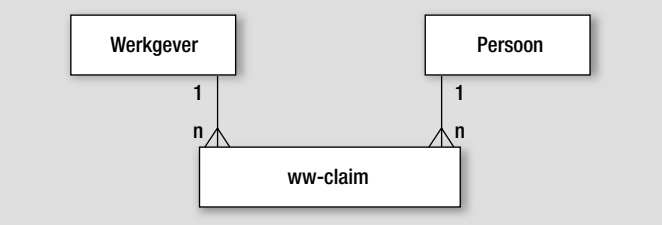
Veel voordelen aan generieke opzet

Het laden van toestanden: de opbouw

Paul Hulst en Fons Rooijers

Met dit artikel besluiten Paul Hulst en Fons Rooijers hun drieluik over het laadproces. De derde laag wordt besproken, waarin de verwerking van gegevens plaats vindt waarvan de historische waarden wel relevant zijn. Een voorbeeld over personen en ww-claims zal gebruikt worden om de afzonderlijke stappen verder uit te werken.

In het eerste artikel is een hiërarchie van sterren voorgesteld waarin een ww-claim verwijst naar exact een persoon en naar exact een werkgever, zie afbeelding 1. Per persoon of werkgever kunnen er meerdere ww-claims zijn. Het zouden bijvoorbeeld ww-claims uit hoofde van de werkloosheidswet kunnen zijn. Alle historische kenmerken van personen en werkgevers zijn opgenomen bij de historische kenmerken van de ww-claim zelf om het onderzoeken van de ww-claim naar kenmerken van personen en werkgevers te vergemakkelijken. Het hoe en waarom van deze opzet is in DB/M5 uitgelegd. De sterstructuur van de ww-claim zou er uit kunnen zien zoals in afbeelding 2 is weergegeven.
In de feitentabel ww-claim vinden we, naast de begin- en einddatum van de toestand, verwijzingen naar de historische kenmerken van de ww-claim en naar de persoons- en werkgeverskenmerken van de claim. Het tussenformaat (de platgeslagen ster) die hoort bij deze structuur bevat de kenmerken zoals in afbeelding 3 staan. Om het voorbeeld beperkt te houden wat betreft ruimtebeslag zal een beperkt aantal historische kenmerken gevolgd worden, te weten de burgerlijke staat van de persoon en het kantoor dat de ww-claim behandelt. Tevens wordt alleen gekeken naar de ster ww-claims.



Afbeelding 1: Hiërarchie van de sterren werkgever, persoon en ww-claim.

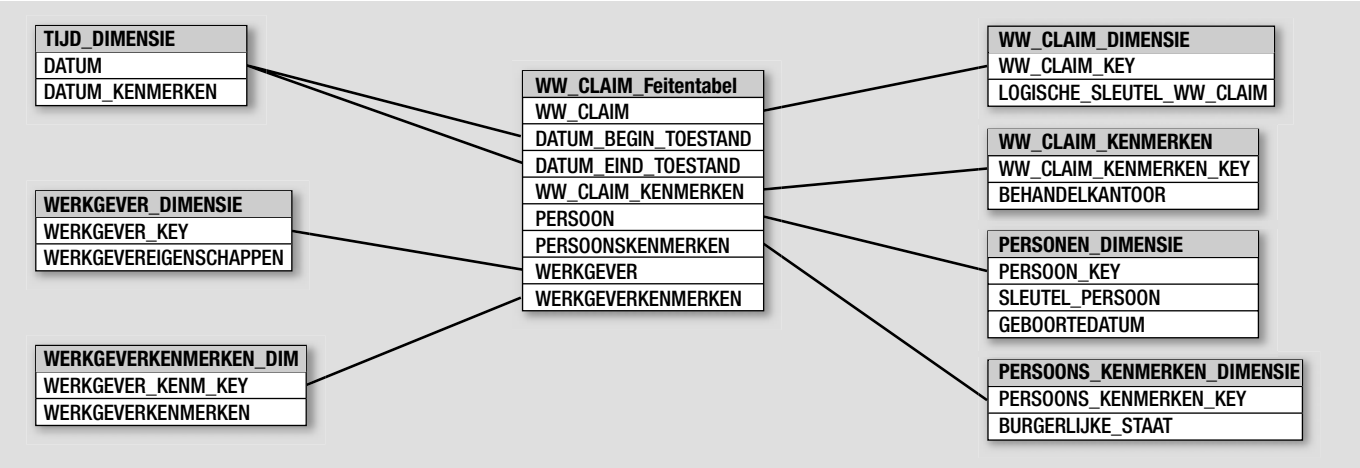
In het datawarehouse is een persoon bekend onder sofi-nummer 123456789. Bekend is verder dat hij ongehuwd is vanaf 18-04-1980 (zijn geboortedatum). Die persoon heeft een claim uit hoofde van de WW lopen vanaf 1 februari 2002, voor één van zijn twee parttime banen. Kantoor Diemen verwerkt de ww-claim die bekend is onder nummer ww22. Op 1 april 2002 worden de mutaties die in maart 2002 zijn ingevoerd, verwerkt in het datawarehouse. Persoon 123456789 is op 2 maart 2002 getrouwd en is per 15 maart 2002 ook uit zijn tweede baan ontslagen. Hij heeft daar een ww-claim voor ingediend, die door Diemen wordt behandeld. Op 20 maart 2002 zijn beide ww-claims overgedragen aan behandelkantoor Amstelveen.

