Toepassingsgebieden van datavirtualisatie in een datawarehouse-omgeving

Virtuele oplossingen zijn flexibel

Rick van der Lans

De term datavirtualisatie komen we steeds vaker tegen. In een notendop komt het erop neer dat met datavirtualisatietechnologie gegevens uit verschillende databases geïntegreerd kunnen worden en dat applicaties van databases ontkoppeld worden; daarmee worden ze tevens afgeschermd van de fysieke eigenschappen van diezelfde databasegegevens.
In dit artikel beschrijven we de toepassingsgebieden van datavirtualisatie in een datawarehouse-omgeving.

Globaal leidt het toepassen van deze technologie tot een vereenvoudiging van een datawarehouse-architectuur en daarmee tot een verhoging van haar flexibiliteit en tot een verlaging van de totale kosten. *Datavirtualisatie* is misschien een relatief nieuw begrip, maar *virtualisatie* is dat zeker niet. De eerste toepassing van virtualisatie in de IT-industrie was waarschijnlijk geheugenvirtualisatie. Met deze techniek wordt er meer geheugen gesimuleerd dan er werkelijk in een machine beschikbaar is. Later zijn andere vormen geïntroduceerd, waaronder netwerk- en storagevirtualisatie. Op een bepaalde manier is ook cloudtechnologie een vorm van virtualisatie. Datavirtualisatie is als het ware een nieuwe vorm hiervan.

Virtualisatie in het algemeen houdt in dat we applicaties toegang tot en gebruik van een resource, zoals geheugen, storage en gegevens, geven zonder dat ze zich moeten bekommeren om onder andere de locatie van de resource, hoeveel er van de resource beschikbaar is en wat de API is. Hetzelfde geldt dus ook voor datavirtualisatie, waarbij de gegevens dan de resource vormen. Met datavirtualisatie zien de applicaties één grote geïntegreerde hoeveelheid gegevens en zien zij niet in welke database welk gegevenselement opgeslagen is, in welk formaat, met welke taal de gegevens benaderd moeten worden, dat ze verspreid zijn over verschillende database servers, enzovoort. De applicaties ervaren het alsof ze één grote geïntegreerde database benaderen, zie afbeelding 1.

Centraal in dit plaatje staat een laag software die verantwoordelijk is voor de datavirtualisatie. Deze laag ontkoppelt *dataconsumenten* van *datastores*. Mogelijke voorbeelden van dataconsumenten zijn productieapplicaties, rapporten en services in een SOA. Voorbeelden van datastores zijn operationele databases, datawarehouses, e-mailbestanden, spreadsheets, XML-documenten, sequentiële bestanden, externe gegevensbronnen en services die gegevens opleveren.

Zoals vermeld, zien de dataconsumenten één grote, geïntegreerde database. Zij kunnen bijvoorbeeld een patiëntendossier opvragen terwijl de benodigde gegevens om dat dossier op te bouwen over verschillende datastores verspreid zijn. Het is de taak van de datavirtualisatielaag om die gegevens uit de relevante datastores op te halen en te integreren. Een ander voorbeeld is dat een rapport courante productiecijfers samen met historische productiecijfers wil analyseren. Ook in dit geval is het de taak van de datavirtualisatielaag om de gewenste gegevens bij elkaar te brengen en geïntegreerd te presenteren alsof alle gegevens uit één database komen.

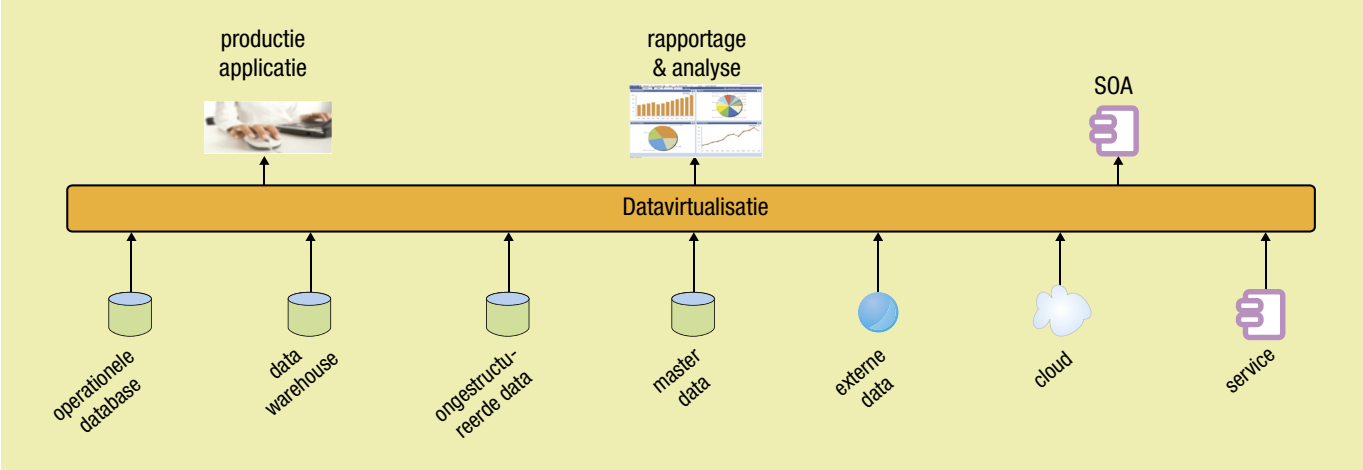
Kortom, de datavirtualisatielaag is verantwoordelijk voor het verwerken van query's. Deze verwerking bestaat minimaal uit de volgende stappen:

- de query moet opgebroken worden in deelquery's die elk naar een andere gegevensbron gestuurd worden;
- deze deelquery's moeten mogelijkerwijs vertaald worden naar een andere taal (de taal die door de gegevensbron ondersteund wordt);
- de resultaten van de deelquery's moeten tot één resultaat geïntegreerd worden;
- het eindresultaat moet naar de applicatie teruggestuurd worden.

Voor het implementeren van een datavirtualisatielaag bestaan diverse producten, zie tabel 1. Lange tijd werden dit *datafederatieservers* of kortweg *federatieservers* genoemd. Tegenwoordig wordt steeds vaker de term *datavirtualisateserver* gebruikt. Er bestaat wel een subtiel verschil tussen deze twee termen, maar dat zullen we in dit artikel buiten beschouwing laten en ze voor het gemak als synoniemen beschouwen.[1]

Naar on-demand transformaties

In een datawarehouse-omgeving is het gebruikelijk om gegevens te transformeren wanneer ze van de ene naar de andere data-



Afbeelding 1: Met datavirtualisatie wordt een verzameling heterogene gegevensbronnen als één geïntegreerde database gepresenteerd.

base gekopieerd worden. Bijvoorbeeld, als gegevens van een productiedatabase naar een datawarehouse gekopieerd worden, moeten er vaak codes omgezet worden, waarden gecombineerd en geknipt en ontbrekende gegevens uit andere systemen gehaald worden. Ook zullen incorrecte waarden opgeschoond moeten worden. We refereren in dit artikel naar dergelijke operaties met de term *transformaties*.

Veel organisaties gebruiken voor dit soort transformaties ETL-producten, zoals Informatica PowerCenter, Oracle Warehouse Builder en Pentaho Data Integration (Kettle). In principe voeren ETL-producten deze transformaties periodiek uit, ofwel ze worden gescheduled. Bijvoorbeeld, elke zondagmiddag om 14.00 uur worden in een keer nieuwe gegevens van het datawarehouse naar een datamart gekopieerd. We noemen dit *scheduled transformaties*; zie afbeelding 2.

Als gegevens door een federatieserver van een datastore naar een dataconsument verzonden worden, zullen deze gegevens ook getransformeerd moeten worden. Deze transformaties worden echter niet gescheduled, maar worden live uitgevoerd. Dus op het moment dat de dataconsument gegevens opvraagt, worden ze uit de datastores gehaald, getransformeerd en naar de consument verstuurd. We noemen dit *on-demand transformaties*. In afbeelding 3 wordt dit geïllustreerd. De resultaten van on-demand transformaties worden dus niet opgeslagen voordat ze beschikbaar komen voor de dataconsumenten, maar worden direct aan hen doorgegeven.

Beide vormen van transformaties hebben hun voor- en nadelen. De twee grote voordelen van on-demand transformaties zijn dat de dataconsumenten met 'versere' gegevens kunnen werken en dat er minder behoefte is om afgeleide datastores (zoals datamarts en kubussen) op te bouwen, iets wat de architectuur flexibeler en goedkoper zal maken. Een nadeel van deze vorm van transformaties is dat de transformaties elke keer wanneer gegevens opgevraagd worden opnieuw uitgevoerd worden en dus tijd en computer resources kosten. Sommige transformaties zouden zelfs zo complex en tijdrovend kunnen zijn dat het onwerkbaar wordt. Met andere woorden, beide vormen zijn noodzakelijk en nuttig.

Voor meer informatie over datavirtualisatie verwijzen we naar www.b-eye-network.com.

