Microsoft en BI: 'one-stop shopping'?

Big Green maakt de cirkel rond

Stefan van Duin

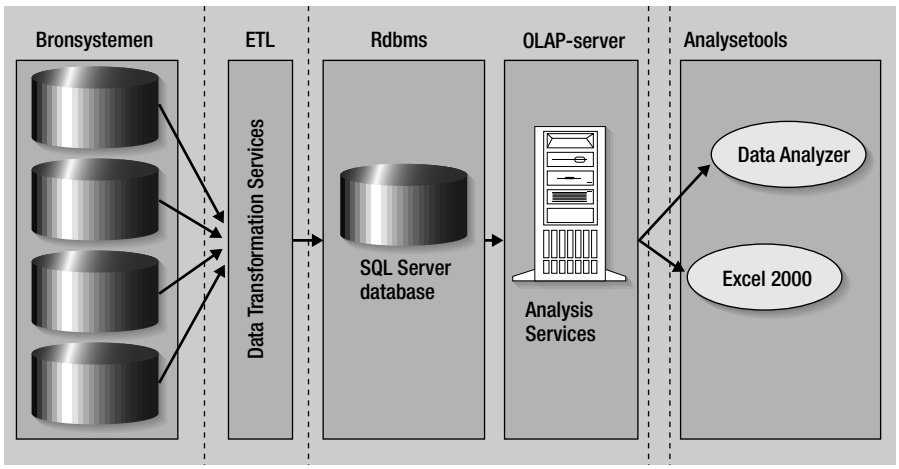
Met de introductie van Data Analyzer, een OLAP-frontend-tool in de Office XP-reeks, heeft Microsoft de cirkel rond. Op alle onderdelen van een business intelligence-architectuur, extractie, transformatie, opslag, analyse en presentatie, kunnen producten van Microsoft worden ingezet. Stefan van Duin beoordeelt of de geboden oplossingen aldaar voldoen. In welke situaties is Big Green in staat een complete toepassing te leveren?

Sinds MS SQL Server 7 bestaat, begeeft Microsoft zich serieus op het terrein van business intelligence. Vervolgens is zij er enerzijds in geslaagd de markt haar de facto standaards op te leggen (OLEDB for OLAP), anderzijds probeert zij nog steeds het imago van de database voor kleine implementaties van zich af te schudden. Microsoft levert inmiddels services voor vrijwel het hele scala aan componenten in een business intelligence-omgeving:
de SQL Server-database voor opslag;

- Data Transformation Services (DTS) voor de extractie, transformatie en het laden (ETL) van brongegevens naar het datawarehouse;
- Analysis Services (voorheen OLAP server) voor de multidimensionele analyse van gegevens;
- Data Analyzer (DA) voor de presentatie van de multidimensionele informatie.

DATABASE

Traditioneel wordt het SQL Server-rdbms vooral gepositioneerd als goedkope, eenvoudig te gebruiken database voor vooral de wat kleinere applicaties. Om het systeem ook op de kaart te zetten als serieuze speler in de markt van grote toepassingen, heeft Microsoft flink geïnvesteerd in de technologie. Dit heeft geresulteerd in TPC-benchmark-resultaten die indruk maken[1];



Figuur 1: Microsofts architectuur voor business intelligence.

SQL Server is tegenwoordig in de transactieverwerkende applicaties een van de best scorende producten. Met initiatieven als het T3-prototype, waarbij in samenwerking met EMC en Unisys een pilot-datawarehouse van 1,2 TB is opgenomen in SQL Server 2000 Analysis Services[2], wil Redmond de mond snoeren van een ieder

Microsoft zoekt nadrukkelijk de samenwerking met toolleveranciers, en claimt geenszins een totaaloplossing aan te bieden

die beweert dat het geen grote databases aankan.

Deze resultaten worden bereikt door gebruik te maken van de nieuwe technologie van *distributed partitioned view*. Een tabel is horizontaal gepartitioneerd over meerdere servers of clusters en is door middel van een view door applicaties te benaderen als ware het één fysieke tabel. De clusters zelf worden onafhankelijk van elkaar beheerd. De technologie voor het ondersteunen -en vooral beheren- van clusters is echter nog verre van volmaakt. De opvolger van SQL Server 2000 (codenaam Yukon) zal daarom verder gaan op de weg van *very large databases* (VLDB). Microsoft belooft in Yukon het integrale beheer van de clusters te regelen.

Big Green richt zich bewust op het Windows-platform en kapitaliseert daarmee de voordelen van concentratie op één omgeving. SQL Server is hard op weg de

meest gebruikte databaseserver te worden op het Windows-platform. Unix- en mainframe-omgevingen zijn echter nog steeds gemeengoed in veel organisaties met grote systemen en grote datawarehouses. Daar zal het voor Microsoft dan ook moeilijk zijn een positie te veroveren.