Resultaten van de eerste twee lagen

De stappen in de eerste laag hebben de aangeleverde bestanden gecontroleerd en ingelezen. Tevens zijn ze gestandaardiseerd naar een formaat dat gebaseerd is op de sterstructuur waarvoor ze als invoer dienen. In de tweede laag zijn nieuwe objecten toegevoegd aan de primaire dimensietabellen en zijn de niet-historische kenmerken bijgewerkt aan de hand van de actuele kennis.
In de _DWL tabel zijn alle mutaties verzameld die iets zeggen over de kenmerken waarvan de veranderingen wel belangrijk worden geacht. In afbeelding 4 is aangegeven welke processen tot nu toe zijn uitgevoerd.

Het laden van toestanden (3)

Dit drieluik begon (DB/M5) met de beschrijving van de informatiebehoefte van het UWV en tevens welke wensen er zijn voor het ontwerp van het datawarehouse en voor de bijbehorende laadprogrammatuur. Tevens is er een laadproces geschetst dat uit drie lagen bestaat. In het tweede artikel (DB/M6) zijn de eerste twee lagen van het laadproces besproken.
De lagenstructuur is geïntroduceerd om de complexiteit van het laadproces te minimaliseren. Dit wordt enerzijds bereikt door het laadproces voor iedere ster in het datawarehouse uit exact dezelfde stappen te laten bestaan. Die stappen zijn op dusdanige wijze in lagen en sublagen gegroepeerd, dat er slechts enkele sublagen zijn waarin de relaties tussen de sterren in het datawarehouse van belang is. In alle overige sublagen kunnen de stappen van de afzonderlijke naast elkaar uitgevoerd worden.



Afbeelding 2: Sterstructuur ww-claim.

In de lagen 1 en 2 is de tweede ww-claim toegevoegd aan dimensie ww-claim met sleutel ww34. Tevens is ww-claims_DWL tabel gevuld met de waarden in afbeelding 5.
De rijen met volgnummers 20 en 21 bevatten de verandering van behandelkantoor. Rij 34 bevat de historische kenmerken van de nieuwe ww-claim. Rij 0 bevat de historie van de persoon voor de nieuwe ww-claim, die in processtap 2.D.4 is opgehaald uit de personen_ster.

