Implementatie datawarehouse-project KLM Passage

Geïntegreerd tool support: praktijk-case

Guido Bakema en Elton Manoku

In het eerste artikel in DB/M 1 werd een model driven-aanpak besproken voor datawarehouse-ontwerp. Eerst wordt een elementair conceptueel informatiemodel opgesteld, dat wordt getransformeerd naar een genormaliseerd en vervolgens naar een gedenormaliseerd dimensioneel model.

Ook deze modellen zijn nog volledig gericht op de gebruikersomgeving en kunnen worden gevalideerd. Via instantane bruggen kan op elk moment worden afgedaald naar het logische niveau (Entity-Relationship-diagrammen of desgewenst UML Class-diagrammen), van waar verder kan worden afgedaald naar het fysieke niveau. Een en ander wordt ondersteund door het FCO-IM modellerings-tool CaseTalk en een FCO-IM Bridge tool set die daarbij naadloos aansluit. De werkwijze en de tools werden uitvoerig getest in de praktijk en met succes toegepast, onder meer in het datawarehouse-project KLM Passage.
De in [1] besproken horizontale en verticale lagen-architectuur wordt in afbeelding 1 in een enigszins vereenvoudigde vorm getoond.

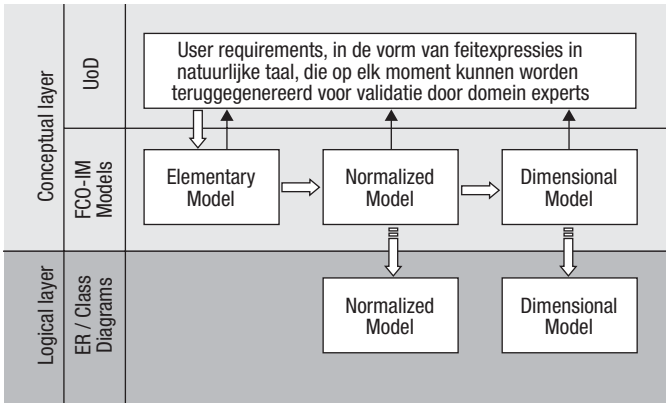
De FCO-IM Bridge toolset kent meerdere repository's

De tool support is van onderscheiden aard [2]:

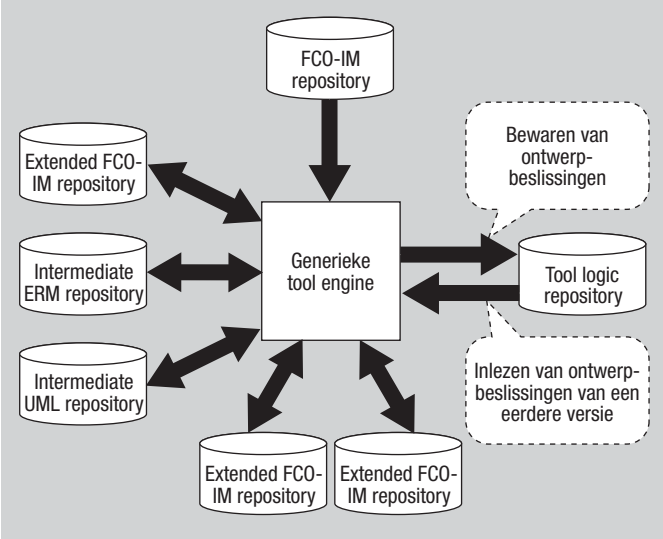
- Met het FCO-IM modellerings-tool CaseTalk worden feitexpressies op type-niveau gebracht, geanalyseerd (geklassificeerd en gekwalificeerd) en vervolgens gediagrammatiseerd en aangevuld met standaard beperkingsregels (uniciteitsregels, totaliteitsregels, waardenregels, etcetera). Ook biedt CaseTalk een aantal transformaties (subtypering, nominalisatie, etcetera), die het mogelijk maken meer structuur en semantiek toe te voegen aan het elementaire model, alvorens dat wordt getransformeerd naar een genormaliseerd model met de in het vorige artikel besproken transformaties (groeperen en reduceren, en desgewenst lexicaliseren);
- Met behulp van de modulen StarBridge en StarSplit van de FCO-IM Bridge tool set kan op basis van een een genormaliseerd conceptueel model een gedenormaliseerd dimensioneel model worden gegenereerd;
- De modulen ERM Bridge en UML Bridge bieden démasqué-algoritmen voor de overstap naar de logische en fysieke lagen. Ze slaan een brug van de FCO-IM wereld naar de wellicht meer vertrouwde wereld van de Entity-Relationship diagrammen en UML Class-diagrammen. Zowel genormaliseerde als gedenormaliseerde (dimensionele) modellen kunnen worden geëxporteerd en vervolgens geïmporteerd in ERM- en UML-tools.

