Snel werken met SQL Server 2008, Data Vault en intelligente package templates

ETL in een paar minuten

Matthijs Vogt, Onno Walraven en Vincent Wylenzek

U heeft op basis van een 'Agile' software-ontwikkelings-methode inmiddels een aantal iteraties opgeleverd, wat uiteindelijk een volledige Business Intelligence solution moet worden. Door middel van voortschrijdend inzicht, het werken met innovatieve producten en nieuwe wensen vanuit de business ontstaat er een cyclisch proces en komt u tot de conclusie dat de reeds ontwikkelde logica prima functioneert, maar dat u het 'een volgende keer' technisch op een meer generieke manier zult opzetten.

De toegepaste kennis zou dan op een eenvoudige manier her-gebruikt en uitgerold kunnen worden bij meerdere klanten. Kwaliteitseigenschappen als beheer- en beheersbaarheid, schaalbaarheid, flexibiliteit, portabiliteit en implementeerbaarheid krijgen dan prioriteit. Herkent u zich in deze hierboven beschreven situatie?
Dit artikel gaat over een dergelijke situatie waar wij tegenaan zijn gelopen en die we vervolgens op een zo generiek mogelijke manier, met behulp van ETL-templates, volwassen hebben gemaakt. ETL-templates maken het mogelijk om op een zeer efficiënte en generieke manier de laadprocessen voor ons EDW op te zetten. Alle technische aspecten en fundamentele keuzes omtrent deze templates worden toegelicht. Aan de hand van een business case wordt duidelijk gemaakt dat een ETL-stap met behulp van een template in slechts enkele minuten op te zetten is.