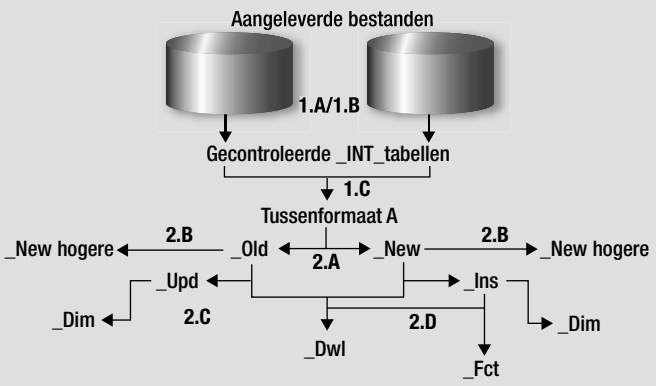
De derde laag

Nadat in de tweede laag de niet-historische kenmerken van de verschillende objecten zijn bijgewerkt, worden in laag 3 de historisch relevant geachte kenmerken verwerkt. Startpunt van deze laag vormt de zogenaamde _DWL tabel, waarin alle relevante mutaties voor de verschillende objecten zijn opgenomen. In de dweiltabel zijn alleen de historische kenmerken opgenomen.
In sublaag 3.A worden de gewenste toestanden bepaald.
De _DWL tabel van een bepaald object bevat op dit moment in het proces een groot aantal mutaties uit verschillende bronnen.
In deze sublaag worden deze mutaties samengevoegd tot voor de objecten gewenste toestanden, rekening houdend met mutaties die elkaar tegenspreken of aanvullen.

Processtap 3.A.1 betreft het ophalen van mutaties uit hogere sterren. Als de ster die nu bijgewerkt wordt verwijzingen naar andere sterren bevat, worden de kenmerken van die hogere ster overgenomen. De historie van de persoon 123456789 is dus redundant opgenomen bij de ww-claims-ster. Als er vervolgens mutaties plaatsvinden op de historische kenmerken van een persoon, moeten die mutaties natuurlijk ook doorgevoerd worden op de ww-claims van die persoon.
In deze processtap worden de mutaties van de persoon toegevoegd aan de _DWL tabel van de ww-claim. De mutaties die zijn uitgevoerd voor persoon 123456789 worden zodoende dus beschouwd als mutaties op de ww-claims ww22 en ww34.
De bron van de persoonsmutaties is de zogenaamde replicatietabel van de personen-ster. Deze tabel wordt aan het einde van deze sublaag 3.A gevuld. De eerste stap in sublaag 3.A van de ww-claim is dus afhankelijk van de laatste stap van sublaag 3.A bij de personenster. Deze paradox zal bij laag 3.A Afhankelijkheid worden opgelost, daarover verderop meer. Het huwelijk van persoon 123456789 wordt nu aan ww-claims_DWL tabel toegevoegd, zie afbeelding 6.
In processtap 3.A.2 wordt de reeds bekende situatie opgehaald. In de _DWL tabel zijn nu alle mutaties op de ww-claims verzameld,

WW_CLAIM_TUSSENFORMAAT	
SLEUTEL_WW_CLAIM	uit ww_claim_dimensie
DATUM_BEGIN_TOESTAND	uit ww_claim_feitentabel
DATUM_EINDE_TOESTAND	uit ww_claim_feitentabel
BEHANDELKANTOOR	uit ww_claim_kenmerken_dimensie
SLEUTEL_PERSOON	uit personen_dimensie
GEBOORTEDATUM	uit personen_dimensie
BURGERLIJKE_STAAT	uit persoons_kenmerken_dimensie
WERKGEVEREIGENSCHAPPEN	uit werkgever_dimensie
WERKGEVERKENMERKEN	uit werkgeverkenmerken_dimensie
RUN_NUMMER	Technische aanduiding
VOLGNUMMER	Technische aanduiding

Afbeelding 3: Het tussenformaat (de platgeslagen ster).