Volledig repository based tool-architectuur

CaseTalk en de FCO-IM Bridge tool set zijn repository based. In CaseTalk gebeurt de modellering en de uitvoering van de beschreven transformaties volledig op basis van wijzigingen in de populatie van deze FCO-IM repository. De FCO-IM Bridge toolset kent meerdere repository's. De modules ERM Bridge, UML Bridge maken gebruik van kopieën van de FCO-IM repository voor het overnemen van de metadata uit CaseTalk. Om de afstand tot de diverse ERM en UML tools (PowerDesigner, ERwin, ...) en hun opvolgende versies zo klein mogelijk te houden, maken ze gebruik van eigen intermediate ERM en UML repository's. De StarBridge/StarSplit-module maakt gebruik van een in vergelijking met die van CaseTalk enigszins uitgebreide FCO-IM repository, waarin ook typische dimensionele informatie wordt vastgelegd.



Afbeelding 1: De horizontale en verticale lagen-architectuur.



Afbeelding 2: De architectuur van de FCO-IM Bridge tool set.

De in-repository en repository-to-repository transformatie-algoritmen bestaan uit enkele honderden SQL-instructies in de vorm van condities (query's) en acties (insert/update/delete instructies), welke zijn vastgelegd in een tool repository, samen met informatie als instructietype, plaats in de desbetreffende transformatie-algoritmen. Het geheel staat onder regie van een generieke tool engine, die feitelijk niets anders doet dan instructies volgordelijk ophalen uit deze tool repository en ze uitvoeren. Het uitvoeren van condities betekent dat er al dan niet een volgende instructie of groep van instructies wordt opgehaald en uitgevoerd. Terug naar het 'stored program principe' in z'n meest simpele vorm: een zo consequent mogelijke data-optiek voor de algoritmische logica, waardoor de FCO-IM Bridge tool set volledig transparant is en daardoor onderhoudbaar en uitbreidbaar [3].
In de tool repository waarin deze algoritmische logica ligt vastgelegd, worden ook de door de analist genomen ontwerpbeslissingen (gekozen algoritme, gemaakte keuzen tijdens de uitvoering van het algoritme, opdeling in submodellen) bewaard. Deze stuurinformatie is daarmee te controleren, eventueel terug te draaien en ook her te gebruiken voor een volgende versie van het datawarehouse, zie afbeelding 2.

Ook user interface-gerelateerde logica en informatie is vastgelegd in de tool logic repository:

- betreffende het dynamiseren van de interface: bijvoorbeeld het afdwingen van zekere volgordelijkheid van interface-stappen op basis van een gemaakte keuze;
- terminologiekeuze en taalafhankelijkheid daarvan: naamgeving van de transformatie-algoritmen en hun delen ('Dimension', 'Do not reduce', 'Snowflake'), de messages, de captions, etcetera.

