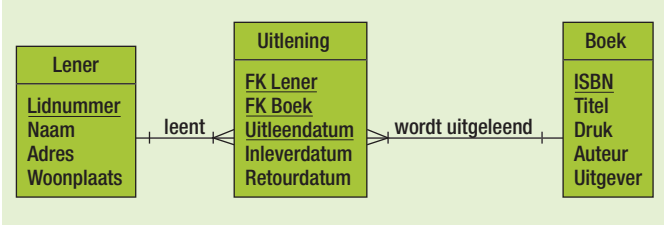
Waarom alles elke keer weer opnieuw doen?

Genereren van mappings

Alexander van Helm en Erik-Jan Koning

In dit artikel wordt beschreven dat het mogelijk is om een groot deel van het datawarehouse-proces te genereren. Met alle kwaliteits- en doorlooptijd winst van dien.

In de praktijk passen we dit succesvol toe bij klanten die werken met Business Objects Data Integrator en Oracle Warehouse Builder. Door het genereren is een besparing van meer dan 20 procent op de bouwtijd gerealiseerd. Bovendien zijn we zeker van de werking van gegenereerde mappings. De programmatuur om mappings daadwerkelijk te genereren is een implementatie van de hier beschreven algoritmes.
Om alle stappen te illustreren gebruiken we het model in afbeelding 1.



Afbeelding 1.

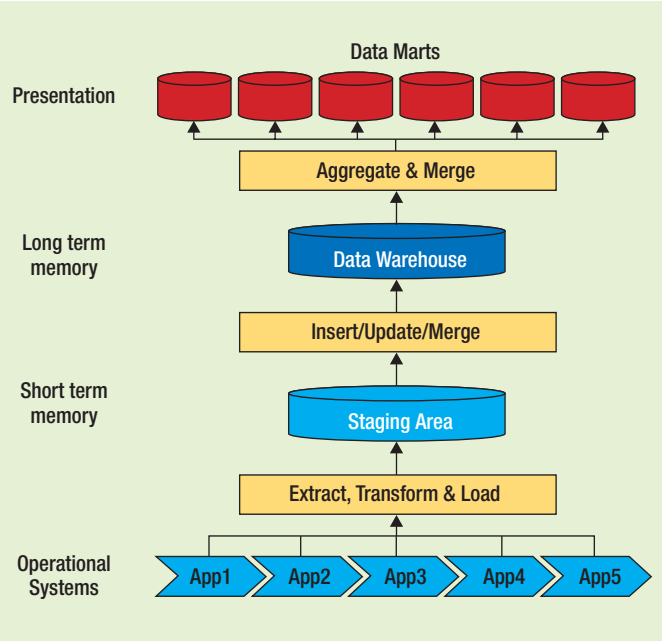
Een bibliotheek heeft een aantal boeken en een aantal leden. De leden lenen boeken op de uitleendatum, het boek wordt terug verwacht op de inleverdatum en is daadwerkelijk ingeleverd op de retourdatum. Natuurlijk zal in de praktijk het model een stuk complexer zijn (van titels bestaan meerdere exemplaren, auteur en uitgever zijn aparte dimensies, er zijn veel meer attributen, etcetera). Het mooie van genereren is dat het dan toch hetzelfde werkt.

Lagenmodel

Om te kunnen genereren is het een vereiste om volgens een architectuur te werken. In dit artikel wordt uitgegaan van de volgende lagenarchitectuur (zie afbeelding 2):

- Informatie laag. Uitgangspunt voor datawarehouses is het informatiegebruik. Dit wordt gefaciliteerd in de datamarts die primair gericht zijn op de presentatie aan specifieke gebruikersgroepen;
- Historische datalaag. Het centrale datawarehouse is het geheugen, het historisch archief van alle gegevens benodigd voor het vullen van datamarts;
- Staging-laag. De aanleveringen om het historisch archief bij te werken staan in de staging-laag.

We gebruiken een en hetzelfde model als grondslag voor zowel het datawarehouse als de staging-laag. Het is van groot belang dat elke entiteit een unieke sleutel heeft, de zogenaamde business key. De informatielaag modelleren we volgens het dimensionele model van Kimball (www.ralphkimball.com).
Omdat je weet welke entiteit in de staging-laag correspondeert met welke entiteit in de historische datalaag, kun je voor elk record vaststellen wat voor mutatie het betreft (insert, update, etcetera). Vanuit een basismodel kun je het datamodel voor zowel de staging-laag als het centrale datawarehouse genereren en de programma's die deze lagen vullen.
Voor het proces hebben we in elke laag specifieke procesattributen nodig. Dit zijn attributen die wat vertellen over de herkomst en de status van de data. Het nut van deze attributen is meervoudig: ondersteuning van het laadproces, assistentie bij testen en tracing, hulp bij onderhoud en het oplossen van fouten, inzicht in



Afbeelding 2: Lagen-architectuur.