Afbeelding 4: Het laadproces van lagen 1 en 2.

sleutel ww-claim	datum begin toestand	datum einde toestand	behandel kantoor	sleutel persoon	burgerlijke staat	datumtijd-mutatie	run-nummer	volg-nummer
ww22	20-03-2002		Amstelveen	123456789		20-03-2002	23	20
ww34	15-03-2002		Diemen	123456789		17-03-2002	17	34
ww34	18-04-1980			123456789	Ongehuwd	17-03-2002	17	0
ww34	20-03-2002		Amstelveen	123456789		20-03-2002	23	21

Afbeelding 5: De ww-claims_DWL tabel.

sleutel ww-claim	datum begin toestand	datum einde toestand	behandel kantoor	sleutel persoon	burgerlijke staat	datumtijd-mutatie	run-nummer	volg-nummer
ww22	02-03-2002			123456789	Gehuwd	04-03-2002	12	87
ww34	02-03-2002			123456789	Gehuwd	04-03-2002	12	87

Afbeelding 6: Nieuwe ww-claims_DWL tabel.

sleutel ww-claim	datum begin toestand	datum einde toestand	behandel kantoor	sleutel persoon	burgerlijke staat	datumtijd-mutatie	run-nummer	volg-nummer
ww22	01-02-2002	31-12-9999	Diemen	123456789	Ongehuwd	nvt	2	nvt
ww34	01-01-1900	31-12-9999	Onbekend	123456789	Onbekend	nvt	17	nvt

Afbeelding 7: De ww-claims_TOP tabel.

niet alleen de directe mutaties uit de bronbestanden, maar ook de mutaties uit sterren waarnaar verwezen wordt. Die mutaties moeten doorgevoerd worden op aanwezige kennis in de feittabel. De rijen in de _DWL tabel vormen immers de mutaties op de reeds aanwezige feitrijen in de feittabel.
Deze processtap haalt die gegevens uit de feittabel voor alle objecten waar mutaties voor aanwezig zijn in de _DWL tabel. Alleen de noodzakelijke toestanden worden opgehaald. Toestanden die eindigen voor de vroegste wijzigingsdatum (= datum_begin_toestand) worden toch niet veranderd en worden dus ook niet opgehaald. De te wijzigen toestanden worden in een aparte tabel geplaatst, genoemd _TOP. De structuur is gelijk aan de structuur van _DWL tabel.

Dat betekent dat de in de feittabel aanwezige technische sleutels moeten worden vervangen door de door de sleutels gerepresenteerde dimensierijen, dit wordt ook wel de-collapsen genoemd. De ww-claims_TOP tabel bevat nu de waarden zoals afgebeeld in afbeelding 7. De rij voor ww-claim ww22 was al in het datawarehouse aanwezig, de rij voor ww-claim ww34 is toegevoegd in processtap 2.D.5.
In processtap 3.A.3 worden de mutaties in tabel _DWL uitgevoerd op de aanwezige kennis in tabel _TOP. Dit wordt het 'dweilproces' genoemd. Hoe dit precies werkt is beschreven door Harm van der Lek in het artikel 'Dweilen met de kraan open' (DB/M 8-2001). De essentie is dat mutaties uit verschillende bronnen in volgorde worden gezet met de reeds aanwezige kenmerken. Vervolgens wordt steeds de nieuwe waarde van een kenmerk gedweild over een reeds aanwezige waarde. Hierbij worden de 'gaten' in de reeds aanwezige kennis opgevuld met de nieuwe waarden.

Dweilproces

Het dweilproces levert een resultaattabel op (_RES) waarin de nieuwe historische opbouw van een object is opgenomen. Het dweilproces heeft ook bepaald hoe de resultaten in de database verwerkt moeten worden. Iedere rij in de _RES tabel heeft daarvoor een van de volgende codes gekregen:

- DE: (delete) verwijderen uit de database;
- DI: (delete/insert) verwijderen van de originele rij en vervangen door een rij met andere kenmerken voor een periode met dezelfde begin- en einddatum;
- UP: (update) wijzigen van de einddatum van de toestand, met andere woorden de periode waarin de kenmerken geldig zijn wordt gewijzigd;
- IN: (insert) toevoegen aan de database;

- NO: (no operation) negeren, deze rij hoeft niet verwerkt te worden in de database.