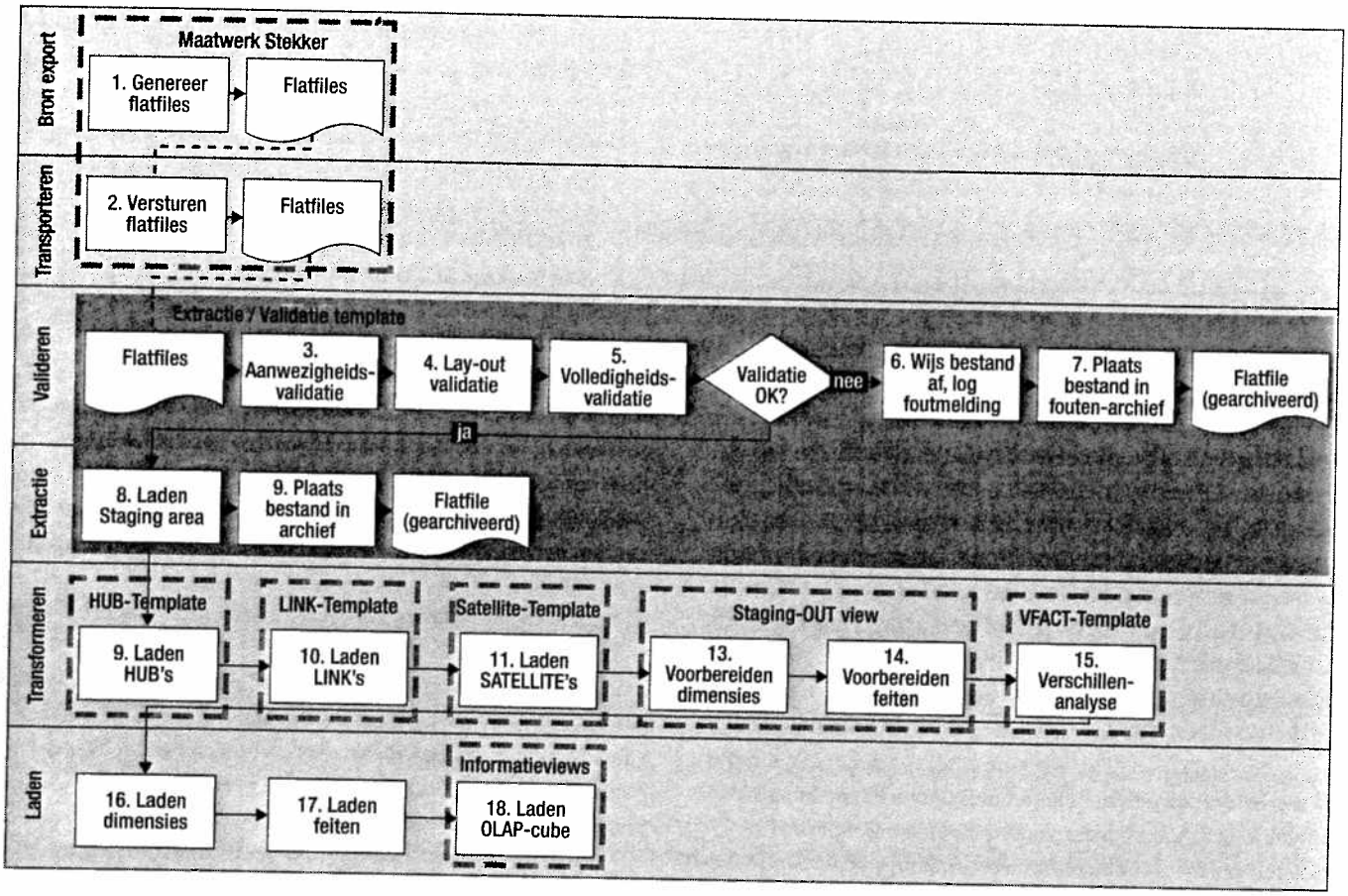
Van maatwerk naar solution

Ordina heeft de afgelopen jaren een groot aantal maatwerk Business Intelligence oplossingen ontwikkeld en succesvol geïmplementeerd bij klanten. Een deel van het maatwerk kon meerdere malen worden hergebruikt als het generiek was opgezet. Helaas was dit in beginsel nog niet het geval en is er een proces gestart om de herbruikbare onderdelen te vertalen van maatwerk naar een complete standaard oplossing.
Voor alle componenten van deze opossing is vervolgens bepaald of deze generiek op te zetten zijn. Met als resultaat dat we een aantal onderdelen, lees modulen, hebben ontwikkeld, welke generiek opgezet zijn en als standaard oplossing zijn in te zetten. De gehele technische opzet van staginglaag tot en met informatielaag is generiek opgezet en daarmee voor elke klant identiek. Het uitrollen van nieuwe releases wordt op deze manier zeer eenvoudig en aan de kwaliteitseis omtrent beheer- en beheersbaarheid wordt optimaal voldaan, omdat er een duidelijk overzicht is van de functionaliteit en versies die bij klanten draaien. Om een aantal veel voorkomende klantspecifieke wensen toe te kunnen passen is een deel van de bouwblokken met parameters in te stellen.
In lijn met de gedachte om het product van CMM niveau 1 (Ad hoc) naar CMM niveau 3 te tillen, hebben we het gehele ETL-proces ook conform een generiek model opgezet. In de volgende paragrafen wordt de opzet van dit generieke proces in de vorm van ETL-templates toegelicht. Eerst volgt een overzicht van de gemaakte keuzes en fundamenten die aan dit proces en dus impliciet aan deze templates ten grondslag liggen.

Waarom Data Vault?

Een generieke oplossing vraagt om een toekomstvaste architectuur; flexibel en historisch volledig. Voor het bouwen van transactionele applicaties zijn door de jaren heen standaarden ontwikkeld om aan deze vragen te voldoen. Hierdoor kunnen nieuwe versies bij klanten uitgeleverd worden zonder verlies van (historische) gegevens. Een datawarehouse is hiervoor een oplossing die zich al vele malen bewezen heeft en die we ook bij dit project zullen toepassen.

Datawarehousing kan worden toegepast met een veelvoud aan modelleringtechnieken. In het kader van volledigheid is Data Vault de beste keus. Omdat bij de gegevensverwerking naar Data Vault zo weinig mogelijk business rules toegepast worden (deze worden toegepast in de vervolgstappen naar de datamarts), blijft zo veel mogelijk van de gegevens bewaard. Er zijn namelijk bijna geen regels van toepassing waardoor gegevens worden afgekeurd. Een van de doelen van Data Vault is dan ook dat *alle* aangeboden gegevens worden opgeslagen. Zo wordt niet 'one version of the truth', maar 'one version of the facts' opgeslagen (zie ook het artikel van Ronald Damhof uit DB/M 5, 2008).
Door het vastleggen van de geldigheid van de attributen wordt het bijhouden van volledige historie gewaarborgd. Hierdoor kan zelfs gesproken worden van *'all* versions of the facts'. Het scheiden van de business keys en attributen door middel van hubs, links en satellites waarborgt de flexibiliteit en toekomstvastheid.



