Alternatieve benadering wordt 'mainstream' technologie

Data Virtualization

Bram Dons

Volgens onderzoekshuis IDC groeit het datavolume elke vijf jaar met een factor tien, wat neerkomt op een jaarlijkse groei van bijna 60 procent. Deze data zijn verspreid over verschillende plaatsen, die daar als gevolg van overnames van ondernemingen en het samengaan van datacenters terecht zijn gekomen.

Het is dan ook niet zo verwonderlijk dat ondernemingen en overheden grote moeite hebben om de juiste informatie uit deze verschillende en verspreide datasystemen op te vragen. Bovendien maken bedrijven als gevolg van de ontwikkeling op internet steeds vaker gebruik van data die zich op remote systemen bevinden; iets wat de zoektocht nog eens extra bemoeilijkt.
De meeste ondernemingen en organisaties bezitten informatie op verschillende plaatsen die (vanwege overheidsrestricties, incompatibele data, of onwil vanwege organisatorische of commerciële redenen) niet in een enkele data store zijn te verenigen. Er bestaat een voortdurende ongelijkheid tussen de manier waarop data worden opgeslagen (format, structuur, API's, enzovoort) en de manier waarop deze data in rapporten, portals en andere applicaties worden gebruikt. Bijvoorbeeld, veel webgebaseerde applicaties gaan van hiërarchisch gestructureerde XML data uit, maar de daarvoor gebruikte data kunnen wel eens uit een relationele database tabel komen. Vergelijkbaar is de SAP-rapportage die vaak op SQL is gebaseerd met data die afkomstig zijn van SAP-specifieke datastructuren.

Niettemin, ondernemingen willen deze verspreide en incompatibele informatie combineren die hen kan helpen bij de realisatie van verschillende zakelijke doelstellingen, waaronder verhoging van de productiviteit en inkomsten, verlaging van de kosten en risico. Een manier om dit te verwezenlijken is hardwarematig verbindingen tussen de verschillende applicaties aan te brengen en de daarbij betrokken databronnen. Dit is natuurlijk een dure methode, het is tijdrovend om te implementeren en bovendien inflexibel in gebruik en onderhoud. Een alternatieve benadering is gebruik maken van datavirtualisatie, ook wel aangeduid als data federation of Enterprise Information Integration (EII). Daarvoor zijn geen specifieke programmatische koppelingen nodig, waardoor het een flexibele oplossing oplevert; let wel, de term 'datavirtualisatie' zou, maar mag niet, verward kunnen worden met 'datacentre virtualisatie'. De uitdrukkingen 'data federation' en 'EII' worden geassocieerd met bepaalde zakelijke problemen, waaronder query processing via operationele en datawarehousing omgevingen of registry-gebaseerde masterdata managementoplossingen. Datavirtualisatie beperkt zich echter niet tot de ondersteuning van genoemde zaken, maar is een uitbreiding daarop. Het biedt een oplossing voor de koppeling van datamarts en warehouses in een uitgebreide analytische omgeving, enterprise sharing van data, ondersteuning van infrastructuur voor real-time zaken en de integratie van interne bronnen die in SaaS-applicaties of Cloud-omgevingen zijn *outsourced* of *hosted*.

Composite Data Virtualization Platform

Datavirtualisatie is nu een 'mainstream' technologie aan het worden die door grote ondernemingen om uiteenlopende redenen wordt toegepast. Vele van dit soort toepassingen zijn echter specifiek bedoeld voor een bepaalde omgeving en ingebed binnen een Business Intelligence- of andere datamanagement-oplossing. De firma Composite denkt met haar query middleware product 'Composite Data Virtualization' zich te onderscheiden van de andere leveranciers door een heterogene, algemeen toepasbare, datavirtualisatie-oplossing te bieden. Composite's Data Virtualization integreert data (afkomstig van meerdere, aparte, bronnen, waar dan ook binnen de onderneming) op een eenduidige, logische, gevirtualiseerde wijze, wat het geschikt maakt voor gebruik in bijna elke zakelijke frond-end oplossing. Het Composite Data Virtualization Platform product (zie afbeelding 1) vormt een compleet ontwikkel- en runtime-platform dat inspeelt op de vijf fundamentele uitdagingen van dataintegratie, te weten: datacomplexiteit, structuur, locatie, volledigheid en actualiteit. Het Platform bestaat uit de Information Server en een aantal aanvullende opties en producten ter ondersteuning van een