De _RES tabel (afbeelding 8) heeft een nieuwe inhoud gekregen. In de vorige processtap is de gewenste situatie bepaald voor een toestandster waarin de mutaties zijn doorgevoerd op de aanwezige kennis. Als een andere toestandsster naar deze toestandsster verwijst, dan moet die toestandsster de wijzigingen overnemen. Processtap 3.A.4 behelst het aanmaken van een replicatietabel. Daartoe wordt nu een tabel (_REP) gevuld met de rijen uit de _RES tabel die gemarkeerd zijn met code IN of DI: dat zijn namelijk de rijen die kenmerken veranderen.
Als er een ster (lager in hiërarchie) zou zijn die verwijst naar de ww-claims-ster, dan zouden dus alle rijen uit de _RES tabel opgenomen worden met uitzondering van de rijen met rijnummers nvt.

Afhankelijkheid

In processtap 3.A.1 is de paradox beschreven dat die stap afhankelijk is van een stap die later in het laadproces plaatsvindt.
De oplossing is dat processtap 3.A.1 afhankelijk is van processtap 3.A.4 van een ANDERE ster. Deze sublaag moet dus in een specifieke volgorde plaatsvinden: als een ster (A) een verwijzing naar een andere ster (B) bevat, dan moet sublaag 3.A voor ster B al gedaan zijn voordat sublaag 3.A voor deze ster A kan beginnen. De volgorde van verwerking verloopt dus volgens de hiërarchie van de verschillende sterstructuren.
Het stelsel van sterren in het data warehouse vormt een hiërarchie. In die hiërarchie mogen geen cirkels voorkomen ($A \rightarrow B \rightarrow C \rightarrow A$), dan worden eerst de sublagen 3.A gedaan voor sterren die geen verwijzingen bevatten, dan de sterren die alleen naar die sterren verwijzen en zo verder.
De tabellen waarin de toestanden worden bewaard, worden zeer groot: 60 miljoen rijen is geen uitzondering. Om het ruimtebeslag te beperken wordt de specifieke combinatie van kenmerken die bij een toestand hoort vervangen door een technische sleutel in de vorm van een nummer. Een nummer neemt natuurlijk aanzienlijk minder ruimte in dan een combinatie van kenmerken die bestaat uit karakterreeksen. Het nummer en de combinatie van kenmerken die het vervangt worden in aparte tabel ondergebracht. Een dergelijk tabel wordt een mini-dimensie genoemd.
In de sublaag 3.B worden combinaties van kenmerken die nog niet voorkomen in de mini-dimensie maar wel in de _RES tabel, opgespoord en toegevoegd aan die mini-dimensie. De processen in deze sublaag kunnen parallel aan elkaar uitgevoerd worden.

Het bijwerken van de feittabellen gebeurt in sublaag 3.C. De mutaties kunnen nu daadwerkelijk uitgevoerd worden. In processtap 3.A.3 Dweilen is bepaald welke operaties uitgevoerd moeten worden en in sublaag 3.B is er voor gezorgd dat alle technische sleutels aanwezig zijn. De rijen worden in drie groepen onderverdeeld:
1. te verwijderen rijen uit de database (codes DE en DI);
2. te wijzigen rijen in de database (code UP);
3. toe te voegen van rijen aan de database (codes IN en DI).

In deze volgorde worden de rijen uit de groepen verwerkt. De reden hiervoor is het voorkomen van meerdere rijen met dezelfde technische sleutel (object, datum_begin_toestand). Voor de rijen uit de laatste groep moet de combinatie van de kenmerken ook nog vervangen worden door de overeenkomende technische sleutel. De processen in deze sublaag kunnen parallel aan elkaar uitgevoerd worden.