Afbeelding 1: ETL-proces. Standaard ETL-templates – Data Vault modellering.

Toegepaste nieuwe features

SQL Server 2008 heeft een aantal interessante nieuwe features (zie ook het artikel van Bram Dons in DB/M 6, 2008), waarvan wij er enkele gebruiken op het gebied van beveiliging, standaardisatie en compressie. Er wordt impliciet gebruik gemaakt van de performanceverbeteringen van SSIS (cached lookup transformations, verbeterde script tasks) en SSAS (improved query performance over star schema, MDX block computations [niet cell-by-cell berekeningen]).

Beveiliging

Gebruikers. Ten opzichte van SQL Server 2005, heeft 2008 een enorme vooruitgang op het gebied van beveiliging geboekt. Het is nu mogelijk om de lokale gebruikersgroep 'BUILTIN/Administrators' de toegang tot de SQL Server instance te onthouden. Als het beheer van onze standaard software door de klant (lokale administrator) wordt uitgevoerd, is dit een belangrijke eis.

Encryptie. Er wordt tevens gebruik gemaakt van encryptie van de SQL Server databases. Encryptie van bijvoorbeeld cliëntgegevens is wenselijk gezien het vertrouwelijke karakter. Met het nieuwe *Transparant data encryption* (TDE) wordt de encryptie niet meer binnen de database gedaan, maar op bestandsniveau. De database files worden door middel van een T-SQL statement versleuteld. Zie het artikel 'Database Encryption in SQL Server 2008 Enterprise Edition' van Sung Hsueh/Microsoft Technet voor details over de technische inrichting.

Met TDE kan er in tegenstelling tot CBE gebruik gemaakt worden van indexen en keys, is er geen toename van storage en een minimale hoeveelheid performanceverlies. De CPU wordt wel extra belast, gemiddeld neemt de CPU utilization met 3 tot 5 procent toe (bron: Microsoft).

Standaardisatie

Policy-based management maakt het mogelijk om naamgevingconventies te hanteren. Met policy-based management kan deze naamgevingconventie als policy ingeregeld worden, hetgeen betekent dat als er een DDL-statement wordt uitgevoerd om een object (tabel, view, procedure et cetera) te creëren, er direct wordt gecontroleerd of het object aan de policy voldoet.

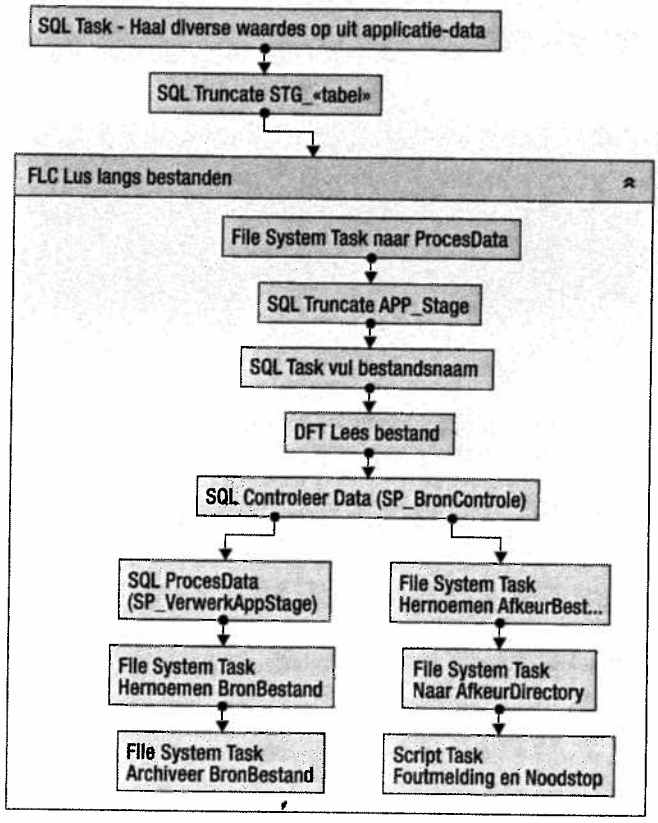
Compressie

In SQL 2008 kan er gebruik gemaakt worden van Page-level encryption (PLE). PLE maakt één bibliotheek per datapage aan, waarin alle begrippen, die meerdere malen voorkomen binnen de page, vastgelegd zijn. De attributen binnen de page worden vervangen door pointers, die verwijzen naar deze bibliotheek.