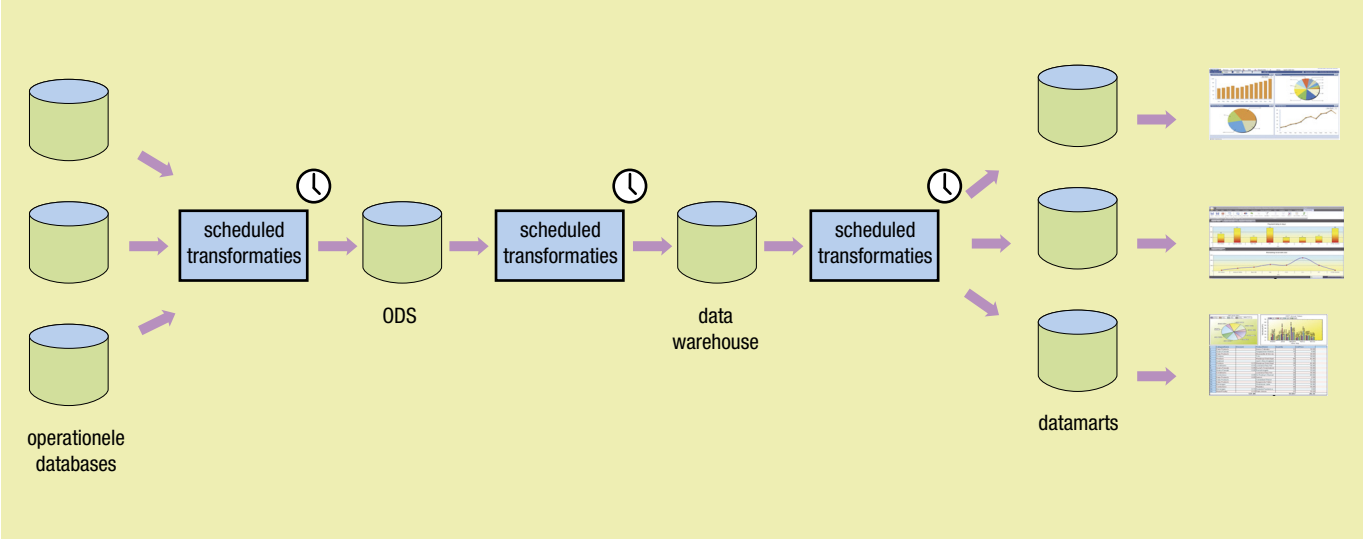
De virtuele tabellen van een federatieserver

De meeste federatieservers hebben een vergelijkbare interne werkwijze. Om te zorgen dat gegevens via een federatieserver beschikbaar komen, moeten eerst tabellen gedefinieerd worden. Deze tabellen worden echter niet met gegevens gevuld zoals in een 'gewone' database, maar worden via specificaties aan werkelijke tabellen en bestanden gekoppeld. Ze worden daarom ook meestal *virtuele tabellen* genoemd. Een groot deel van de specificaties voor een virtuele tabel heeft betrekking op hoe de gegevens uit de verschillende bronnen geïntegreerd moeten worden en naar de structuur van die virtuele tabel getransformeerd moeten worden.

Voor het definiëren van de transformaties om de virtuele inhoud van een tabel op te bouwen, gebruiken sommige producten SQL, andere XQuery, en er zijn er die flow-achtige talen gebruiken, zoals we ze kennen van ETL-producten. Het verschil is dat bij een ETL-product het resultaat van een flow leidt tot het opslaan

Datavirtualisatieservers
Composite Information Server
Dataflux Federation Server
Denodo Platform
Entropysoft Content Federation Server
IBM InfoSphere Federation Server
Informatica Data Services
Ipedo XIP
iWay Data Hub
Oracle BI Server
Queplix Virtual Data Manager
Software AG Information Integrator

Tabel 1: Overzicht van enkele federatieservers alias datavirtualisatieservers.



