Nico Jacobs en Wim Uyttersprot zijn beiden verbonden aan U2U.

Microsoft SQL Server 2005 Analysis Services

VAN GEGEVENSBEHEER NAAR INFORMATIEBEHEER

Databases zijn een vast onderdeel in zowat elk IT-project. Naar mate de hoeveelheid opgeslagen data toeneemt, worden taken als rapportage en analyse alsmaar belangrijker. Vandaar dat Microsoft in de nieuwe versie van Microsoft SQL Server ook veel aandacht heeft besteed aan de OLAP-component Analysis Services. In dit artikel beschrijven de auteurs de belangrijkste kenmerken van Microsoft SQL Server 2005 Analysis Services, met de nadruk op de nieuwe aspecten.

SQL Server: meer dan transacties alleen

Hoewel databases voornamelijk gebruikt worden voor het verwerken van transacties, is de laatste jaren een sterke groei merkbaar in applicaties die de databases gebruiken voor rapportagedoeleinden of voor analyse van gegevens. Hoewel beide dezelfde gegevens als basis hebben, zijn er toch essentiële verschillen tussen beide soorten toepassingen. Daarom zijn er in de databasewereld twee soorten databases: OnLine Transaction Processing (OLTP) databases, zoals SQL Server, en OnLine Analytical Processing (OLAP) databases, zoals Analysis Services. OLAP-databases zijn gericht op het *verwerken* van gegevens. Vaak gaat het hier om geaggregeerde waarden zoals som, gemiddelde, enzovoort. We zien dat OLAP-databases, bij data- en rekenintensieve queries, de gegevens op een andere manier structureren en bepaalde geaggregeerde waarden op voorhand berekenen. Hierdoor verloopt het raadplegen via queries van dergelijke databases efficiënter dan wanneer men dezelfde query op een OLTP-database zou verrichten.

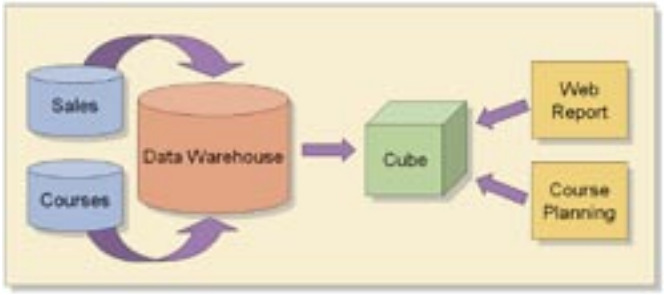
Microsoft heeft al jaren een OLAP-component die samen met SQL Server geleverd wordt: Analysis Services. Dit is een populair product geworden bij OLAP-gebruikers. De cijfers van OlapReport (http://www.olapreport.com/market.htm) tonen aan dat Analysis Services 2000 veruit de meest gebruikte OLAP-engine is. Microsoft SQL Server Analysis Services 2005 heeft inmiddels verschillende grote stappen vooruit gezet in vergelijking met de 2000-editie, zowel op het gebied van robuustheid, performance en functionaliteit als gebruiksvriendelijkheid. Hoe verloopt nu het opstellen en gebruiken van een OLAP-database?