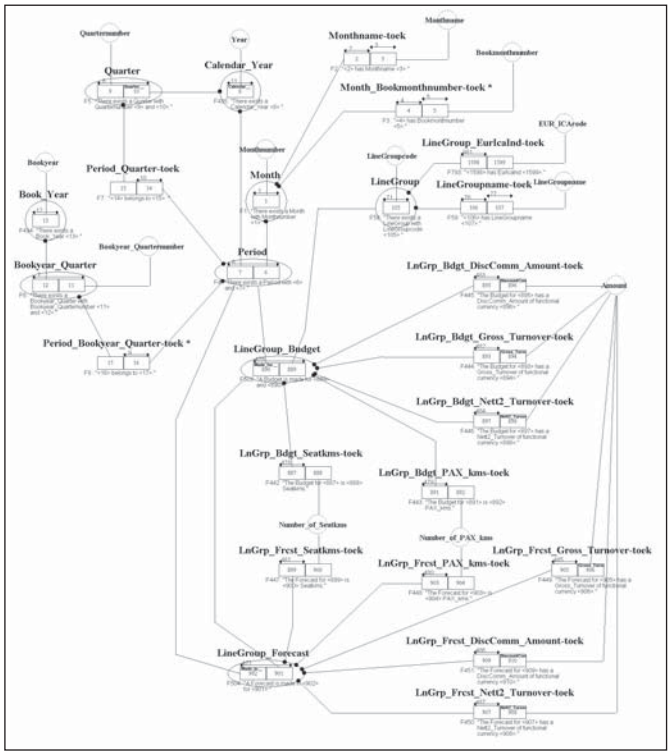
De CaseTalk modellerings-tool gebruikt de basis FCO-IM Repository, waarin de volgende conceptuele modellen opgeslagen zijn: elementaire en gegroepeerde & gereduceerde FCO-IM informatiegrammatica. De ER Tool Repository en UML-Class Diagram Tool Repository zijn externe ER of UML Tool repository's (Allfusion ERwin 4.1, Sybase PowerDesigner 9) waarin de geëxporteerde modellen opgeslagen kunnen worden.

Toepassing in het DWH-project KLM Passage

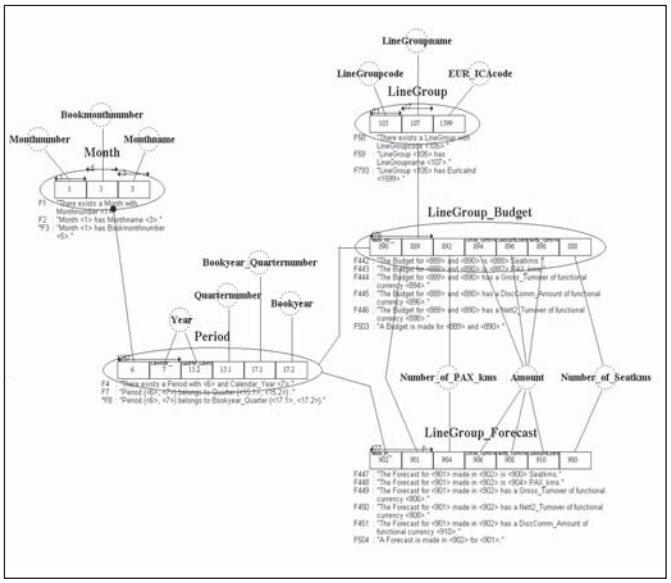
Het KLM Passage DWH is een corporate datawarehouse ontwikkeld voor de Passenger Division van Royal Dutch Airlines KLM. Voor het project werd efficiënt en – voor zover mogelijk – 'single-point-of-definition' metadata-beheer cruciaal geacht. De in het eerste artikel beschreven aanpak werd in dit project eerst uitvoerig getest en vervolgens integraal toegepast. Peter Alons van Atos Origin en Karel Duran van KLM waren hierbij betrokken als data managers. Een en ander heeft geleid tot het door Atos Origin gelanceerde Metadata Framework [4], [5].

"There exists a Month with Monthnumber 04."
"Month 04 has Monthname April."
"Month 04 has Bookmonthnumber 01."
"Period (01, 1998) belongs to Bookyear_Quarter (4, 1997/1998)."
....
"There exists a LineGroup with LineGroupcode A."
"LineGroup A has LineGroupname Europe."
"A Budget is made for LineGroup A and Period (04, 1998)."
....
"The Budget for LineGroup A and Period (04, 1998) has a DiscComm_Amount of functional currency 5."
"The Budget for LineGroup A and Period (04, 1998) has a Gross_Turnover of functional currency 25."
....

Afbeelding 3: Enkele feitexpressies uit het submodel 'Forecast' van het KLM Passage DWH.



Afbeelding 4: Het elementaire submodel 'Forecast' van het datawarehouse model KLM Passage.



Afbeelding 5: Het genormaliseerde submodel 'Forecast'.