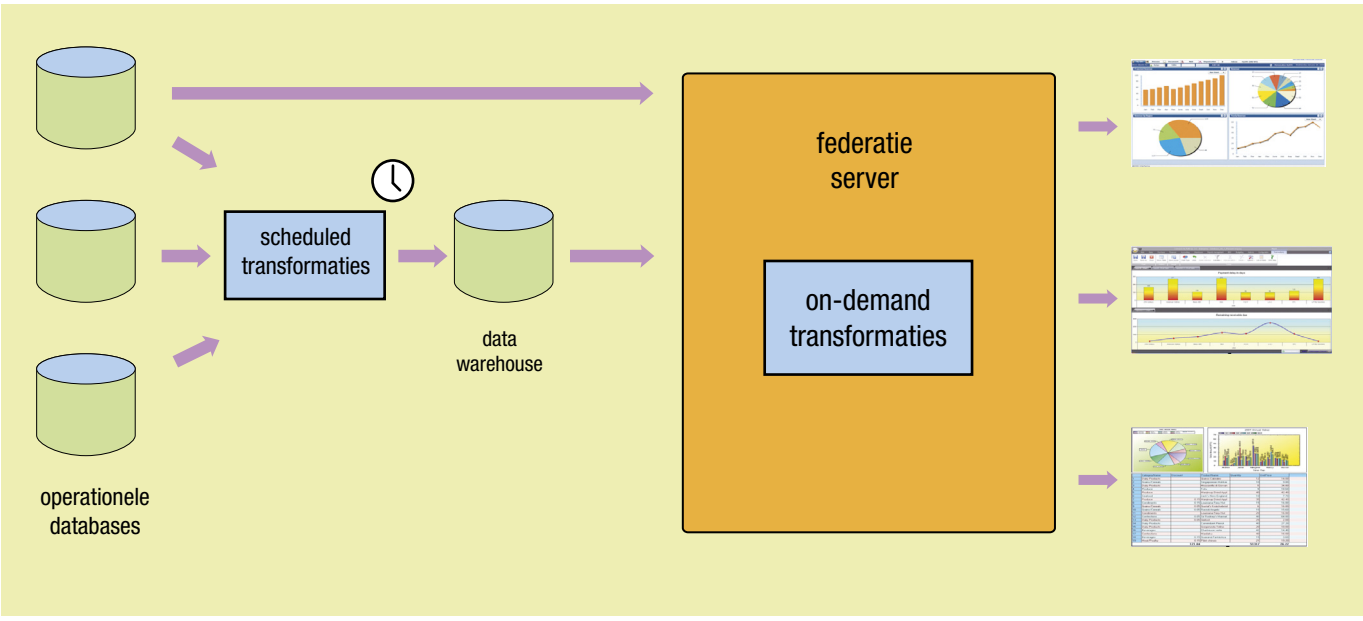
Afbeelding 2: Scheduled transformaties waarmee periodiek datastores bijgewerkt worden.

van het resultaat, terwijl bij een federatieserver het resultaat naar de applicaties wordt doorgestuurd.

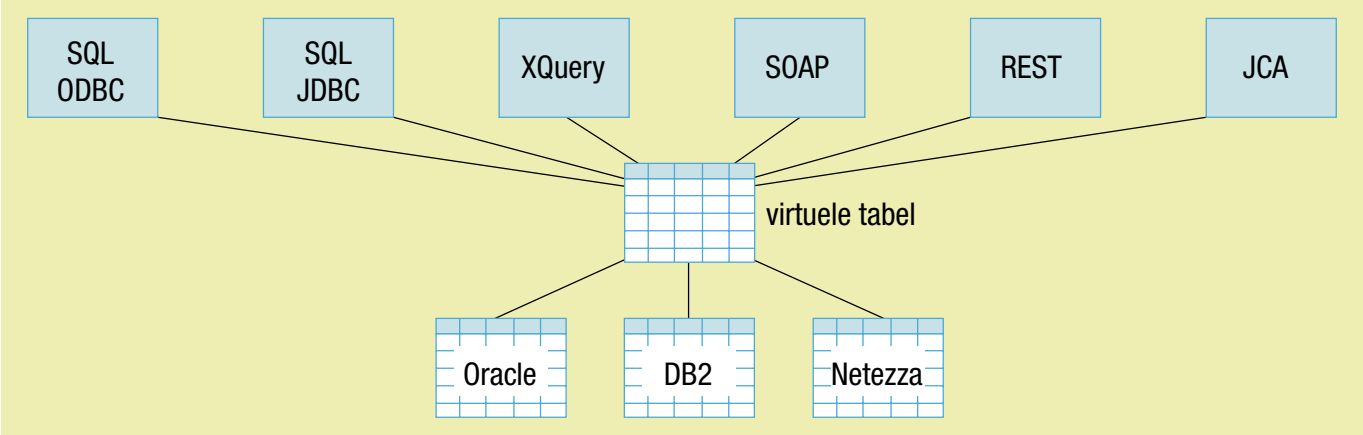
Als een virtuele tabel gedefinieerd is, kan deze via verschillende technische interfaces benaderd worden. Meestal worden hiervoor interfaces als SQL/ODBC, SQL/JDBC, XQuery, SOAP, REST, JMA en JMS ondersteund. Afbeelding 4 is een weergave van de relatie tussen een virtuele tabel en de gegevensbronnen enerzijds en de technische interfaces anderzijds.
Om modulair te kunnen werken, kunnen virtuele tabellen gestapeld worden. De ene virtuele tabel kan dan als gegevensbron voor een andere virtuele tabel fungeren.
Voor meer informatie over de interne werkingswijze van federatieservers verwijzen we naar de whitepaper *Developing a Data Delivery Platform*.[2]