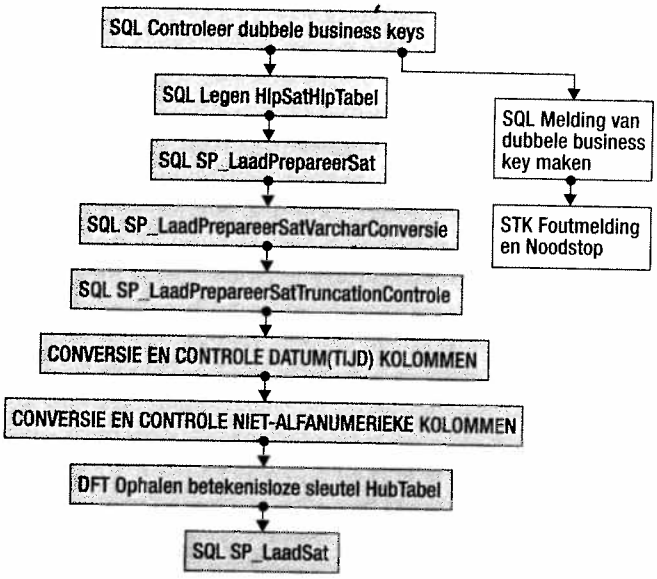
Vooral bij herhaaldelijk voorkomende attributen, wat bij een datawarehouse veelvuldig het geval is, levert dit een ruimte-besparing op.
Het performanceverlies is beperkt bij PLE en aangezien wij niet direct vanuit de front-end verbinding maken met de datamart, maar alles via de OLAP-cubes loopt, ervaart de eindgebruiker geen performanceverlies.

SSIS Template packages

In afbeelding 1 is het ETL-proces te zien. De met een stippellijn gemarkeerde vakken bevatten de naam van een reeds gerealiseerde template-package of template-view.
Extractie. Omdat door ons gekozen is voor een generieke implementatie van de ETL-processen, is ontkoppeling van de generieke ETL en de bronsystemen noodzakelijk. Deze ontkoppeling wordt gerealiseerd door middel van aanlevering in column-delimited flat files. De bronbestanden worden geëxtraheerd naar stagingtabellen die qua structuur overeenkomen met de flatfiles. Op de flatfiles worden aanwezigheids-, lay-out- en volledigheidscontroles uitgevoerd. Fouten worden gelogd en het bestand wordt gearchiveerd.
Omdat de controles die toegepast worden voor ieder bestand gelijk zijn en de datatypes van ieder veld in iedere tabel gelijk zijn, leent dit zich uitstekend voor een ETL-template. Deze template werkt op basis van een aantal parameters en het invullen van een aantal variabelen. De ontwikkelaar die de template toepast hoeft hiermee nog maar een minimale inspanning te leveren.
Laden van het EDW – Hub. Vanuit de stagingtabellen worden de verschillende hubs in het EDW gevuld. Omdat de hubs alle aangeleverde business keys moeten bevatten worden er meerdere tabellen aan de hub aangeboden. Hieronder valt natuurlijk de aanlevering van de referentiegegevens van de business key, bijvoorbeeld een lijst met kostenplaatsen uit een financieel systeem. Hieronder vallen ook de business keys uit de transactionele gegevens, bijvoorbeeld de kostenplaats uit een financiële transactie. Hierin zit een volgordelijke afhankelijkheid – de referentiegegevens hebben voorrang boven de transactiegegevens. Omdat dit proces ook generiek op te zetten is, hebben we besloten om een ETL template voor deze stap te introduceren. De template kan worden toegepast door het invullen van een aantal variabelen.
Laden van het EDW – Link. De verschillende linktabellen worden gevuld op basis van één staging tabel. Deze stagingtabel bevat alle relaties of business events tussen de verschillende business keys. Ook dit proces is generiek op te zetten. Daarom hebben we ook hier besloten om een ETL-template voor deze stap te introduceren. Ook voor deze template geldt dat bij het implementeren alleen een aantal variabelen hoeft te worden ingevuld.
Laden van het EDW – Satellite. Bij het laden van de satellitetabellen worden de attributen van een business key historisch vastgelegd. Deze attributen worden gekoppeld aan de betekenisloze sleutel die in de bijbehorende hub of link aan de business key is toegekend. De attributen zelf worden in deze stap geconverteerd naar het juiste datatype.
Omdat het opzoeken van de surrogate key in de hub of link een generiek proces is, is voor deze stap een ETL template aangemaakt. Deze ETL template kan worden toegepast door een aantal variabelen in te vullen. Dataconversies van numerieke waarden moeten als maatwerk worden geïmplementeerd.
Staging out. De logica en daarmee de onderliggende Data Vault datastructuur wordt afgeschermd door middel van views op het EDW. Hierin worden de hubs en links met de bijbehorende satellites weer aan elkaar gekoppeld, zodat een consistente gegevensset zichtbaar wordt. Voor het reguliere ETL-proces zijn staging-out views ontwikkeld, welke de meest recente geldende situatie aan de vervolgstappen van het ETL-proces aanbieden. Daarnaast zijn er ook staging-out views ontwikkeld om het historisch verloop van de business keys en bijbehorende attributen weer te kunnen geven.
Voorbereiden van de dimensies. In deze stap wordt de transformatie naar stermodel uitgevoerd, op basis van de staging-out views. Daarnaast wordt de dimensie bijgewerkt tot de nieuwe situatie.
Voorbereiden van de feitentabellen. In deze stap wordt de transformatie naar stermodel uitgevoerd, op basis van de staging-out views. Daarnaast wordt de feitentabel opgebouwd en bijgewerkt tot de nieuwe situatie.