complete live-cycle software ontwikkelomgeving. De Information Server is een op Java gebaseerde server die zich op een 'non-invasively' manier toegang verschaft tot bestaande data, verspreide data verenigt, complexe data vereenvoudigt en deze als data services of relationele views toont aan de gebruiker. Het platform kent twee ontwikkelomgevingen: Studio, voor de traditionele databasegeoriënteerde ontwikkelaar en Designer, voor de servicesgeoriënteerde ontwikkelaar die bekend is met een Eclipse-gebaseerde omgeving.

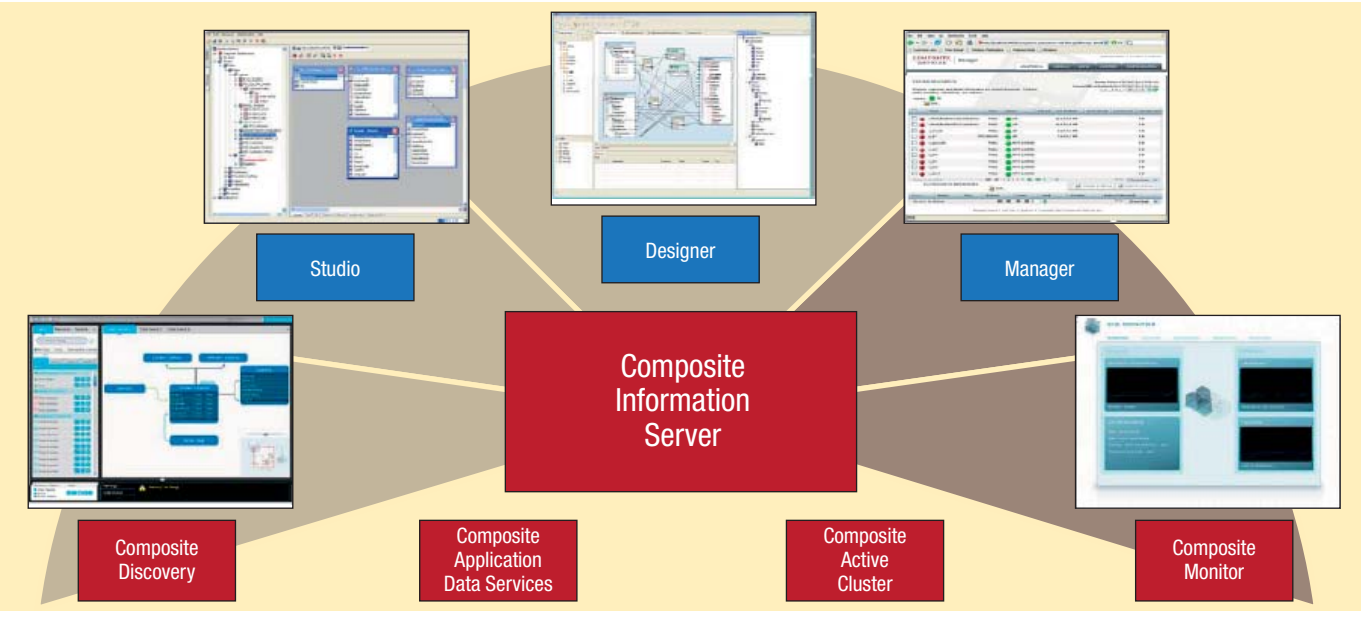
Werking Data Virtualization

In tegenstelling tot andere leveranciers, die recentelijk simpelweg datavirtualisatie aan hun bestaande traditionele ETL- en BI-oplossingen hebben toegevoegd, heeft Composite Software bijna tien jaar besteed om de moeilijkste datavirtualisatieproblemen voor de grootste ondernemingen op te lossen, aldus de firma Composite. De firma zegt een standaard te hebben gesteld voor de snelste query optimalisatiealgoritmen en -technieken, gecombineerd met een hoog schaalbare architectuur en voorzien van data caching-opties. De vraag is hoe de firma Composite dat voor elkaar heeft gekregen.