Open versus closed federatieservers

Er bestaan diverse rapportage- en analyseproducten die een vorm van federatie ondersteunen. Bijvoorbeeld, het self-service BI-product QlikView is zeker in staat om gegevens uit een verzameling heterogene datastores te benaderen en te integreren. Hetzelfde geldt voor SAP/BusinessObjects, IBM/Cognos en vele andere producten. Bijvoorbeeld, het Universe concept in BusinessObjects kunnen we als federatietechnologie beschouwen. Echter, alle specificaties die we in dit soort producten opvoeren zijn alleen bruikbaar voor deze producten zelf (of alleen voor de producten van de leverancier), zie afbeelding 5. Ofwel, de transformatiespecificaties die in deze producten opgevoerd zijn, zijn niet *shareable*. Als een organisatie verschillende producten gebruikt, zullen de transformatiespecificaties in al deze producten herhaald moeten worden. Dat is



Afbeelding 3: Bij on-demand transformaties worden gegevens direct naar de dataconsument doorgestuurd.



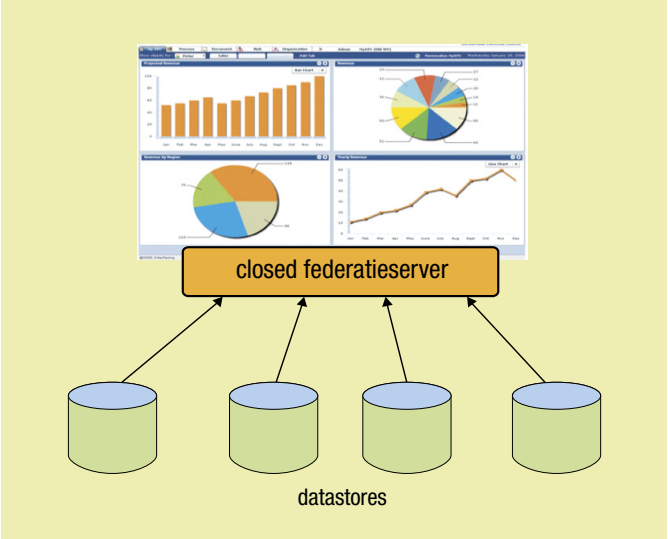
Afbeelding 4: De relatie tussen een virtuele tabel en de interfaces enerzijds en de gegevensbronnen anderzijds.

de reden waarom we dit *closed federatieservers* noemen.
Voor een *open federatieserver* geldt het tegenovergestelde. Buiten dat deze veel soorten gegevensbronnen kan benaderen, kunnen de gecreëerde virtuele tabellen door allerlei soorten producten benaderd worden, zie afbeelding 6. Het effect is dat hier de transformatiespecificaties wel shareable zijn. Veronderstel dat we in een virtuele tabel de definitie opnemen dat verkoopregio Noord-Drenthe niet bevat, dan zal elk product gebruik van deze specificatie maken, of dat nu BusinessObjects, Microsoft Excel of SAS Analytics is. Dit verhoogt de onderhoudbaarheid van de omgeving en het verlaagt de kans dat gebruikers, die verschillende producten gebruiken, inconsistente resultaten zien.