Lener
Lidnummer
Naam
Adres
Woonplaats
DWH_key
Vor_key
Geldig_van
Geldig_tot
Huidig_record_ind
Verwijderd_ind
DM_verwerkt_ind
Run_opvoer
Run_afvoer
Bron
Aanlevermoment
Laatste_Mutatiecode

Uitlening
FK_Lener
FK_Boek
Uitleendatum
Inleverdatum
Retourdatum
DWH_key
Vor_key
Geldig_van
Geldig_tot
Huidig_record_ind
Verwijderd_ind
DM_verwerkt_ind
Run_opvoer
Run_afvoer
Bron
Aanlevermoment
Laatste_Mutatiecode

Boek
ISBN
Titel
Druk
Auteur
Uitgever
DWH_key
Vor_key
Geldig_van
Geldig_tot
Huidig_record_ind
Verwijderd_ind
DM_verwerkt_ind
Run_opvoer
Run_afvoer
Bron
Aanlevermoment
Laatste_Mutatiecode

Afbeelding 3.

mutatiegraden en oorzaken, inzicht in performance enzovoort. Per laag beschrijven we hier welke attributen extra nodig zijn.

Historische datalaag

De historische datalaag wordt uitgebreid met de volgende velden:

- DWH_Key. Technische identificatie per record, meestal gevuld vanuit een sequence;
- Vor_key. Welke rij is de voorganger van dit record. Gebruikt om ketens van wijzigingen te bekijken;
- Geldig_van. Vanaf wanneer is een rij de werkelijkheid;
- Geldig_tot. Tot wanneer is een rij de werkelijkheid;
- Huidig_record_indicator. Is een rij de huidige werkelijkheid?;
- Verwijderd_indicator. Is een rij verwijderd uit de administratie? Merk op, dat we de rij niet fysiek verwijderen, omdat het feit dat de rij heeft bestaan nuttige naslaginformatie is;
- DM_verwerkt_indicator. Is een rij al doorgezet naar de informatielaag? Benodigd om aan delta verwerking te doen;
- Run_opvoer. In welke technische run is deze rij ontstaan;
- Run_afvoer. In welke technische run is deze rij afgesloten;
- Bron. Uit welk systeem is de informatie aangeleverd;
- Aanlevermoment. Wanneer is de informatie aangeleverd;
- Laatste Mutatiecode. Wat is de laatste wijziging geweest aan dit record. (I: insert, een nieuw gegeven. D: delete, logische verwijdering van het gegeven. O: overwrite, wijziging die als correctie behandeld moet worden. Hier wordt geen historie van bijgehouden. U: update. Wijziging van gegevens die wel historisch wordt bijgehouden).

Uit ons voorbeeld hebben de tabellen in de historische datalaag de vorm als in afbeelding 3.

ISBN	Titel	Druk	Auteur	Uitgever	DWH_Key
0471200247	The Data Warehouse Toolkit	2	Ralph Kimball	Wiley	1
0201784203	Business Intelligence Roadmap	1	Larissa Moss	Addison-Wesley	2
0764599445	Building the Data Warehouse	4	Bill Inmon	Wiley	3

Afbeelding 4: Boek.

Lidnummer	Naam	Adres	Woonplaats	DWH_Key
10001	Henk van de Ploeg	Dorpsstraat 3	Viergraven	1
10002	Wilhelm Driessen	Kerkweg 17	Achterwijk	2

Afbeelding 5: Lener.

Lener	Boek	Uitleendatum	Inleverdatum	Retourdatum	DWH_Key
10001	0764599445	23-8-2006	13-9-2006	8-9-2006	1
10002	0201784203	15-9-2006	6-10-2006		2

Afbeelding 6: Uitlening.

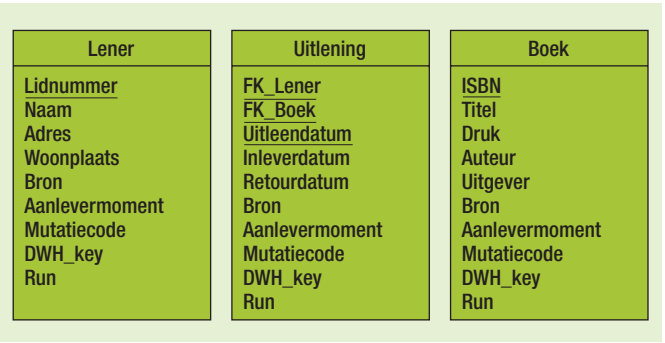
We gaan er vanuit dat de records al bekend waren in het datawarehouse, zie afbeelding 4, 5 en 6. Er zijn drie boeken, twee leners en twee uitleningen. Het eerste boek is alweer retour, terwijl het tweede boek nog uitgeleend is.

