wordt het 'overerven' van historie van de persoon en werkgever door de claim genoemd. Er ontstaat hiermee een hiërarchie van sterren die wordt bepaald door de relaties tussen de verschillende objecten. Zo kan een werkgever bij meerdere claims betrokken zijn. Een persoon kan ook meerdere malen in de WAO belanden en dus meerdere claims hebben, zie afbeelding 3.

Het laadproces moet regelen dat de persoonshistorie en de werkgevershistorie die bij de claim is opgeslagen, exact gelijk is aan de historie in de persoon-ster en de historie in de werkgever-ster.
Als voorbeeld; stel een persoon is geboren op 8 april 1952. Deze persoon woont een bepaalde periode in Amsterdam en is verhuisd op 15 januari 1997 naar Leiden en trouwt op 15 juli 2002.
De werkgever van de persoon is begonnen op 1 januari 1995 als



Afbeelding 3: Een persoon kan meerdere claims hebben.



Afbeelding 4: De arbeidsongeschiktheid wordt per 1 januari 2003 omgezet.

Afbeelding 5: De historie van persoon en werkgever in de claimbeschrijving.

houtverwerkend bedrijf. Per 1 januari 2000 verandert de werkgever zijn werkzaamheden en wordt een staalverwerkend bedrijf. De persoon is op 1 januari 1996 volledig arbeidsongeschikt geworden. De arbeidsongeschiktheid wordt per 1 januari 2003 omgezet naar 50 procent. Deze situatie is weergegeven in afbeelding 4. Men wil nu in de beschrijving van de claim de historie van de persoon en de werkgever opnemen, zie afbeelding 5. De gegevens van de claim, de bijbehorende persoon en werkgever worden in één tabel bij elkaar gebracht, zodat de gebruiker alle relevante gegevens met één vraag kan verkrijgen en niet afzonderlijke vragen aan drie tabellen hoeft te stellen en de antwoorden vervolgens weer bijeen te voegen.

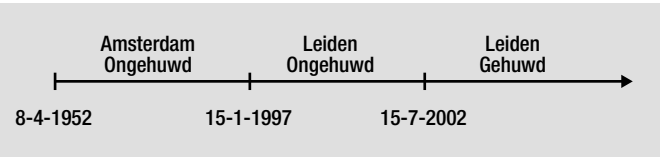
Laadproces

Het laadproces moet dus zorgen voor een juiste reeks perioden (= historie) per object en voor dezelfde historie op verschillende niveaus. Het laadproces moet daarnaast een viertal andere problemen oplossen; mutaties met terugwerkende kracht, verschillende bronnen voor hetzelfde object, meerdere uitspraken over hetzelfde kenmerk en referentiële inconsistentie tussen bronsystemen. Deze problemen worden hierna verder besproken.

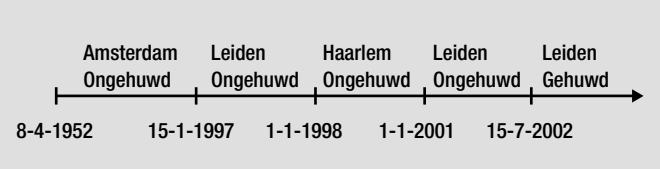
Mutaties met terugwerkende kracht zijn mutaties op een object die betrekking hebben op een vroegere periode van dat object. Bijvoorbeeld, van een persoon zijn de in afbeelding 6 weergegeven toestanden reeds vastgelegd in het datawarehouse. Bekend wordt dat de persoon van 1 januari 1998 tot en met 31 december 2000 in Haarlem gewoond heeft. De nieuwe situatie van de persoon zal dan moeten worden zoals in afbeelding 7.