Zoals gezegd, het Platform is tegelijkertijd een complete software ontwikkel- en run-time omgeving. Tijdens de ontwikkelfase is Composite Discovery behulpzaam bij het zoeken naar en modelleren van sleutelentiteiten en relaties. In de bouwfase maken ontwikkelaars van Composite Informations Server's twee eenvoudige ontwikkelomgevingen gebruik (relationeel en XML). Deze zijn voorzien van automatische codegenerators voor de creatie van kwalitatief hoge, semantisch betekenisvolle, standard compliant views en data services. Geavanceerde tools helpen bij de complexe federatieve en transformatieve functies. Standaard adapters maken een eenvoudige toegang tot data mogelijk en de publicatie van de ontwikkelingsactiviteiten. De Manager bewaakt de beveiliging, bestuurt de metadata, broncode en voert andere beheertaken uit. De Composite Applications Data Services, met pre-built objecten voor de toonaangevende ERP suites en SQL naar MDX vertalers, dragen bij aan de automatisering en versnellen de *critical view* en *data service* ontwikkelingsactiviteiten. Tijdens run-time voert de Information Server's query engine nauwkeurig query's uit, biedt toegang, abstraheert en levert de data op aanvraag af aan de zakelijke oplossingen. Verschillende caching-opties bieden extra snelheid en flexibiliteit. Verder is de Information Server gebaseerd op een volledig schaalbare architectuur, waarbij de Active Cluster optie zorgt voor load balancing, high availability en een fail-over voorziening voor toepassing in een 24x7 enterprise omgeving.