Van OLTP naar OLAP

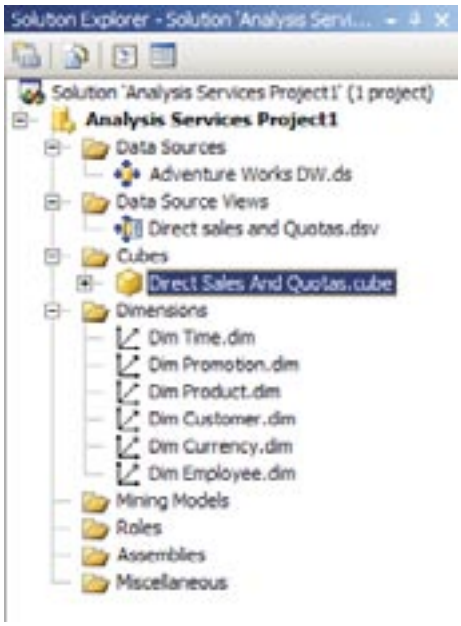
Net als met een OLTP-database begin je een OLAP-database door gegevens te importeren vanuit een relationele database. Zoals we ondertussen van Microsoft gewend zijn, kan je hierbij gebruik maken van elke gegevensbron die OLE DB of ODBC ondersteunt. Afhankelijk van het aantal gegevensbronnen en de kwaliteit van de gegevens, zullen er eerst stappen nodig zijn om de gegevens foutenvrij te maken (data cleansing) en in een centrale database (datawarehouse) te plaatsen. Hieruit worden dan de kubussen (cubes) gecreëerd die door de verschillende applicaties kunnen worden gebruikt (verderop in dit artikel meer over dit kernconcept). Dit wordt geïllustreerd in afbeelding 1. Afbeelding 2 toont de Solution Explorer met daarin de belangrijkste componenten van een Ana-

Microsoft SQL Server 2005 Analysis Services
Integratie in Microsoft Visual Studio
IntelliCube
Meer Fact-tabellen
XML-representatie
Perspectives
Vertalingen
Proactive caching
Key Performance Indicator
Bijkomende datamining-algoritmen

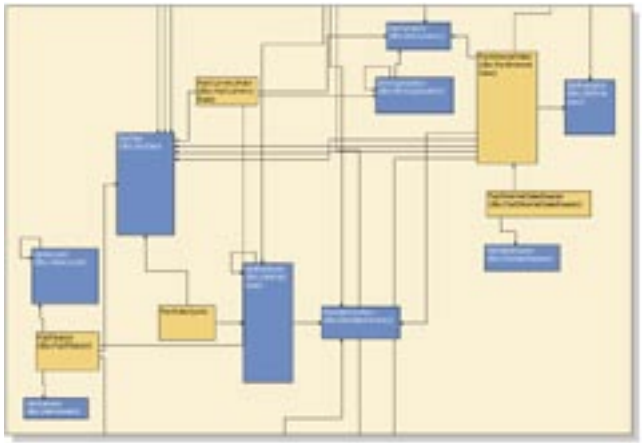
Tabel 1. Overzicht van de nieuwe kenmerken van Analysis Services 2005



Afbeelding 1. Schema dat de basis vormt voor het opstellen van de kubussen



Afbeelding 2. De belangrijkste aspecten van een Analysis Services-project



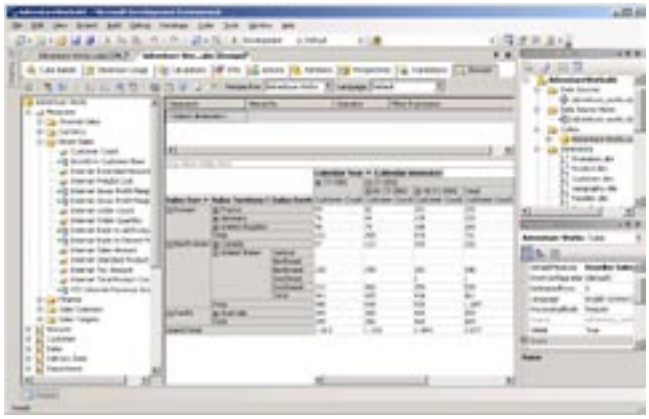
Afbeelding 3. Fragment van een schema met een derde normaalvorm (3NF)

lysis Services 2005-project. Een nieuwe term in Analysis Services 2005 is de Data Source View. Hierbij selecteert de ontwerper de tabellen uit de onderliggende OLTP-database en beschikt hij over de optie om gerelateerde tabellen automatisch mee te selecteren. Het resultaat hiervan is een schema dat de basis vormt voor het opstellen van de kubussen.