Staging-laag

De staging-laag wordt per entiteit uitgebreid met:

- Bron. Uit welk systeem is de informatie aangeleverd;
- Aanlevermoment. Wanneer is de informatie aangeleverd;
- Mutatiecode. Wat voor een wijziging betreft het;
- DWH_key. Technische sleutel van het huidige record in de historische datalaag met dezelfde business key.

In het voorbeeld hebben de tabellen de vorm in afbeelding 7. De nieuwe aanlevering bevat de gegevens in afbeelding 8 en 9.



Afbeelding 7.

In eerste instantie lijkt het aantal velden dat toegevoegd wordt af te schrikken. Hoe zorg je ervoor dat alle velden juist gevoed en opgeslagen worden? Door het proces te automatiseren is het aantal velden dat bijgehouden moet worden geen zorg meer.

Mapping van bron naar staging-laag

Op basis van een volledige aanlevering kan de delta-bepaling gegenereerd worden. De pseudo-code voor het genereren van een delta is als volgt:

1. Vergelijk de staging-laag met de huidige records van de historische datalaag. Met de huidige records van de historische datalaag worden die records bedoeld die de huidige status weergeven en die niet logisch verwijderd zijn;
2. Stel sets samen op basis van deze vergelijking:
   - Nieuw. Rij komt voor in aanlevering, niet in de huidige records van de historische datalaag en de unieke identificatie komt ook niet voor;
   - Wijziging. Rij komt voor in aanlevering, niet in de huidige records van de historische datalaag, maar de unieke identificatie komt al wel voor;
   - Verwijdering. Unieke identificatie komt wel voor in de huidige records van de historische datalaag, maar niet in de aanlevering.
3. Schrijf de sets weg in de staging-laag met de waarde zoals te zien in afbeelding 10.

Dit levert voor onze aanlevering de waarden getoond in afbeelding 11 en 12.

Lidnummer	Naam	Adres	Woonplaats
10003	Karel Leeflang	Slotlaan 90	Drift
10001	Henk van de Ploeg	Molenweg 12	Gaasperen

Afbeelding 8: Lener.

Lener	Boek	Uitleendatum	Inleverdatum	Retourdatum
10002	0201784203	15-9-2006	6-10-2006	6-10-2006
10001	0471200247	6-10-2006	27-10-2006	

Afbeelding 9: Uitlening.

Kolom	Nieuw	Wijziging	Verwijdering
Entiteit kolommen	Uit aanlevering overnemen	Uit aanlevering overnemen	Uit huidige records historische datalaag
Bron	Constante	Constante	Constante
Aanlevermoment	Constante	Constante	Constante
Mutatiecode	I	U	D
DWH Key	Leeg	Uit huidige records historische datalaag	Uit huidige records historische datalaag
Run	Constante*	Constante*	Constante*

* De constante is afhankelijk van het ETL-tool wel of niet te gebruiken.

Afbeelding 10.

Lidnummer	Naam	Adres	Woonplaats	Bron	Aanlevermoment	Mutatiecode	DWH_Key	Run
10003	Karel Leeflang	Slotlaan 90	Drift	App1	09-10-2006 08:00	I		2
10001	Henk van de Ploeg	Molenweg 12	Gaasperen	App1	09-10-2006 08:00	U	1	2

Afbeelding 11: Lener.

Lener	Boek	Uitleendatum	Inleverdatum	Retourdatum	Bron	Aanlevermoment	Mutatiecode	DWH_Key	Run
10002	0201784203	15-9-2006	6-10-2006	6-10-2006	App2	09-10-2006 08:00	U	2	2
10001	0471200247	6-10-2006	27-10-2006		App2	09-10-2006 08:00	I		2

Afbeelding 12: Uitlening.

Kolom	I	U1	U2	D1	D2
Entiteit kolommen	Staging laag	Ongewijzigd	Staging laag	Ongewijzigd	Staging laag
DWH_Key	Sequence	Ongewijzigd	Sequence	Ongewijzigd	Sequence
Vor_key	Leeg	Ongewijzigd	DWH Key	Ongewijzigd	DWH Key
Geldig_van	Aanlevermoment	Ongewijzigd	Aanlevermoment	Ongewijzigd	Aanlevermoment
Geldig_tot	Oneindig*	Aanlevermoment	Oneindig*	Aanlevermoment	Oneindig*
Huidig_record_ind	J	N	J	N	J
Verwijderd_record_ind	N	N	N	N	J
DM_verwerkt_ind	N	Ongewijzigd	N	Ongewijzigd	N
Run_opvoer	Run	Ongewijzigd	Run	Ongewijzigd	Run
Run_afvoer	Leeg	Run	Leeg	Run	Leeg

* Oneindig is een constante met een zo hoog mogelijke datum.