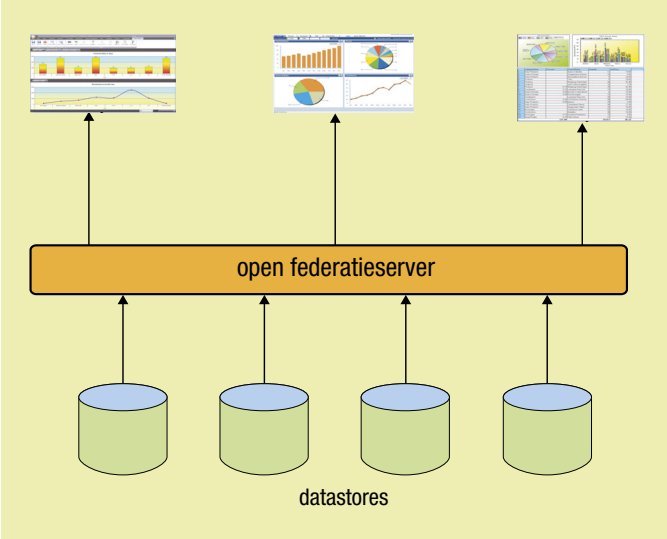
Toepassingsgebieden in een DWH-omgeving

De droomarchitectuur voor een datawarehouse-omgeving waarbij datavirtualisatie wordt toegepast, is in afbeelding 7 weergegeven. Naast de operationele databases en het datawarehouse bestaan er verder geen andere datastores. De ultieme droom zou zijn dat ook het datawarehouse niet nodig is en dat de operationele databases alle gegevens voor alle applicaties bezitten. Helaas is dat vanwege vele redenen voor de meeste organisaties niet mogelijk. Een van die redenen is dat er in hun operationele databases geen historie bijgehouden wordt.
Als er inderdaad een datawarehouse aanwezig is, zal deze met scheduled transformaties periodiek bijgewerkt moeten worden. Alle dataconsumenten (in deze omgeving zijn dat voornamelijk rapportage- en analyseproducten) zien één grote database. Als ze gegevens opvragen, zien zij niet uit welke datastores de gegevens komen. Uiteraard is dat ook niet belangrijk voor ze, mits de resultaten maar aan hun eisen voldoen. Dit kunnen eisen zijn die betrekking hebben op het kwaliteitsniveau van de gegevens, de gewenst performance, de actualiteit van de gegevens, enzovoort.
Afbeelding 7 geeft ook direct de primaire toepassingsgebieden van datavirtualisatie in een datawarehouse-omgeving weer:
Data-integratie. Het verzorgen van data-integratie, ofwel het verzorgen dat gegevens van verschillende bronnen samengebracht worden;
Data transformatie. Het verzorgen van alle benodigde transformatie- en opschoningsoperaties om de gegevens om te zetten naar een vorm die de dataconsumenten nodig hebben;
Vereenvoudiging van de architectuur. Door het ontkoppelen van de dataconsumenten en de datastores wordt het mogelijk bepaalde datastores te saneren en daarmee de gehele architectuur te vereenvoudigen.
Maar veel organisaties hebben al een complexe architectuur bestaande uit vele datastores en data-integratieoplossingen. Als daar een federatieserver ingevoerd wordt, zal de architectuur er meer uitzien zoals in afbeelding 8.

Het toepassen van datavirtualisatie in een meer doorsnee datawarehouse-omgeving maakt het mogelijk bepaalde zaken te doen, die zonder datavirtualisatie minimaal lastig zouden zijn. Hieronder staat een lijst van additionele toepassingsgebieden, waarmee datavirtualisatie zich onderscheidt van andere data-integratieoplossingen.
Virtueel datamart. Een organisatie ontwikkelt meestal datamarts om tabelstructuren aan te bieden die passen bij de behoeften van bepaalde gebruikers en rapportageproducten. Het nadeel hiervan is dat de meeste datamarts fysieke oplossingen zijn, die opgebouwd en beheerd moet worden. Een federatieserver kan gebruikt worden om het bestaan van een datamart te simuleren. In een federatieserver kunnen virtuele tabellen opgebouwd worden met dezelfde structuur als die voor een fysiek datamart. Maar in plaats dat de gegevens werkelijk opgeslagen worden, worden ze bijvoorbeeld vanuit een centraal datawarehouse via on-demand transformaties omgezet naar de verzameling virtuele tabellen. In plaats van daadwerkelijk een datastore voor een fysieke datamart te creëren en ETL-scripts te schrijven voor het laden van de datamart, is het alleen nodig om de benodigde tabelstructuren in de federatieserver te definiëren. Kortom, het ontwikkelen van flexibele en goedkope datamarts wordt hiermee mogelijk.
Virtueel centraal datawarehouse. Stel dat een Business Intelligence architectuur van een organisatie uitsluitend uit datamarts bestaat en geen centraal datawarehouse kent, maar dat er wel rapporten ontwikkeld moeten worden die de gegevens uit meerdere data-



Afbeelding 5: De specificaties van een closed federatieserver zijn uitsluitend bruikbaar voor één product.

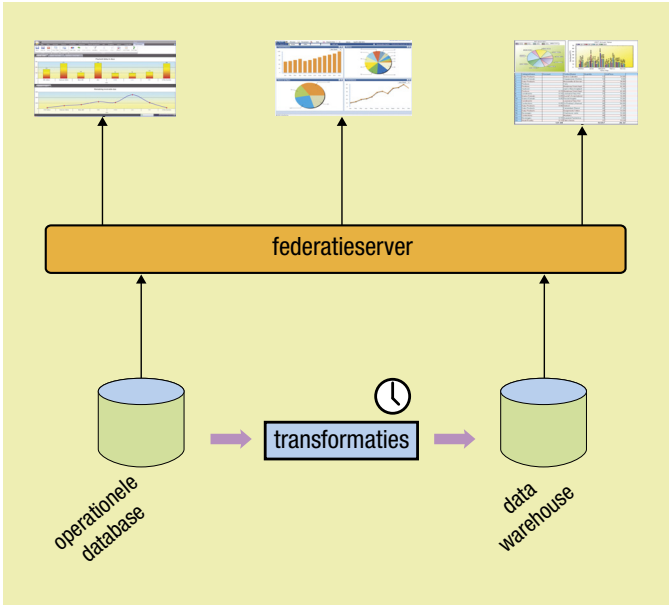


Afbeelding 6: De specificaties van een open federatieserver zijn bruikbaar voor elk product.