Kubus, de belangrijkste bouwsteen

Zodra de gegevens bekend zijn, kan de belangrijkste stap gezet worden in de opbouw van de OLAP-database, namelijk het definiëren van de relaties tussen de verschillende tabellen. Daar waar bij een OLTP-database de relaties tussen de data opgegeven worden 'at query-time', worden deze bij een OLAP-database 'at design time' vastgelegd. Het is immers deze extra kennis die de OLAP-database toelaat efficiënter queries te beantwoorden. De belangrijkste concepten zijn measures en dimensies. Measures zijn de velden waarover je informatie zoals totalen, gemiddelden enzovoort wilt krijgen, vergelijkbaar met de velden in een SELECT-statement. Dimensies bestaan uit de velden die je als groeperingcriterium wilt gebruiken. Deze zijn gelijk aan de velden die je in een GROUP BY-statement gebruikt. Wanneer je bijvoorbeeld geïnteresseerd bent in verkoopcijfers en verkoopaantallen over product, regio en/of tijd vormen verkoopcijfers en verkoopaantallen jouw measures, terwijl product, regio en tijd velden de dimensies vormen.

Op dit punt zijn er vele belangrijke nieuwigheden. Om te beginnen wordt de taak van het opstellen van het schema van de OLAP-database geautomatiseerd. Dit is het vastleggen van measures, dimensies en hun onderlinge relaties. De OLAP-engine analyseert de gegevens in de onderliggende OLTP-database en extraheert hieruit eigenschappen van deze velden, evenals de relaties tussen deze velden. Op basis hiervan worden de beschikbare velden opgesplitst in measures en dimensies. Een dimensie (of beter: een



Afbeelding 4. Het aanmaken, deployen en browsen van een OLAP-cube is geïntegreerd in Visual Studio

Afbeelding 5. Fragment uit de XML-representatie van een kubus

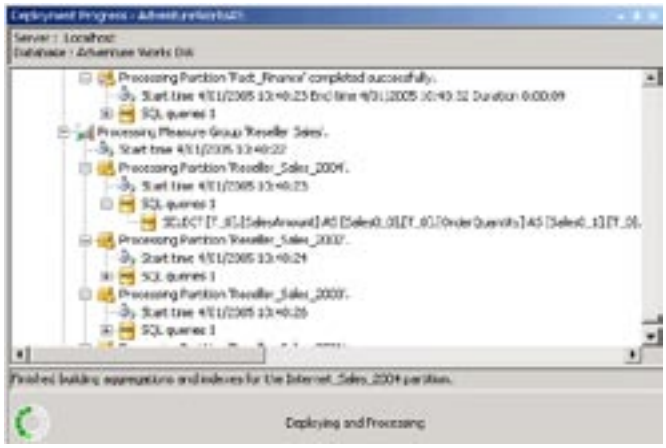
hiërarchie in een dimensie) kan uit meer niveaus bestaan. Een regiodimensie kan bijvoorbeeld uit een stad-, een provincie- en een landniveau bestaan. Analysis Services 2005 detecteert deze afhankelijkheden automatisch en plaatst dergelijke velden op de juiste manier in een hiërarchie (IntelliCube). Natuurlijk kun je het resulterende schema nog handmatig aanpassen of het volledige ontwerp handmatig maken.

Afbeelding 3 toont een fragment van een schema met een derde normaalvorm (3NF): de structuur van de onderliggende gegevens kan vanaf Analysis Services 2005 complexer zijn dan de klassieke ster- en sneeuwvlokschema's.

In het verleden waren er beperkingen op de structuur van de tabellen waaruit deze dimensies werden afgeleid. Er kon maar één tabel zijn die de measure-velden bevatte. Daarnaast waren de relaties tussen deze 'fact table' en de dimensievelden beperkt tot de zogenaamde sterschema's en sneeuwvlokschema's. Analysis Services 2005 laat echter toe om verscheidene fact-tables te gebruiken in het schema en kan alle informatie aan die in de derde normaalvorm gerepresenteerd is; zie afbeelding 4. Ook is het vanaf nu mogelijk om in een hiërarchie niet-rechtstreeks gerelateerde velden op te nemen. Een belangrijke stap op gebied van stabiliteit is dat erg grote dimensies nu geen probleem meer vormen voor de Multidimensionale OLAP-database (MOLAP). Dit vormt dus geen reden meer om naar de minder goed presterende Relationele OLAP (ROLAP) over te stappen (zie de sectie over proactive caching voor meer informatie over deze twee manieren van dataopslag). Dit alles maakt het mogelijk een bredere waaier van gegevens op te nemen in je OLAP-database.