Afbeelding 2: Extractie template-package.



Afbeelding 3: Sattelite template-package.

Verschillenanalyse. Vergelijking van de opgebouwde nieuwe situatie, met de situatie uit de datamart voor het bepalen van de deltawaarden.
Laden van de dimensies. Het bijwerken van de dimensietabellen, vanuit de prepare-laag naar de datamart-laag.
Laden van de feitentabellen. Het toevoegen van de deltarecords uit de prepare-laag aan de feitentabel in de datamart. Delta-records zijn alle records, waarin één van de attributen een wijziging bevat ten opzichte van de vorige aanlevering.

Implementatie ETL-Template

"Met goed bruikbare templates wordt het ontwikkelwerk significant verkort. Voor een aantal onderdelen heeft dit de ontwikkeltijd met 75 procent verkort." Een quote van Pim Dettmers (Data Management Engineer).
Bij het inrichten van de Hub satellite package (zie afbeelding 3) zijn er randvoorwaarden en daarom dient de gebruiker een aantal variabelen in te vullen:

- *Entiteit.* De business entiteit die gebruikt wordt om tabelnamen af te leiden. Op basis van deze variabele wordt een aantal andere variabelen automatisch van waarde voorzien;
- *Satellite kolommen*; de business attributen.

Een Hub satellite package bevat acht stappen.
1. Controleer dubbele business keys.
In de staging area tabel wordt gecontroleerd of business keys meerdere malen voorkomen. Zo ja, dan wordt de verwerking afgebroken en wordt een foutmelding gegenereerd.
Stapkenmerken: 100 procent generiek opgezet; geen stored procedure.

Product_code	Product_ID	Omschrijving	Omschrijving_doel	...	...

Voorbeeld hulptabel, stap 2.

Product_code	Product_ID	Omschrijving	Omschrijving_doel	...	...
A		Dit is een omschrijving			
B		Omschrijving			

Voorbeeld gevulde hulptabel, stap 3.

Product_code	Product_ID	Omschrijving	Omschrijving_doel	...	...
A		Dit is een omschrijving →	Dit is een omschrijving		
B		Omschrijving →	Omschrijving		

Voorbeeld conversie alfanumerieke kolommen, stap 4.