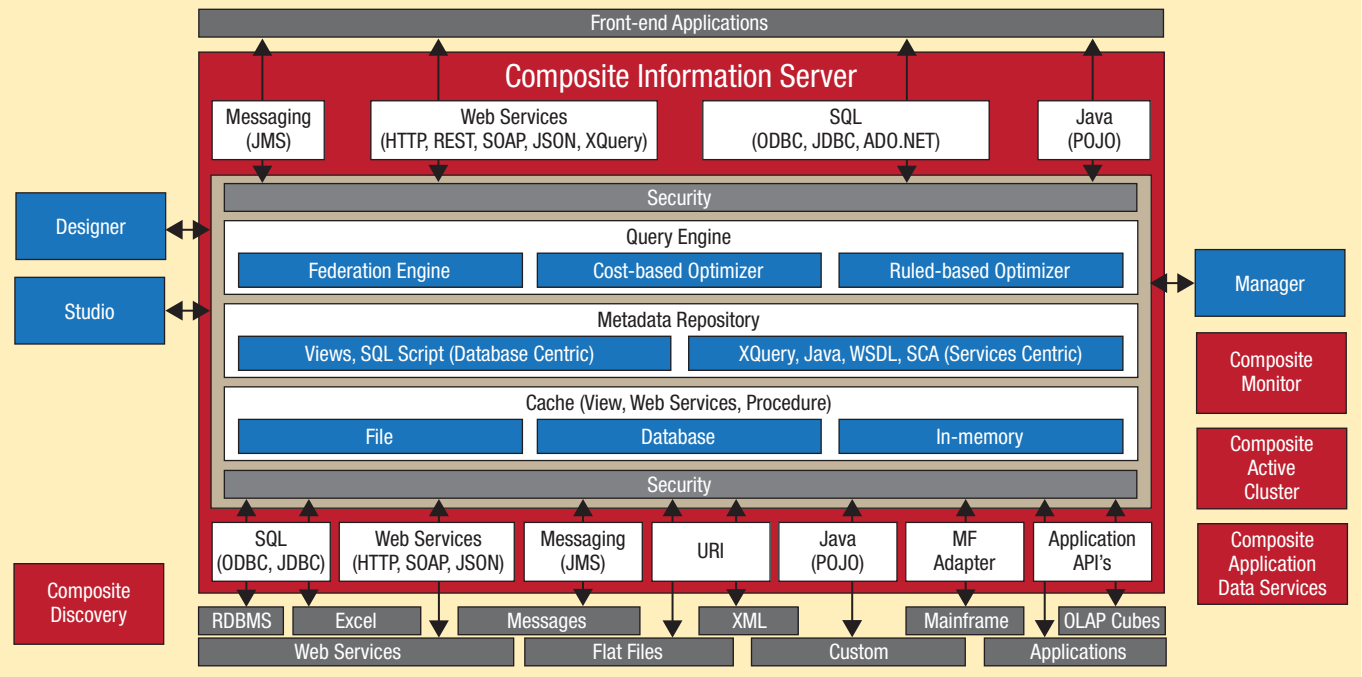
Composite Information Server en Discovery zijn pure Java applicaties die zijn geïmplementeerd als een Linux-gebaseerde appliance met browser-gebaseerde clients en maken van AJAX gebruik voor de ondersteuning van de interactieve zaken. Het platform draait op Windows 2000/2003/2007/2008 en Vista, Solaris 9, Linux (Red Hat en SUSE), AIX en HP-UX. Studio en Designer ondersteunen slechts de Windows 2000, XP en Vista clients. LDAP en Active Directory worden ondersteund en product features als pass-through login, single sign-on en SSL. De ondersteunde publishing interfaces zijn: JDBC, SOAP, ADO.NET en REST. Verder worden de diverse versies van de meest bekende data sources ondersteund, waaronder: DB2, SQL Server, MySQL, Informix, Oracle, Sybase, Teradata, XML en Tibco.

Composite Data Virtualization architectuur

De Information Server vormt de kern van het Composite Platform. Het verbindt meerdere complexe datastructuren en 'streamt' de daaruit verkregen query-resultaten naar de clients. De onderste laag van de Information Server stack wordt gevormd



Afbeelding 1: Composite Data Virtualization Platform (Bron: Composite).



Afbeelding 2: Composite Information Server Architectuur (Bron: Composite).

door de data source toegangslaag. Deze voorziet in een protocol- en/of leverancierspecifieke toegang tot alle afzonderlijke data sources, nodig voor een datavirtualisatie implementatie. In de ontwerpfase onderzoekt het de data source om de structuur en datatypes aan de Information Server door te geven. Tijdens runtime normaliseert en 'streamt' de Server de data in een format dat efficiënt door de Information Server kan worden gemanipuleerd. Elk in de Server beschikbare data source type maakt van een specifieke daaraan gekoppelde driver gebruik. Bij het toevoegen van een nieuwe data source worden de driver's parameters aan een specifiek type data source gekoppeld (bijvoorbeeld een SQL database). Bij het tot stand komen van de verbinding onderzoekt de Server de structuur en de datatypes en slaat de daarvan afgeleide metadata op in de Server's metadata repository. Daarna zijn de metadata beschikbaar om, op basis van de corresponderende data source, views en data services te creëren.