Het tweede probleem betreft verschillende bronnen voor hetzelfde object. De vier verschillende uitvoeringsinstellingen waaruit het UWV is voortgekomen, gebruiken ieder een eigen systeem voor de opslag van gegevens. Vanuit die systemen zullen de kenmerken van hetzelfde object gecombineerd moeten worden tot een eenduidig geheel. Het laadproces dient zo te worden ingericht dat het niet uitmaakt of de gegevens over de kenmerken voor een object uit één of uit meerdere (verschillende) bronnen voor hetzelfde object kunnen worden geladen. Zo zijn er bijvoorbeeld afzonderlijke bronnen voor de persoonskenmerken burgerlijke staat en nationaliteit enerzijds en adresgegevens anderzijds. Beide bronnen bevatten dezelfde logische sleutel (sofi-nummer) waarmee de kenmerken gecombineerd kunnen worden.
Meerdere uitspraken over hetzelfde kenmerk vormt het derde probleem binnen het laadproces. Het kan gebeuren dat er op enig moment meerdere uitspraken over hetzelfde kenmerk beschikbaar zijn, ieder met een eigen ingangsdatum. Het laadproces moet dan de verschillende uitspraken in de juiste volgorde zetten, zodat de afzonderlijke toestanden gegenereerd kunnen worden voor een historisch kenmerk en de meest recente uitspraak geselecteerd kan worden voor een niet-historisch kenmerk. Als er meerdere uitspraken zijn over één kenmerk met dezelfde ingangsdatum, moet het moment van inbrengen van die mutatie uitsluitsel brengen over de vraag welke uitspraak gelijk heeft.
Met een voorbeeld; er is een bron met informatie betreffende de burgerlijke staat en de woonplaats van personen. Die informatie wordt periodiek aangeleverd en bevat dan alle mutaties die in die periode zijn doorgevoerd in het bronsysteem.
Van een persoon A is al bekend dat hij sinds 1-1-2000 in Appingedam woont en ongehuwd is. Op mutatiedatum 03-04-2002 wordt voor deze persoon bekend dat vanaf 15-2-2000 een burgerlijke staat 'gehuwd' geldt. Uit dezelfde bron komt op mutatiedatum 10-08-2002 de mededeling dat de burgerlijke staat gewijzigd is in 'gescheiden' per 25-2-2002. Tevens bevat het bestand een mededeling met mutatiedatum 15-08-2002 dat persoon A op 20-1-2002 verhuisd is naar Haarlem en tot slot is er de mededeling met mutatiedatum 16-08-2002, dat de persoon sinds 20-1-2002 woonachtig is in Amsterdam. Blijkbaar is de mutatie van 15-08-2002 een vergissing geweest.
Het laadproces dient er voor te zorgen dat de verwerking van deze vier mutaties de juiste toestanden oplevert. Deze mutaties kunnen weergegeven worden in een tabel, zie afbeelding 8. Het resultaat van bovenstaande voorbeelden dient de combinatie van gegevens op te leveren zoals in afbeelding 9.

Het vierde probleem is de referentiële inconsistentie tussen bronsystemen. In de verschillende registratiesystemen kunnen inconsistenties ontstaan, omdat de systemen niet aan elkaar gekoppeld zijn. Zo kan bijvoorbeeld de claimregistratie een claim opnemen zonder dat de persoonsregistratie de persoonsgegevens al heeft geregistreerd. In dat geval wordt dus een claim aangeboden aan het datawarehouse, terwijl de persoon (en de persoonskenmerken) nog niet is aangeboden aan het bronsysteem en dus evenmin is opgenomen in het datawarehouse.
De relatie tussen de claim en de persoon moet behouden blijven. Als van de betreffende persoon later wel kenmerken bekend worden, kunnen die kenmerken 'overerft' worden door de claim. Het laadproces zal met de mogelijke inconsistenties tussen de bronsystemen en de aanleveringen aan het datawarehouse rekening moeten houden.



Afbeelding 6: De in het datawarehouse vastgelegde toestanden.



Afbeelding 7: De nieuwe situatie.

Van	Tot	Situatie		MUTATIEDATUM
		Burgerlijke staat	Woonplaats	
15-2-2000		Gehuwd		03-04-2002
20-1-2002			Haarlem	15-08-2002
20-1-2002			Amsterdam	16-08-2002
25-2-2002		Gescheiden		10-08-2002

Afbeelding 8: Tabel met mutaties.