Integratie in Visual Studio 2005

Het opstellen van schema's gebeurt ook op een gebruiksvriendelijkere manier. Om te beginnen is de functionaliteit van de Analysis Manager en andere tools van Analysis Services 2000 nu geïntegreerd in de Business Intelligence Development Studio van Visual Studio 2005, zodat ontwikkeling voor Analysis Services in dezelfde vertrouwde omgeving kan gebeuren als ontwikkeling van



Afbeelding 6. Informatie over het deployment-proces wordt in Visual Studio weergegeven

windows- of webapplicaties; zie afbeelding 1. Dit past in de ver doorgevoerde vereenvoudiging van de IDE. De SQL Server OLTP-functionaliteit wordt immers ook in een omgeving gebundeld, namelijk SQL Management Studio. In Analysis Services 2005 is er ook geen permanente verbinding meer nodig met de onderliggende OLTP-database: Alleen bij de start van de schemacreatie en bij het automatisch opstellen van dimensies is er nog OLTP-toegang nodig. Verder gebeurt het ontwikkelen van de schema's volledig asynchroon. Een ander belangrijk verschil met de vorige versie van Analysis Services is dat het resultaat van de ontwikkeling van de OLAP-database een XML-bestand is; zie afbeelding 5.

Dat betekent dat het erg gemakkelijk wordt om een OLAP-database via programmering te gaan ontwikkelen. Zoals we ook van andere producten gewend zijn, kunnen we vanuit Visual Studio de ontwikkelde database ook gaan deployen. Het XML-bestand wordt dan naar de Analysis Services gestuurd waar de nodige data uit de onderliggende database opgehaald worden en de eigenlijke OLAP-database geconstrueerd wordt. Gedetailleerde informatie over dit deployment-proces wordt hiërarchisch weergegeven in afbeelding 6. Wat hier een beetje ontbreekt is een globale indicator die aangeeft welk percentage van het deployment-proces al is afgerond.

Perspectives

Een vaak voorkomend probleem bij het opstellen van een OLAP-schema is dat er vele dimensies en measures zijn die potentieel nuttig zijn. Wanneer deze allemaal in de OLAP-database opgenomen worden, betekent dit overweldigend veel werk voor de analist of de persoon die de rapporten ontwerpt. Deze stellen meestal een of meer pivot-tabellen op, waarbij ze measures in het datagrid plaatsen en een of meer dimensies op de assen van de tabel. Als OLAP-ontwikkelaars kunnen we het deze gebruikers gemakkelijker maken door een perspective aan te maken. Een perspective is een kijk op een kubus: het definieert de subset van measures en dimensies die aan een gebruiker gepresenteerd worden wanneer deze de database gaat queriën. Op deze manier kunnen we verschillende perspectives op een kubus baseren, die toelaten dat verschillende gebruikersgroepen elk hun eigen kijk op de kubus hebben. Wanneer deze perspectives direct op de kubus gemapt worden, heeft het aanmaken van deze perspectives geen invloed op de opslag, noch op de verwerkingstijd van de kubus. Let wel op, dit mechanisme is niet bedoeld als een beveiligingsmechanisme. Wanneer je bepaalde gebruikers de toegang tot bepaalde velden wilt ontzeggen, moet je hiervoor gebruik maken van het rolgebaseerde veiligheidsmodel op de onderliggende kubus, zoals dit al in Analysis Services 2000 het geval was.