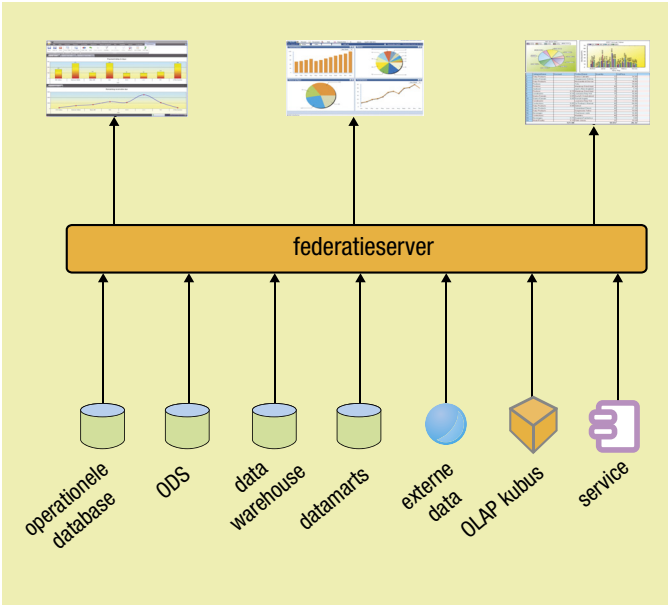
marts integreren. In dit geval kan een federatieserver uitkomst bieden. Er kan een virtueel centraal datawarehouse gecreëerd worden dat gegevens gebruikt die in de fysieke datamarts opgeslagen zijn. De federatieserver zal de tabellen uit meerdere datamarts samenvoegen op het moment dat een rapport om de gegevens vraagt. Let wel, een eis is dan wel dat de te benaderen datamarts integreerbaar zijn, dat wil zeggen dat ze kolommen bezitten waarop 'gejoined' kan worden (eventueel na transformaties).

Overkoepelend centraal datawarehouse. In sommige grote organisaties zijn onafhankelijk van elkaar verscheidene Business Intelligence architecturen ontwikkeld, elk met een eigen centraal datawarehouse. Als een rapport gecreëerd moet worden op basis van de gegevens die opgeslagen zijn in al deze centrale datawarehouses, kan dat simpelweg met een federatieserver gerealiseerd worden. In feite is dit vergelijkbaar met het vorige toepassingsgebied.

Extended datawarehouse. Voor rapporten die gegevens uit het datawarehouse moeten integreren met gegevens uit andere gegevensbronnen, zoals operationele databases, externe bronnen en lokale bestanden, zou een federatieserver een geïntegreerd beeld van al die datastores kunnen bieden. Op deze manier breidt de federatieserver het datawarehouse uit met andere data stores zonder dat gegevens uit die andere datastores naar het datawarehouse gekopieerd moeten worden. Met deze aanpak kan het datawarehouse bijvoorbeeld uitgebreid (extended) worden met gegevens uit privé-spreadsheets, externe websites en andere bestanden.



Afbeelding 7: De droom datawarehouse-omgeving.



Afbeelding 8: Datavirtualisatie in een realistische DWH-omgeving.