Product_code	Product_ID	...	...	Aantal	Aantal_doel
A				40,00 →	40,00
B				100,0023 →	

Voorbeeld conversie niet-alfanumerieke kolommen, stap 6.

Product_code	Product_ID	...	...	...	...
A →	123				
B →	423				

Voorbeeld opzoeken betekenisloze sleutel, stap 7.

2. Legen satellite hulptabel.
Op basis van de waarde uit een variabele wordt de betreffende hulptabel geleegd.
Zie voorbeeld hulptabel, stap 2.
Stapkenmerken: 100 procent generiek opgezet; geen stored procedure.

3. Overhalen van data uit de betreffende staging area tabel naar de satellite hulptabel.
Op basis van de waarden uit de variabelen wordt de hulptabel gevuld met alle data uit de betreffende staging area tabel.
Zie voorbeeld gevulde hulptabel, stap 3.
Stapkenmerken: 100 procent generiek opgezet; stored procedure aangezien het aantal kolommen dat verwerkt wordt per satellite verschilt; aangeboden satellites worden geteld en de telling wordt gelogd.

4. Conversie van datatypes van alfanumerieke business attributen op basis van datatypes in het EDW.
In de systeemtabellen wordt gecontroleerd welke business attributen in het EDW een alfanumeriek datatype hebben. Deze attributen worden in de hulptabel geconverteerd naar het nieuwe datatype (bepaling via datadictionary). In de hulptabel is voor elk business attribuut een extra kolom opgenomen met het datatype dat overeenkomt met het EDW. De geconverteerde waardes worden in deze doelkolommen in de hulptabel geplaatst.
Zie voorbeeldtabel conversie alfanumerieke kolommen, stap 4.
Stapkenmerken: 100 procent generiek opgezet; stored procedure aangezien het aantal kolommen dat verwerkt wordt per satellite verschilt;

5. Controle afkappen van alfanumerieke business attributen.
In de systeemtabellen wordt gecontroleerd welke business attributen in het EDW een alfanumeriek datatype hebben. Van deze attributen worden de lengtes van de waardes in de beide kolommen (bijvoorbeeld omschrijving en omschrijving_doel) in de hulptabel vergeleken. Records waarbij de lengtes van attributen tussen kolommen niet overeenkomen worden gekopieerd naar een afkeurtabel en worden gekoppeld aan een melding in de loggingtabel.
Stapkenmerken: 100 procent generiek opgezet; stored procedure aangezien het aantal kolommen dat verwerkt wordt per satellite verschilt.

6. Conversie van datatypes van niet-alfanumerieke business attributen op basis van datatypes in het EDW.
De niet-alfanumerieke attributen worden in de satellite hulptabel geconverteerd naar het datatype dat overeenkomt met het EDW. In de hulptabel is voor elk business attribuut een extra kolom opgenomen met het datatype dat overeenkomt met het EDW. De geconverteerde waardes worden in deze kolommen in de hulptabel geplaatst. Van attributen waarbij de conversie niet uitgevoerd kan worden, blijft de doelkolom leeg. Records die niet (volledig) geconverteerd kunnen worden, worden gekopieerd naar de afkeurtabel en gekoppeld aan een melding in de loggingtabel.
Zie voorbeeld conversie niet-alfanumerieke kolommen, stap 6.

Het vervolg van dit artikel staat op pagina 33.

mando's te beheren, bijvoorbeeld met behulp van Remote Server Administration Tools (RSAT) die op een Windows Vista SP1 draaien. De installatie van Failover Clustering wordt geactiveerd via de 'Add Features' wizard en in de Hyper-V Manager met een Virtual Network. Na de gebruikelijke creatie van een VM moet deze via Failover Cluster Management nog 'highly available' worden gemaakt.
Microsoft heeft eind vorig jaar de bètaversie van Windows 2008 R2 vrijgegeven. Een van de belangrijkste nieuwigheden van R2 is de 'Live Migration', de real-time verhuizing van VM's; zie voor een evaluatie van R2 het artikel 'Realtime VM's migreren met Cluster Shared Volumes' in Storage Magazine 2, 2009.