Vertalingen

In een veeltalige regio zoals West-Europa is het niet ongewoon dat mensen die een verschillende taal spreken eenzelfde kubus raadplegen. Analysis Services 2005 ondersteunt daarom vertalingen. Voor alle objecten (zoals measures, dimensies en dimensieniveaus) kunnen we nu in verschillende talen een naam opgeven. Wanneer een client gegevens uit de kubus opvraagt, zal deze in zijn connectie een Locale Identifier (LCID) meegeven. Aan de hand daarvan zal Analysis Services in zijn respons de namen gebruiken die in de bijbehorende taal ingegeven zijn. Is er voor deze taal geen vertaling voorzien in de kubus, dan zal het systeem terugvallen op de vertaling die als standaardvertaling gemarkeerd is. Voor de eigenlijke data is het ook mogelijk om in een vertaling te voorzien. Dit gebeurt door voor elk object waarvoor we een vertaling willen, in de corresponderende tabel in de relationele database een kolom te maken met de vertaling. Dit gebeurt nu al in grotere OLTP-databases voor bijvoorbeeld productbeschrijvingen en dergelijke.

Digitale dashboards en rapportering

Zoals al eerder vermeld, vormt rapportage een belangrijke toepassing van OLAP-databases. Een extreme vorm van rapportage zijn digitale dashboards: beleidsmensen steunen op enkele belangrijke bedrijfsindicatoren bij het nemen van hun beslissingen. Er zijn verschillende toepassingen die de visualisatie van dergelijke Key Performance Indicators (KPI) mogelijk maken. In Analysis Services 2005 is het mogelijk een Multi Dimensional Expression (MDX) formule op te geven die een dergelijke KPI bepaalt. Naast de eigenlijke waarde verwacht het systeem ook een doel (de waarde die we willen bereiken), een status (een formule die aangeeft hoe dicht we al bij dit doel zitten) en een trend (een formule die aangeeft in welke mate we de goede kant op gaan). Ook kunnen we opgeven met welke iconen we de informatie willen visualiseren. Let wel op: de eigenlijke visualisatie ligt bij de client-applicatie. De instellingen van onze server worden gewoon als property naar de client gestuurd. Afbeelding 7 toont de visualisatie van twee KPI's in Visual Studio.

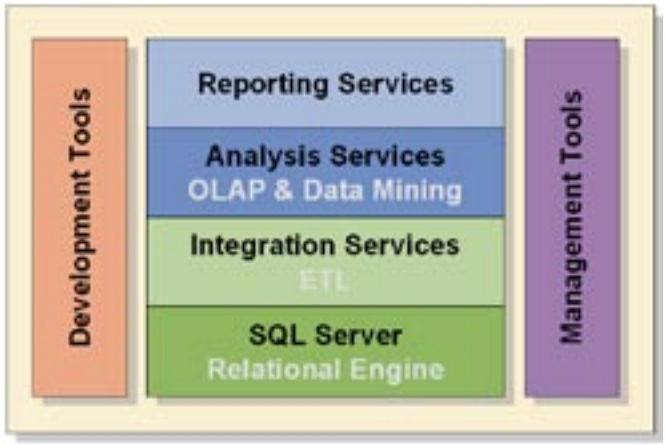
Uitgebreidere rapportagemogelijkheden zijn beschikbaar via SQL Server 2005 Reporting Services waarmee een brede waaier aan rapporten kan worden opgesteld. Dankzij een nauwe integratie van ActiveViews zal de constructie van rapporten nu ook door de zakelijke eindgebruiker kunnen gebeuren. Afbeelding 8 illustreert de samenhang tussen al deze componenten.

Proactive caching

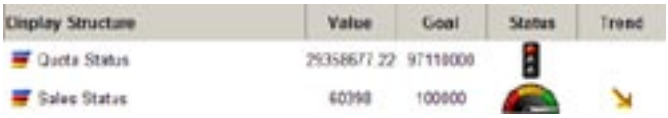
Een OLAP-kubus wordt gebouwd bovenop een relationele database. De OLAP-kubus moet daarom af en toe synchroniseren met de relationele data. In Analysis Services 2000 waren er drie opslagmodellen:

- MOLAP: waarbij de data uit de relationele database in een multidimensionale database gerepliceerd worden
- ROLAP: waarbij elke OLAP-query vertaald wordt naar een query tegen de relationele database
- HOLAP: de hybride vorm waarbij geaggregeerde waarden berekend en bewaard worden in een multidimensionale database, maar de basisdata niet gedupliceerd worden. Indien deze nodig zijn blijft een toegang tot de relationele database nodig.