De gebruikte modellerings-tools zijn het FCO-IM modelleringstool CaseTalk, de FCO-IM Bridge tool set [2] en ERwin 4.1. Het aantal feittypen in het elementaire FCO-IM model bedraagt ruim 600. Het logische Entity-Relationship model dat hieruit werd gegenereerd telt ongeveer 180 entiteittypen. Het gegenereerde dimensionele model heeft 16 fact tables en 37 gedenormaliseerde dimension tables met 177 verwijzingen vanuit de fact tables naar de dimension tables. Op dit moment zijn 13 afzonderlijke data marts – elk bestaande uit 1 of 2 sterren – in gebruik. Deze data marts worden wekelijks of maandelijks geladen met data uit vijf verschillende bronsystemen en van het Internet.
Om te demonstreren hoe de StarBridge- en StarSplit-algoritmes werken, lichten we uit het KLM Passage-model het submodel 'Forecast'. Dit submodel leidt tot het kleinste van de 13 data marts, bestaande uit twee fact tables en twee conforme dimensietabellen waarnaar vanuit beide fact tables wordt verwezen. Na het interviewen van potentiële gebruikers van het datawarehouse werden, rekening houdend met de in de bronsystemen aanwezige informatie, de in afbeelding 3 getoonde feitexpressies opgesteld. Deze feitexpressies zijn elementair en hebben de vereiste granulariteit.

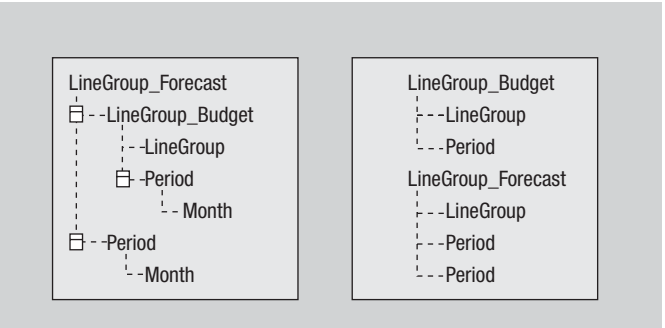
De volgende stappen zijn de klassificatie en kwalificatie van deze feitexpressies en het toevoegen van standaard beperkingsregels. Hiervoor werden opnieuw domeinexperts geïnterviewd. De harde en zachte semantiek liggen nu vast in de FCO-IM repository van CaseTalk, dat vervolgens het elementaire FCO-IM model van afbeelding 4 laat zien.
Voor het verkrijgen van een genormaliseerd FCO-IM model werden de feittypen gegroepeerd en gereduceerd. Ook deze stap werd uitgevoerd met CaseTalk, zie afbeelding 5. Dit diagram toont (in FCO-IM vermomming) de entiteittypen 'LineGroup_Forecast', 'Period', 'LineGroup_Budget', 'Month' and 'LineGroup'.

De feitexpressies kunnen nog steeds worden teruggegenereerd ter validatie. Met behulp van de module ERM Bridge uit de FCO-IM Bridge tool set kan een Entity-Relationship-format worden gegenereerd voor import in het ERM-tool ERwin.

De volgende stap is de transformatie van het genormaliseerde model naar een dimensioneel model. Dat kan met de module StarBridge van de FCO-IM Bridge tool set. Deze module helpt de ontwerper door voorstellen te doen voor kandidaat fact tables en dimension tables, door conformiteit te realiseren via het afsplitsen van een mini-tabel of een aggregaattabel aan te maken etcetera. De door de ontwerper gemaakte keuzen worden uitgevoerd en onthouden voor eventueel hergebruik in een volgende versie.
In dit voorbeeld suggereert het StarBridge-algoritme dat 'LineGroup_Forecast' een fact table wordt en 'LineGroup_Budget' en 'Period' dimensies, zie afbeelding 6a.