Conclusie

Met de komst van Windows Server 2008 Hyper-V heeft Microsoft een belangrijke stap gezet in de markt van 'Type 1' hypervisor technologie; daarbij draait de hypervisor direct op de onderliggende hardware op een server, in tegenstelling tot de 'Type 2' waarbij deze binnen een operating system draait (Microsoft Virtual Server, Virtual PC en VMware Server). De grootste concurrenten van Hyper-V zijn de Citrix XenServer en VMware ESX Server. Van deze twee producten is VMware toch wel het verst ontwikkeld en met de onlangs aangekondigde nieuwe release van ESX Server, VMware vSphere 4 genaamd, is de afstand met Microsoft nog weer eens behoorlijk vergroot. Want zo boden onder meer ESX Server 3.5 en Citrix 5.0 al geruime tijd de real-time verplaatsing van VM's, respectievelijk 'VMotion' en 'Xenmotion' genaamd en Microsoft Storage VMotion. Eind dit jaar hoopt Microsoft pas deze feature officieel in Windows Server 2008 te kunnen bieden. Voordeel van Hyper-V is wel dat de virtualisatie zonder extra kosten met Windows Server 2008 wordt meegeleverd. Bovendien kunnen naast de VM's nog, zij het in beperkte mate, andere applicaties op de Windows server worden gedraaid. De mate waarin dit mogelijk is, hangt natuurlijk af van restcapaciteit van de server. In onze test hebben we gezien dat de meest voorkomende databases in een op Linux of Windows gebaseerde VM zonder problemen kunnen worden toegepast. Hyper-V is een eenvoudig te configureren en toe te passen virtualisatie-omgeving die in combinatie met de nieuwe Windows VMM 2008 een goed te beheren virtuele omgeving biedt. Hyper-V is voor high availability te combineren met Microsoft Clustering Services. Naast deze op beperkte schaal toepasbare high availability voorziening biedt Hyper-V nog slechts een beperkte load balancing feature, in vergelijking met de nieuw aangekondigde Citrix Xenserver versie 5.5, de bestaande VMware ESX 3.x en de zojuist vrijgegeven VMware vSphere 4.

Noot

1. Deze term refereert aan modificaties aan het OS om de hypervisor er op te wijzen dat het efficiënter kan draaien wanneer het als een guest wordt gedetecteerd binnen een hypervisor-omgeving.

Bram Dons is onafhankelijk IT consultant.

Vervolg van ETL in een paar minuten.

Stapkenmerken: niet generiek opgezet, in verband met de performanceproblemen die dit met zich mee breng; de ontwikkelaar dient deze stap toe te voegen, wanneer er sprake is van aanwezigheid van niet-alfanumerieke business attributenkolommen.

7. Opzoeken van de betekenisloze sleutel van de bij de satellite behorende Hub.
Voor de overgebleven records in de hulptabel wordt de betekenisloze sleutel van de bij de satellite behorende Hub opgezocht. Zie voorbeeld opzoeken betekenisloze sleutel, stap 7, pagina 17.
Stapkenmerken: 100 procent generiek opgezet; geen stored procedure.

8. Verwerken van de records in de satellite hulptabel in het EDW.
De records uit de hulptabel worden verwerkt in het EDW.
Satellites (in het EDW) waarvan een gewijzigde versie aanwezig is in de hulptabel worden afgesloten. Nieuwe satellites of nieuwe versies van satellites worden toegevoegd aan het EDW.
Stapkenmerken: 100 procent generiek opgezet; stored procedure aangezien het aantal kolommen dat verwerkt wordt per satellite verschilt; nieuwe satellites worden geteld en de telling wordt gelogd; gewijzigde satellites worden geteld en de telling wordt gelogd.

Conclusie

Door gebruik van de template packages wordt de doorlooptijd van ETL-ontwikkeling verkort en is er eenduidigheid in het ETL-proces. De kwaliteitseisen worden gewaarborgd, doordat alle packages voldoen aan de naamgevingconventie en standaard-procedures hierin zijn uitgewerkt. Bepaalde generieke logica is opgenomen in stored procedures. Hierdoor kan deze logica op één centrale plek beheerd worden. Wijzigingen zijn direct merkbaar in alle geïmplementeerde ETL-packages.

Matthijs Vogt, Onno Walraven en **Vincent Wylenzek** zijn allen werkzaam bij Ordina als respectievelijk Data Management Engineer, Senior Business Intelligence Consultant en Business Intelligence Consultant.