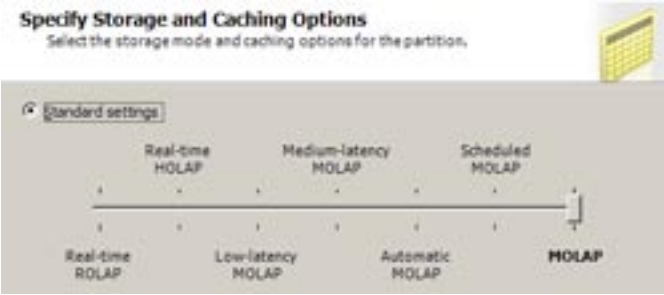
Met de uitzondering van ROLAP zonder aggregaatswaarden, was in elk van deze scenario's een expliciete opdracht tot synchronisatie nodig; iets wat vaak met DataTransformation Service (het huidige Integration Services) gebeurde. In Analysis Services 2005 echter is proactive caching ingebouwd, waardoor het systeem zelf de onderliggende database kan monitoren en, als er wijzigingen optreden, deze automatisch doorvoert in de OLAP-database. Wanneer SQL Server de onderliggende database is, gebeurt dit via een notificatie-



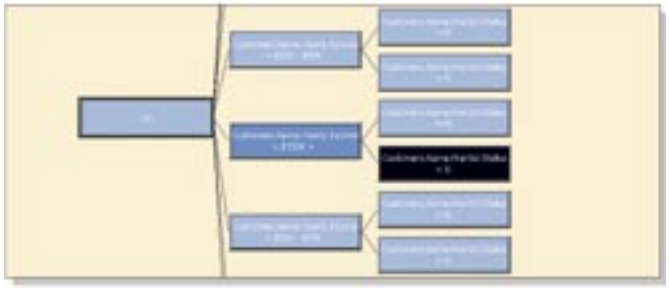
Afbeelding 8. Een overzicht van de belangrijkste SQL Server 2005-componenten.



Afbeelding 7. Voorbeeldvisualisatie van Key Performance Indicator in Visual Studio



Afbeelding 9. Analysis Services 2005 biedt een brede waaier aan opslag- en updatemogelijkheden



Afbeelding 10. Een deeltje van een Analysis Services-beslissingsboom: ongehuwde rijke klanten zijn voornamelijk in een 'Silver Card' geïnteresseerd.

systeem, waarvoor de gebruiker zelf niets hoeft te doen. Wanneer dit een andere database is, heeft de gebruiker de keuze tussen zelf een notificatiesysteem implementeren of gebruikmaken van polling. Ook hier zijn er allerlei opties instelbaar over hoe frequent een dergelijke update moet gebeuren. Het resultaat is dat er niet drie, maar een heel scala aan opslagopties beschikbaar is; zie afbeelding 9.

Het is belangrijk hierbij op te merken dat de beperkingen op MOLAP-kubussen en dimensies weggewerkt zijn via de 'spill to disk' aanpak; in Analysis Server 2000 zijn deze niet bruikbaar op erg grote hoeveelheden gegevens.

Datamining: zoeken in de hooiberg

Wat we tot nu toe van Analysis Services besproken hebben, is handig om de invloed van bepaalde variabelen (zoals product of regio) op bepaalde maatstaven (zoals verkoopcijfers of winstmarges) te onderzoeken. Maar er zijn in een kubus vaak vele variabelen, en voor sommige taken is het moeilijk om handmatig de juiste analyses te maken. Weet iemand bijvoorbeeld welke de belangrijkste factoren zijn die het verschil maken tussen goede en minder goede klanten? Weet iemand welke producten er vaak samen aangekocht worden? Door alle mogelijke combinaties tegen elkaar af te zetten, kun je uit je gegevens antwoorden op dit soort vragen afleiden. Maar gegeven de grote hoeveelheid variabelen is dit in de praktijk onmogelijk. Daarom beperken analisten zich vaak tot het verifiëren van voor hen bekende afhankelijkheden. Het is echter ook mogelijk om de computer zelf vele combinaties van variabelen te laten analyseren en de statistisch relevante correlaties te combineren in een model. Dit noemt men datamining.