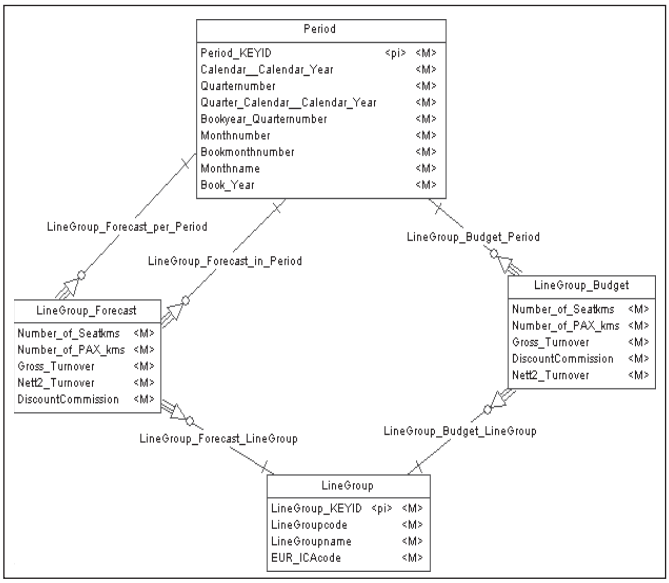
De feitexpressies zijn elementair en hebben de vereiste granulariteit

De ontwerper kan nu vrijelijk beslissen deze al dan niet te denormaliseren. In dit geval werd echter, rekening houdend met de zachte semantiek, besloten dat 'LineGroup_Budget' ook een fact table zou moeten worden. Door deze beslissing blijft 'LineGroup' een dimensie. 'Month' wordt gedenormaliseerd en in dimensietabel 'Period' opgenomen. Het boomdiagram ziet er dan uit als in afbeelding 6b. Het definitieve submodel, dat na toepassing van het StarSplit-algoritme ontstaat, heeft nu twee fact tables ('LineGroup_Budget' en 'LineGroup_Forecast') en twee conforme dimensies ('LineGroup' en 'Period'), waarbij vanuit 'LineGroup_Forecast' twee keer wordt verwezen naar 'Period'. Dit gekozen dimensionale deelmodel is nu (samen met de andere afgesplitste data marts) als een conceptueel dimensioneel deelmodel opgeslagen in de extended FCO-IM repository van het FCO-IM Bridge tool. Nog steeds kunnen ter validatie de initiële feitexpressies worden teruggegenereerd.



Afbeeldingen 6a en 6b: Boomdiagrammen van het submodel 'Forecast' voor en na denormalisatie.

De volgende stappen zijn de conversie naar een logisch en van daaruit naar een fysiek dimensioneel model. Voor het eerste zorgt de module ERM Bridge uit de FCO-IM Bridge tool set, dat een logisch Entity-Relationship-model genereert dat als XML-file kan worden geïmporteerd in het ERM-tool ERwin. De ontwerper kan nu voor het vervolg profiteren van de door ERwin geboden mogelijkheden voor de fysieke implementatie. Hierbij gaat ook de 'zachte semantiek' niet verloren. De module ERM Bridge brengt ook de volledige zachte semantiek (naamgeving en verwoordingen) over vanuit de FCO-IM repository naar een intermediate ERM repository en van daaruit naar het format van het gekozen ERM-tool ERwin in de vorm van namen voor entiteittypen en attributen, namen voor attributen en hun rollen en commentaar. Het dimensionele model in ER-notatie ziet er nu uit als in afbeelding 7: een data mart bestaande uit twee fact tables met respectievelijk drie en twee verwijzingen naar hun twee dimensietabellen.



Afbeelding 7: Entity-Relationship diagram van data mart 'Forecast'.

Literatuur

1. *Guido Bakema, Elton Manoku: Geïntegreerd tool support voor datawarehouse-ontwerp. DB/M, februari 2005.*
2. *Bommeljé Crompvoets and partners: CaseTalk en FCO-IM Bridge tool set. www.casetalk.com.*
3. *Guido Bakema: Metadata Management en Applicatiegeneratie, van visie naar toepassing. Tinfon, november 2004.*
4. *Peter Alons: Single point of definition voor metadata. DB/M, december 2000.*
5. *Peter Alons: Beter modelleren begint op conceptueel niveau. DB/M, februari 2001.*

Guido Bakema en Elton Manoku

Guido Bakema (guido.bakema@han.nl) is lector Data Architectures & Metadata Management aan de Hogeschool van Arnhem en Nijmegen en studieleider van masters-opleidingen in Information Systems Development. Elton Manoku (elton.manoku@han.nl) is als docent/onderzoeker werkzaam bij de Hogeschool van Arnhem en Nijmegen aan de ontwikkeling van geïntegreerde tool sets.