Van	Tot	Situatie		
		Burgerlijke staat	Woonplaats	
1-1-2000	14-2-2000	Ongehuwd	Appingedam	
15-2-2000	19-1-2002	Gehuwd	Appingedam	
20-1-2002	24-2-2002	Gehuwd	Amsterdam	
25-2-2002		Gescheiden	Amsterdam	

Afbeelding 9: De combinatie van gegevens als resultaat.

Oplossingsrichting

Het laadproces dat past bij het hierboven beschreven ontwerp en dat een oplossing biedt voor de beschreven problemen dient, samenvattend, te voldoen aan de volgende voorwaarden:

1. Aangezien het UWV een recent fusieproduct is, zijn er nog vele bronsystemen die allemaal gegevens aanleveren. Die gegevens moeten verwerkt kunnen worden zonder voor iedere bron een volledig laadproces te bouwen. Dit zou een veel te omvangrijk en daardoor onderhoudsgevoelig systeem tot gevolg hebben;
2. Het laadproces moet de referentiële integriteit tussen de sterren in stand houden;
3. Het laadproces moet rekening houden met mutaties met terugwerkende kracht;
4. Het laadproces dient de 'overerving' van de gegevens te regelen zodat de sterstructuren consistent blijven.

Deze voorwaarden hebben geleid tot een laadproces dat is opgesplitst in een aantal processen die in lagen zijn gegroepeerd. De eerste laag behelst de standaardisatie van aangeleverde bestanden. In deze groep processen worden de aangeleverde bestanden getransformeerd naar een standaardstructuur. Die standaardstructuur wordt bepaald door de historische en niet-historische kenmerken van de sterstructuur, waarin het betreffende bestand moet worden verwerkt. De standaardstructuur is het ideale mutatiebestand voor een bepaalde sterstructuur. Die structuur wordt door Harm van der Lek de *One Attribute Set Interface* genoemd [6]. In het laadproces wordt de structuur ook wel als het 'tussenformaat' aangeduid, aangezien het tussen de bronsystemen en de sterstructuren in staat. Door de standaardisatie is de verwerking van de gegevens onafhankelijk gemaakt van de toevallige vorm waarin ze worden aangeleverd.
De tweede laag is de verwerking van die gegevens waarvan geen historie wordt bijgehouden. In deze tweede laag worden de gegevens verwerkt in de primaire dimensies van de verschillende sterstructuren. Omdat van deze gegevens alleen de actuele waarde relevant is, kunnen de oude gegevens worden overschreven.
Dan volgt de derde laag, de verwerking van die gegevens waarvan de historische waarden wel relevant zijn. In deze laag ten slotte worden de tijdvakken bepaald waarin de historische kenmerken ongewijzigd zijn. Deze laag moet een oplossing bieden voor de overerving van historie (het 'doorschrijven' van mutaties op een object naar de daar aan gerelateerde objecten), voor mutaties met terugwerkende kracht en voor de 'geschiedenis van de historie'.

Conclusie

Het gelijktijdig verwerken van overerving en mutaties met terugwerkende kracht leek een onoplosbare kluwen van problemen en deeloplossingen te zijn. De totaaloplossing die is gevonden behelst het indelen van de deeloplossingen in lagen en het afwerken ervan in een specifieke volgorde. In twee komende artikelen in dit tijdschrift zal de totaaloplossing verder uitgewerkt worden.

Paul Hulst en **Fons Rooijers** zijn betrokken bij het datawarehouse-project van het UWV.
E-mail: phulst@deloitte.nl en fons.rooyers@uwv.nl

Literatuur

1. *Harm van der Lek: Geschiedenis van de historie, DB/M 7-2001.*
2. *Bijvoorbeeld een universe voor Business Objects.*
3. *R. Kimball: The Data Warehouse Toolkit, 1996. Naast het stapelen onderkent hij ook andere manieren voor het bijhouden van de historie.*
4. *Michael Schmitz: Sterren en dimensies - Ontwerp en onderhoud van datawarehouses, Deel III, DB/M Essay 2000.*
5. *R. Kimball: The Datawarehouse Toolkit, 1996.*
6. *Harm van der Lek: Het pletten van een ster, DB/M 6-1998.*