Operationeel datawarehouse. Door een federatieserver toegang te verlenen tot zowel een datawarehouse als operationele databases, kunnen rapporten historische gegevens uit een datawarehouse koppelen aan actuele gegevens van operationele databases die volledig up-to-date zijn. Hierbij wordt een operationeel datawarehouse gesimuleerd (soms wordt dit wel een online of near-online datawarehouse genoemd).
Self-service rapportage en analyse. Self-service rapportage en analyse houdt in dat de gebruikers zelf hun rapporten mogen ontwikkelen en hun eigen analyses mogen uitvoeren. De IT-afdeling blijft uiteraard wel verantwoordelijk voor het opzetten van de datastores voor de gebruikers. Met een federatieserver kunnen virtuele tabellen snel opgezet en aangepast worden; dit alles omdat ze virtueel zijn en niet fysiek. Dit snel kunnen reageren past prima bij het karakter van self-service BI.
Virtuele sandbox. Bij sandboxing wordt een stand-alone omgeving gecreëerd waarin analisten gegevens kunnen onderzoeken en antwoorden kunnen vinden op onduidelijke vragen. Momenteel wordt hiervoor meestal een fysieke omgeving ingericht bestaande uit een server en database- en analysesoftware. Gegevens moeten er naar toe gekopieerd worden. Het inrichten van zo'n sandbox kost veel tijd en geld. Met een federatieserver kan simpelweg een virtuele sandbox opgezet worden. Deze oplossing vereist minder werk vooraf en een kleinere investering dan wanneer een daadwerkelijke fysieke sandbox opgebouwd wordt.
Prototyping. Veronderstel dat het voor een organisatie noodzakelijk is om een complexe oplossing met behulp van datamarts en ETL-scripts te ontwikkelen. Het kan nuttig zijn om eerst een oplossing te ontwikkelen zonder de inzet van extra datastores, maar een die gebruik maakt van een federatieserver en dus on-demand transformaties. Met dit prototype wordt het bijvoorbeeld snel duidelijk welke problemen er op het gebied van transformaties en opschonen bestaan. Bovendien kunnen gebruikers met dit prototype snel inzage krijgen in de kwaliteit van de rapporten. Naderhand kan de definitieve versie ontwikkeld worden met behulp van de inzichten die met het prototype verkregen zijn, desnoods met een fysiek datamart.
Weggooi-rapporten. Af en toe moeten nieuwe rapporten met spoed ontwikkeld worden. In de meeste gevallen zullen deze rapporten slechts eenmalig gebruikt worden en hebben zij daarna hun nut gehad, vandaar de term weggooi-rapport. Gezien de urgentie is er geen tijd om een volledige fysieke omgeving in te richten bestaande uit, bijvoorbeeld, een datamart en ETL-scripts. Met een federatieserver kan deze omgeving, waarschijnlijk bestaande uit enkele virtuele tabellen, snel en eenvoudig ontwikkeld worden. Nadat het rapport is samengesteld, kunnen de virtuele tabellen weer verwijderd worden. Er is dan een minimale investering nodig geweest.

Tevens bestaan er nog vele toepassingsgebieden die buiten het Business Intelligence domein vallen. We noemen er hier enkele:
Enterprise Data Sharing. Het kan zijn dat een productieapplicatie enkele heterogene operationele databases moet benaderen. Een federatieserver zou alle technische en semantische verschillen tussen de verschillende systemen kunnen verbergen en een geïntegreerd en consistent beeld van de operationele data kunnen bieden. De operationele applicatie hoeft zich daar dan niet mee bezig te houden;
Data Services. De meeste federatieservers staan toe dat een virtuele tabel via SOAP en/of REST benaderd wordt. Met andere woorden, met federatieservers kunnen service interfaces, voor het bevragen en manipuleren van opgeslagen gegevens, redelijk snel ontwikkeld worden. Dit versnelt de ontwikkeling van een servicegeoriënteerde architectuur;
Data Mashups. Voor mashups die gericht zijn op het bevragen en manipuleren van gegevens uit verschillende gegevensbronnen, kan een federatieserver de ontwikkeling versnellen. De federatieserver zal de toegang tot alle data stores afhandelen, zal de benodigde integratie, opschoning en transformatie van gegevens voor haar rekening nemen, terwijl de mashup zich volledig op aspecten rondom de gebruikersinterface van de applicatie kan richten.

Afsluiting

In dit artikel is getoond dat datavirtualisatie veel toepassingsgebieden voor een datawarehouse-omgeving kent bovenop de bekende toepassingsgebieden die voor elk soort data-integratieoplossing gelden. Deze extra toepassingsgebieden maken dat datavirtualisatie van een datawarehouse-omgeving een flexibele omgeving maakt die past bij de nieuwe wensen van organisaties. Het invoeren van datavirtualisatie hoeft geen revolutie te zijn. Stap voor stap kan deze technologie ingevoerd worden. Met elke stap komt een datastore via de federatieserver beschikbaar en kunnen steeds meer rapporten via de federatieserver gegevens ophalen.

Op het moment dat alle rapporten die een bepaalde datastore via de federatieserver werken benaderen, dan kan bepaald worden of deze datastore uitgefaseerd kan worden en volledig virtueel opgebouwd kan worden. Hiermee begint dan de vereenvoudiging van de architectuur. En hoe meer datastores zijn afgebouwd, hoe flexibeler de architectuur zal worden en hoe meer de kosten van de gehele omgeving zullen afnemen. Simpelweg: virtuele oplossingen zijn flexibeler dan fysieke oplossingen en datavirtualisatie bewijst dat.

Noten

1. www.b-eye-network.com/channels/5087/view/14815
2. www.compositesw.com/index.php/resources/data-virtualization-leadership-series

Rick van der Lans is zelfstandig IT-consultant.