DATA TRANSFORMATION SERVICES

Microsoft Data Transformation Services (DTS) is de component die het datawarehouse kan voeden. Enerzijds van bron naar warehouse, anderzijds het genereren van OLAP-kubussen. Daartoe levert het een grafische interface, waarmee verschillende logische eenheden van taken (*packages*) kunnen worden ontwikkeld en in een workflow geplaatst. DTS ondersteunt onder meer:

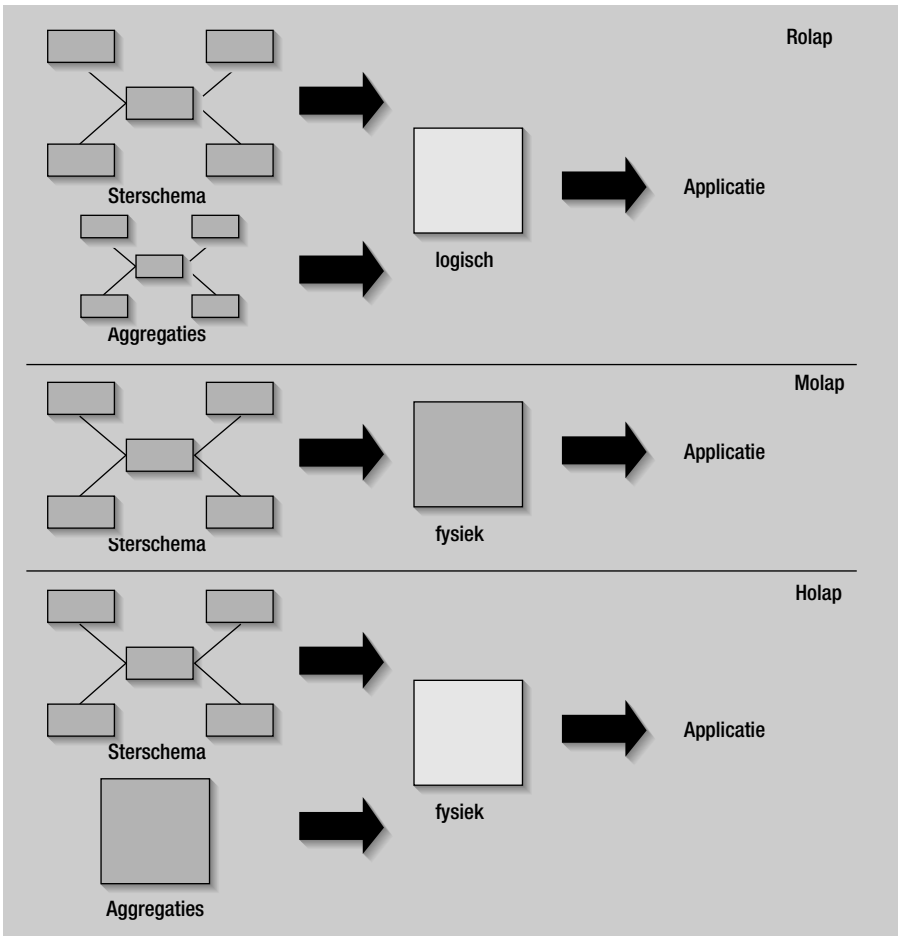
- importeren en exporteren van data naar externe bronnen;
- transformeren van gegevens;
- kopiëren van database-objecten;
- zenden en ontvangen van (e-mail-) berichten;
- uitvoeren van een set van SQL-statements of ActiveX script.

DTS kan verschillende doel- of brondatabases benaderen via OLEDB. De meest gangbare databases, zoals Oracle en DB2, zijn te benaderen via meegeleverde OLEDB-drivers, of via OLEDB for ODBC.

Met DTS is het ongetwijfeld mogelijk de periodieke populatie van een datawarehouse uit bronsystemen te realiseren. Het biedt een goede omgeving om transformatietaken in te delen (*scheduling>P>) en de uitvoering (*execution) ervan te volgen. Een goede kennis van ActiveX en/of SQL blijkt echter noodzakelijk, omdat de geboden tools zich beperken tot technische functionaliteit, waardoor het ontwikkelen van scripts noodzakelijk is.

Als vervanging van een high-end ETL-tool kan DTS dan ook niet dienen. Zo zal de ontwerper/ontwikkelaar tevergeefs zoeken naar wizards voor slowly changing dimensions, snowflake dimensions en fact-table populatie. Tevens is het niet direct mogelijk de output van de ene stap als input van de andere stap te gebruiken (*piping*), waardoor de wat ingewikkeldere dataflows in een ETL-proces niet te realiseren zijn.

Doordat de brondatabases beperkt zijn tot OLEDB ODBC of flat-file, kan het ontsluiten van echte legacy-systemen een potentieel probleem zijn. Denk hierbij aan mainframebestanden met complexere bestandsstructuren, zoals meerdere record layouts in één bestand of meerdere waarden met de OCCURS-clausule in Cobol, of de EBCDIC-ascii-conversie met packed-velden[3].