Afbeelding 13.

ISBN	Titel	Druk	Auteur	Uitgever	DK	VK	Geldig_van	Geldig_tot
0471200247	The Data Warehouse Toolkit	2	Ralph Kimball	Wiley	1		02-10-2006 08:00	31-12-9999 23:59
0201784203	Business Intelligence Roadmap	1	Larissa Moss	Addison-Wesley	2		02-10-2006 08:00	31-12-9999 23:59
0764599445	Building the Data Warehouse	4	Bill Inmon	Wiley	3		02-10-2006 08:00	31-12-9999 23:59

Afbeelding 14: Boek.

Lidnummer	Naam	Adres	Woonplaats	DK	VK	Geldig_van	Geldig_tot
10001	Henk van de Ploeg	Dorpsstraat 3	Viergraven	1		02-10-2006 08:00	09-10-2006 08:00
10002	Wilhelm Driessen	Kerkweg 17	Achterwijk	2		02-10-2006 08:00	31-12-9999 23:59
10003	Karel Leeflang	Slotlaan 90	Drift	3		09-10-2006 08:00	31-12-9999 23:59
10001	Henk van de Ploeg	Molenweg 12	Gaasperen	4	1	09-10-2006 08:00	31-12-9999 23:59

Afbeelding 15: Lener.

Lener	Boek	Uitleendatum	Inleverdatum	Retourdatum	DK	VK	Geldig_van	Geldig_tot
10001	0764599445	23-8-2006	13-9-2006	8-9-2006	1		02-10-2006 08:00	31-12-9999 23:59
10002	0201784203	15-9-2006	6-10-2006		2		02-10-2006 08:00	09-10-2006 08:00
10002	0201784203	15-9-2006	6-10-2006	6-10-2006	3	2	09-10-2006 08:00	31-12-9999 23:59
10001	0471200247	6-10-2006	27-10-2006		4		09-10-2006 08:00	31-12-9999 23:59

Afbeelding 16: Uitlening.

Mapping van staging-laag naar datawarehouse

In de staging-laag hebben we de tabellen voorbereid om verwerkt te worden in de historische datalaag. Op basis van de mutatiecode en de DWH-key is de verwerking nu als volgt:

1. Outer join de staging-laag met de huidige records uit de historische datalaag.
2. Verwerk de records op basis van de mutatiecode als volgt:
   - I: Voeg alle rijen toe (I);
   - U: Stapel de wijziging; Sluit het huidige record af op basis van de DWH-key (U1); Voeg een nieuw voorkomen toe (U2);
   - D: Stapel de verwijdering; Sluit het huidige record af op basis van de DWH-key (D1); Voeg een nieuw voorkomen toe (D2).

De vulling van de attributen is nu als in afbeelding 13. En dit levert uiteindelijk het resultaat op van boek, lener en uitlening, getoond in afbeelding 14, 15 en 16.

De praktijk

Er kan zowel voor ETL-tools als voor stored procedures gegenereerd worden. Door de mappings in de ETL-tools te genereren blijft het mogelijk om gebruik te maken van zowel de metadata-, de audit- en de performance-mogelijkheden van die ETL-tools.

In de praktijk hebben we kunnen bewijzen dat het genereren van mappings transparantie vergroot en het aantal fouten verkleint. Daarnaast is in de projecten een enorme verkorting in doorloop-tijd gerealiseerd. De conclusie luidt: genereren loont!.

Alexander van Helm en Erik-Jan Koning

Drs. A. van Helm (alexander.van.helm@kadenza.nl) en drs. E.J. Koning (erik.jan.koning@kadenza.nl) zijn beiden werkzaam als BI & DWH Architect bij Kadenza.

	HRI	VRI	DVI	ROr	RAr	Bron	Aanlevermoment	LMC
	J	N	J	1		Appl3	02-10-2006 08:00	I
	J	N	J	1		Appl3	02-10-2006 08:00	I
	J	N	J	1		Appl3	02-10-2006 08:00	I

	HRI	VRI	DVI	RO	RA	Bron	Aanlevermoment	LMC
	N	N	J	1	2	Appl1	02-10-2006 08:00	I
	J	N	J	1		Appl1	02-10-2006 08:00	I
	J	N	N	2		Appl1	09-10-2006 08:00	I
	J	N	N	2		Appl1	09-10-2006 08:00	U

	HRI	VRI	DVI	RO	RA	Bron	Aanlevermoment	LMC
	J	N	J	1		Appl2	02-10-2006 08:00	I
	N	N	J	1	2	Appl2	02-10-2006 08:00	I
	J	N	N	2		Appl2	09-10-2006 08:00	U
	J	N	N	2		Appl2	09-10-2006 08:00	I