Controles

Essentieel voor acceptatie van het datawarehouse door gebruikers is dat de gegevens consistent zijn. Dat betekent onder andere dat ouder/kind-relaties moeten kloppen, de toestanden elkaar exact opvolgen en de begindatum van de ene toestand gelijk is aan de einddatum van de vorige toestand min één dag. Als een ster een verwijzing naar een andere ster bevat, dan moet de historie van een object in de verwijzende ster exact gelijk zijn aan de historie van het bijbehorende object in de andere ster. In het gegeven voorbeeld moet de historie van de persoon in de personenster gelijk zijn aan de historie van dezelfde persoon voor wie een ww-claim is geregistreerd.
Dit zijn enkele voorbeelden van de controles die na afloop van het

sleutel ww-claim	datum begin toestand	datum einde toestand	behandel kantoor	sleutel persoon	burgerlijke staat	datumtijd-mutatie	run-nummer	volg-nummer	db_code
ww22	01-02-2002	01-03-2002	Diemen	123456789	Ongehuwd	nvt	2	nvt	UP
ww22	02-03-2002	19-03-2002	Diemen	123456789	Gehuwd	04-03-2002	12	87	IN
ww22	20-03-2002	31-12-9999	Amstelveen	123456789	Gehuwd	20-03-2002	23	20	IN
ww34	01-01-1900	17-04-1980	Onbekend	123456789	Onbekend	nvt	17	nvt	UP
ww34	18-04-1980	01-03-2002	Onbekend	123456789	Ongehuwd	17-03-2002	17	0	IN
ww34	02-03-2002	14-03-2002	Onbekend	123456789	Gehuwd	04-03-2002	12	87	IN
ww34	15-03-2002	19-03-2002	Diemen	123456789	Gehuwd	17-03-2002	17	34	IN
ww34	20-03-2002	31-12-9999	Amstelveen	123456789	Gehuwd	20-03-2002	23	21	IN

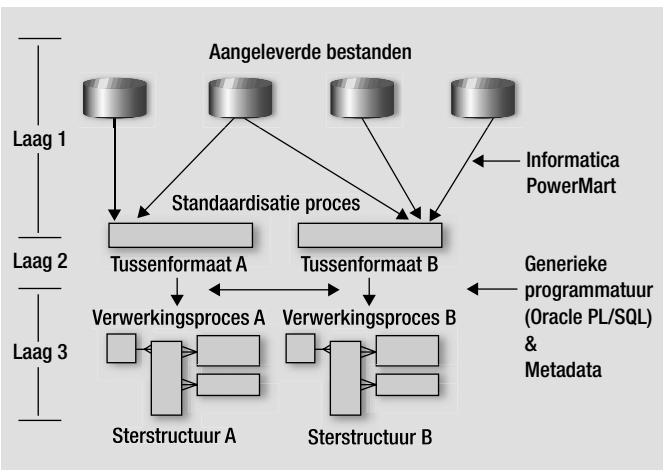
Afbeelding 8: De _RES tabel met nieuwe inhoud.

laadproces worden uitgevoerd om te garanderen dat de gegevens goed zijn verwerkt. Voor deze controles is een apart programma in gebruik. In dat programma zijn query's opgenomen die de boven beschreven controles uitvoeren. Het programma voert de query's uit en rapporteert welke rijen niet voldoen.

Het laadproces is met behulp van twee programma's gebouwd; Informatica PowerMart en Oracle PL/SQL. Daarnaast er een programma in gebruik voor het onderhouden van de metadata. Informatica PowerMart wordt gebruikt voor de eerste laag. Het is een grafisch programma waarmee uitstekend de complexe transformaties in sublaag 1.C gebouwd en onderhouden kunnen worden. Tevens kan het goed omgaan met diverse soorten bronnen. De processtappen in lagen 2 en 3 zijn, dankzij het tussenformaat, afhankelijk van de ster die ze bijwerken en niet van de bronbestanden. En het tussenformaat wordt bepaald door de betreffende ster. Dit heeft het mogelijk gemaakt om een generieke oplossing te kiezen, gebaseerd op metadata. In een apart programma worden de kenmerken in een ster beschreven en of er historie moet worden bijgehouden van dat kenmerk. Tevens is de hiërarchie van de sterren opgenomen.