In Analysis Services 2000 waren er al twee datamining-algoritmen ingebouwd: een classificatiealgoritme en een clusteringalgoritme. Veronderstel dat je als supermarkt graag wilt weten welke klanten een bepaald soort voordeelkaart willen aanschaffen, zodat je bijvoorbeeld specifieke marketingcampagnes op deze klanten kunt richten). Als je historische gegevens over je klanten verzameld hebt, kan Analysis Services met zijn classificatiealgoritme hieruit een beslissingsboom afleiden. Afbeelding 10 toont een voorbeeld van zo'n beslissingsboom.

Analysis Services 2005 breidt de collectie van datamining-technieken uit Analysis Services 2000 drastisch uit. Zo is er een algoritme toegevoegd om associatieregels (ook wel basket analysis genaamd) mee af te leiden. Dit zijn regels die bijvoorbeeld correlaties in aankoopgedrag kunnen detecteren, zoals het beroemde voorbeeld van Amazon (www.amazon.com). Hierbij wordt bij elk boek op de webpagina getoond in welke andere boeken een gebruiker waarschijnlijk ook geïnteresseerd zal zijn, gebaseerd op historisch aankoopgedrag van andere klanten. Andere nieuwe algoritmen zijn er bijvoorbeeld voor regressie (voorspellen van continue waarden) of tijdreeksen (voorspellen van een volgende waarde in een sequentie van waarden).

Vele mogelijkheden

Analyse en rapportage worden steeds belangrijkere facetten in de moderne bedrijfscultuur. OLAP-databases voorzien in een gebruiksvriendelijkere en efficiëntere toegang tot deze geaggregeerde gegevens. Microsoft Analysis Services 2005 vormt een gebruiksvriendelijke en krachtige OLAP-server. De asynchrone ontwikkeling die in Visual Studio gebeurt, en die kan steunen op wizards zoals IntelliCube, maakt het ontwikkelingsproces gemakkelijker. Wanneer het eindresultaat hiervan een XML-document is, betekent dit dat (onderdelen van) de OLAP-kubussen ook met programmering gegenereerd kunnen worden. Daarnaast kunnen met perspectives en vertalingen grote kubussen voor verschillende gebruikers op maat worden gemaakt. Bovendien kan met proactive caching de voordelen van ROLAP (up-to-date gegevens) gecombineerd worden met de voordelen van MOLAP (performance).Tot slot bevat Microsoft Analysis Services 2005 nog tal van extra kenmerken, zoals meer datamining-mogelijkheden, complexere relationele basisgegevens en KPI's. Andere kenmerken waar we in dit artikel niet dieper op zijn ingegaan zijn de debug-mogelijkheden voor MDX-expressies, triggers, CLR stored procedures, caching van calculated cells en calculated members. Kortom, een heel arsenaal aan mogelijkheden om analyse of rapportagetoepassingen efficiënt te ontwikkelen en deployen. Ook belangrijk om te vermelden is dat de aanpak die in dit artikel geschets wordt, deel uitmaakt van UDM, een data-model dat het mogelijk maakt om allerhande databronnen op een uniforme manier aan te spreken. Meer informatie over UDM vindt u op de MSDN-site van Microsoft.

Dr. Nico Jacobs is trainer en consultant bij U2U (www.u2u.net), waar hij zich voornamelijk in Microsoft SQL Server en ADO.NET specialiseert. Zijn e-mailadres is nico@u2u.be.

Wim Uyttersprot is director van U2U (www.u2u.net), Microsoft Partner voor Learning Solutions en gespecialiseerd in .NET Training en Consultancy. Zijn e-mailadres is wim@u2u.be.

Nuttige internetadressen

- http://msdn.microsoft.com/sql/2005/2005articles/default.aspx
- http://msdn.microsoft.com/SQL/2005/2005Webcasts/
- http://www.microsoft.com/Sql/reporting/
- http://www.microsoft.com/BI
- http://www.olapreport.com/market.htm
- http://www.sqlpass.org/