Een van de belangrijkste access layer taken is om zowel de vorm als het type data in een van de drie datavormen te normaliseren: tabellen (rijen en kolommen), hiërarchisch (bijvoorbeeld een XML document), of scalar (enkelvoudige waarde). De op basis van deze standaardvormen genormaliseerde data zijn daarna eenvoudig te manipuleren, te transformeren en te combineren met andere data, zonder rekening te hoeven houden met de leverancier- of protocolspecifieke details. Let wel, hoewel vorm en datatype door de access layer zijn getransformeerd, blijven de oorspronkelijke waarden onveranderd tijdens de doorloop in de data virtualization layer. De Server kan dus data van uiteenlopende data sources integreren.
In het kort geven we een opsomming; voor een volledige lijst zie de data sheet op de Composite website:

- Relationele database: data sources in tabelvorm, bijvoorbeeld Oracle, DB2 en SQL Server, waarbij alle relationele databases van een leverancierspecifieke JDBC-driver gebruik maken;
- Web services: data die beschikbaar zijn via gestandaardiseerde web service interfaces, waaronder SOAP, XML/HTTP en REST;
- Multi-dimensionale sources: data afkomstig van SAP BW en Oracle EssBase data sources met behulp van SQL query's transformatie naar MDX query's; daarbij worden in feite multi-dimensionale data getransformeerd naar plakjes tabeldata zodat de server deze kan samenvoegen, abstraheren of voor andere functies gebruiken;
- Packaged Applications: de meeste industriestandaard applicaties, zoals SAP en Siebel, gebruiken hun eigen leverancierspecifieke API's. De Server voegt aan deze API's een aantal Application Data producten toe, waarmee deze API's en key data-entiteiten als common objecten kan gebruiken;
- Mainframes: toegang tot mainframe databases (bijvoorbeeld DB2) via JDBC of speciale driver die beschikbaar is gekomen door een partnerschap van Composite met Data Direct;
- Files: toegang tot bepaalde gestructureerde bestanden, waaronder delimited tabular, XML en Excel spreadsheets;
- Common: voor sommige data sources zijn geen standaardgebaseerde API's beschikbaar, Information Server biedt de mogelijkheid om custom drivers te ontwikkelen op basis van Java.

Query Processing

De kern van de Composite Server architectuur wordt gevormd door een hoogpresterende query processing engine. Deze maakt van gedistribueerde, door queryplan geoptimaliseerde, data streaming technologieën gebruik om snel de samengestelde resultaten van een query op te kunnen vragen. De engine past

uitgekiende rule- en cost-based query optimalisatiestrategieën toe die voor elke query een plan opstellen. Door gebruik te maken van SQL Pushdown, parallelverwerking, gedistribueerde joins, caching en geavanceerde query-optimalisatie kunnen efficiënte joins methods worden toegepast. Deze presteren beter dan de handgecodeerde query's zoals die doorgaans worden geschreven door een doorsnee ontwikkelaar. Algemeen doel van deze optimalisatietechnieken is om de hoeveelheid data die voor een query over het network moet worden verstuurd, zo laag mogelijk te houden.

Caching

De Information Server maakt om twee redenen gebruik van zogenoemde 'result-set' caching. Ten eerste; het cached resultaat is onmiddellijk beschikbaar voor een nieuw request waardoor de opvraagvertraging wordt verminderd. De tweede reden is om de potentiële impact op de onderliggende data sources te verminderen. Tijdens de query planningsfase kan de query optimizer herkennen dat bepaalde resources al in cache zijn opgeslagen en op basis daarvan het queryplan aanpassen om snel data uit de cache op te kunnen halen, in plaats vanuit de oorspronkelijke data source. De cache policy voor een resource wordt tijdens de ontwikkelfase bepaald, waardoor de ontwerper de storagelocatie kan voorschrijven, de updatestrategie en tijd.