Genlaad

Deze informatie wordt gebruikt door het programma Genlaad, geschreven in Oracle PL/SQL. Het leest de informatie uit de metadata-database en genereert daar run-time dynamisch SQL voor, rekening houdend met de afhankelijkheden tussen de lagen en processtappen en in de juiste volgorde van sterren, indien relevant. Deze code wordt vervolgens door de Oracle database uitgevoerd. Afbeelding 9 toont de opzet.
Er zijn veel voordelen aan deze generieke opzet verbonden.
In een eerdere fase van het project is de programmatuur voor de lagen 2 en 3 met de hand gebouwd in Informatica PowerMart. Dat leverde veel code op en het kostte veel tijd en moeite bij het maken en testen. Die code hoeft nu niet meer gemaakt te worden. Het testen van de generieke code is eerder al zeer uitvoerig gedaan en hoeft dus niet meer herhaald te worden voor een specifieke ster. Het testproces kan dus beperkt blijven tot laag 1. Het is daardoor mogelijk geworden om snel wijzigingen door te voeren in de sterstructuren. Eén of meer kolommen extra in een ster betekent het invoeren van de gegevens in de metadata-programma. In Informatica PowerMart moet de bron beschreven worden en de transformatie van de brongegevens naar het tussenformaat aangegeven worden door lijntjes te trekken. Deze programmatuur maken kost weinig tijd.
Door de programmatuur van het programma Genlaad zelf te verbeteren, wordt die verbetering in één keer voor alle sterren doorgevoerd. Ook dat scheelt veel tijd in vergelijking tot een aanpak, waarbij voor iedere ster aparte programmatuur op dezelfde gewijzigd moet worden.



Afbeelding 9: Laadproces met programma's.

Voor de beschrijving van de metadata wordt gebruik gemaakt van een zelf ontwikkeld programma waarin de sterstructuren worden beschreven. Van deze sterstructuren wordt (naast de beschrijving van de elementen in de sterstructuur) ook vastgelegd welke hiërarchie tussen de sterren bestaat en welke overerving van gegevens tussen de verschillende sterren plaats vindt.

Conclusie

In dit drieluik is een laadproces voor toestandgeoriënteerde sterstructuren beschreven en zijn de eisen benoemd die aan het laadproces gesteld worden, gedicteerd door het betreffende datawarehouse. Deze eisen zijn:

1: Aangezien het UWV een recent fusieproduct is, zijn er nog vele bronsystemen die allemaal gegevens leveren. Die gegevens moeten verwerkt kunnen worden zonder voor iedere bron een apart laadproces te bouwen. Dit zou een veel te omvangrijk en daardoor onderhoudsgevoelig systeem tot gevolg hebben. Aan deze eis wordt voldaan door het formeren van het tussenformaat en door het laten genereren van een deel van de programmatuur aan de hand van metadata;
2: Het laadproces moet de referentiële integriteit tussen de sterren in stand houden. Met behulp van de geautomatiseerde controles achteraf kan de consistentie van de sterstructuren gegarandeerd worden;
3: Het laadproces moet rekening houden met mutaties met terugwerkende kracht. Het 'dweilproces' zorgt ervoor dat die mutaties in de sterstructuren verwerkt worden;
4: Het laadproces dient de 'overerving' van de gegevens te verzorgen met behoud van referentiële integriteit. De overerving van gegevens wordt verzorgd door de processtappen 2.D.4 Ophalen historie hogere sterren en 3.A.1 Ophalen van mutaties uit hogere sterren.

Dit laadproces zal verder ontwikkeld worden om een betere invulling te geven aan de huidige eisen en om te voldoen aan nieuwe eisen ten aanzien van functionaliteit.

Fons Rooijers (fons.rooijers@uwv.nl) en **Paul Hulst** (phulst@deloitte.nl) zijn beide betrokken bij het datawarehouse-project van het UWV. Fons Rooijers werkt bij het UWV, Paul Hulst bij Deloitte & Touche.