Composite Studio

De Studio vormt voor de ontwerper en beheerder de primaire modelleer-, view- en resource managementomgeving. De omgeving is gebaseerd op een boomstructuur waarin de beschikbare fysieke data sources zijn afgebeeld en biedt een werkblad waarin de query's kunnen worden gecreëerd en getest, en een data services gedeelte waarin onder meer views, data services of caches worden getoond. Studio is een datamodelleeromgeving die veel overeenkomt met de 'look and feel' van andere omgevingen waarmee IT-ontwikkelaars al bekend zijn.
De relationele views, of alleen views, zijn binnen Information Server resources als tabulaire data geïntegreerd. In feite zijn ze conceptueel equivalent aan traditionele database views en zien er voor de gebruiker dan ook uit als een gewone databasetabel. De view kan een uiteindelijke integratie voorstellen maar kan ook weer worden gepubliceerd voor gebruik van clients in een SQL query. Anderzijds, een view kan ook gepresenteerd worden als tussenresultaat voor weer andere integraties (views en andere resources). Views in Composite worden gedefinieerd als een enkel SQL statement. De Studio ontwikkelomgeving biedt een grafische editor om views te creëren met behulp van 'drag-and-drop' technieken waarna het SQL resultaat automatisch wordt gecreëerd; een SQL statement is naar wens ook handmatig te creëren. Het belangrijkste verschil in de ontwikkeling tussen een traditionele view en een met Information Server gecreëerde view is dat de Composite views data kunnen lezen die van meerdere data sources afkomstig zijn. Dit gedrag is in Composite SQL geïmplementeerd door in elke source in de Composite Information Server te refereren met de FROM clausule van de view's SQL statement; de view's SQL conformeert zich aan de SQL-92 standaard.

Composite Designer

Composite Designer is een op Java gebaseerd Eclipse Framework dat door applicatie- en Java-ontwikkelaars wordt gebruikt voor ontwerp, ontwikkeling en testen van data services. De Designer ondersteunt zowel Contract-First en Design-by-Example methoden, evenals SCA en elke XML-standaard. Data services zijn web service-operaties die data aan de requestor retourneren en de basis vormen voor geïntegreerde data binnen de context van een SOA. Van de buitenkant zien data services er als elke andere web service-operatie uit. Het gebruikt een XML-document als input en produceert XML als output. Naast Composite ondersteunt de Information Server met Designer ook de ontwikkeling van data services: Composite voor de relatief rechttoe rechtaan data services; Designer voor de meer complexe services. Data services kunnen als SOAP of REST-models worden gecreëerd waarbij het transport over HTTP of JMS kan plaats vinden.

Additionele platformopties

Composite biedt een aantal optionele producten als uitbreiding op de Information Server die gezamenlijk de Data Virtualization Platform architectuur compleet maken. De Discovery optie onderzoekt data en brengt verborgen relaties naar voren tussen entiteiten in de verschillende enterprise datasilo's. Discovery indexeert de metadata en data van de verschillende entiteiten en laat daar algoritmen op los om vast te stellen of er mogelijk een relatie bestaat. De Application Data Services maken het gebruik van packaged applicaties mogelijk, waaronder SAP, Siebel en Oracle E-Business Suite en levert deze als voorgebouwde, herbruikbare, data services af. De Composite Monitor voorziet in een real-time overzicht van het complete Data Virtualization Platform. De Systems Management optie bewaart het overzicht over de actieve views en resources en is gefocust op de configuratie, monitoring en onderhoud van de draaiende Information Server.

Tenslotte de Active Cluster optie. Daarmee is de schaalbaarheid van Information Server toepassingen te vergroten zodat gebruikers requests sneller kunnen worden afgehandeld. Voor ondersteuning van high availability wordt een cluster gecreëerd op aparte servers, wat Composite beschermt tegen server hardwarefouten. De Server clusterarchitectuur is van het type peer-to-peer waarbij elke node zijn eigen kopie van de metadata repository bevat die continu met alle andere nodes wordt gesynchroniseerd. Er kunnen maximaal zestien nodes in een cluster bestaan.
Informatie op internet: www.compositesw.com

Bram Dons is onafhankelijk IT-consultant.