FIGUUR 2: MOLAP, ROLAP EN HOLAP.

DATA ANALYSIS SERVICES

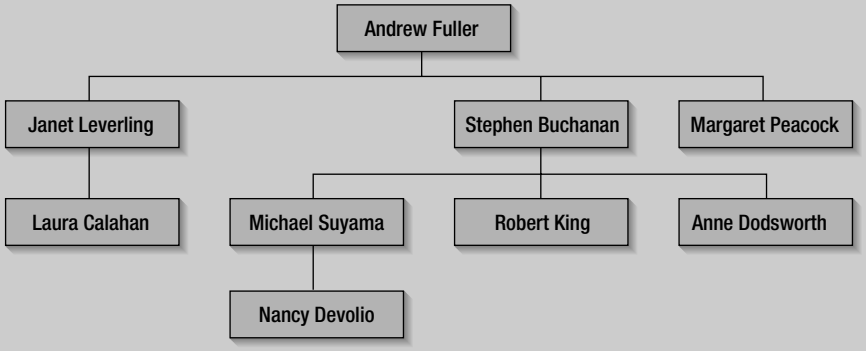
Microsoft Data Analysis Services (AS) is de servercomponent, die de relationele data voorziet van een multidimensionele blik. Het uitgangspunt van AS is een relationele database, dat overigens niet per se een SQL Server-database hoeft te zijn. Uitgangspunt is gegevensopslag in een multidimensioneel datamodel, zoals een ster- of sneeuwvlokschema.

In Analysis Services worden één of meer kubusdefinities vastgelegd, op basis van het datawarehouse. In een dergelijke kubusdefinitie is onder meer vastgelegd hoe fact- en dimension data zijn opgebouwd. Zo valt aan te geven of een dimension is opgebouwd volgens de principes van hetzij een sterschema of à la sneeuwvlok (genormaliseerd) of door middel parent/child-relaties.

Zodra een kubus is gedefinieerd, kan men op drie manieren overgaan tot het voeden met data:

- relationeel OLAP (Rolap): gegevens en aggregaties zijn en blijven opgeslagen in de relationale database;
- multidimensioneel OLAP (Molap): gegevens en aggregaties worden opgeslagen in een (door AS te genereren) multidimensionele database;
- hybride OLAP (Holap): gegevens blijven opgeslagen in de relationele database, aggregaties worden multidimensioneel opgeslagen.

Emp_ID	Name	Mgr)–ID
1	Andrew Fuller	(null)
2	Janet Leverling	1
3	Stephen Buchanan	1
4	Margaret Peacock	1
5	Laura Calahan	2
6	Michael Suyama	3
7	Robert King	3
8	Anne Dodsworth	3
9	Nancy Davolio	6



FIGUUR 3: PARENT/CHILD-DIMENSIES.

Rolap is bruikbaar voor grote hoeveelheden gegevens die minder frequent worden opgevraagd, of realtime OLAP, omdat geen extra transformatiestap nodig is. Holap is typisch geschikt voor situaties waarin vaak aggregaties worden geraadpleegd, terwijl de achterliggende detailgegevens zeer veel rijen bevatten.

De gegenereerde kubussen kunnen overigens gepartitioneerd worden opgeslagen, wat inhoudt dat delen (partities) van de kubus fysiek op verschillende plaatsen opgeslagen zijn en dat de opslageigenschappen per partitie geoptimaliseerd kunnen worden. Zo is de kubus per periode te verdelen en kunnen de oudere, minder gebruikte perioden op een andere locatie staan dan de veelgebruikte, meest recente periode. Bovendien is een query, die data uit meerdere (fysiek gescheiden) partities omvat, parallel uit te voeren.

AS heeft duidelijk gekeken naar de best-of-breed van andere OLAP-leveranciers. Het komt tegemoet aan de gangbare uitdagingen in de multi-dimensionele wereld, waarvan we de meest in het oog springende beschrijven:

- parent/child-dimensies;
- ragged dimensies;
- conformed dimensies;
- multi-pass query's;
- write-back.

Parent/child

In parent/child-dimensies is een hiërarchische structuur vastgelegd tussen de leden van de dimensies. Een voorbeeld hiervan is een werknemer/manager-relatie, waarin in een werknemer-regel een verwijzing als manager is opgenomen naar een andere werknemer-regel. Deze verwijzing is recursief, omdat de betrokken manager op zijn beurt een manager heeft. De dimensie heet *unbalanced* als de diepte van de hiërarchie niet overal gelijkwaardig is. Dit type dimensie komt ook vaak voor in organisatiestructuren (afdelingen met sub-afdelingen).

Ragged dimensies

In een aantal gevallen worden niveaus in sommige lijnen van de parent/child-hiërarchie overgeslagen of heeft een 'kind' meerdere 'ouders', die ieder een vastgelegd deel van het kind claimen. *Ragged* noemt men deze dimensies.

Een voorbeeld vinden we in de hiërarchie afdeling-team-persoon. In sommige gevallen behoort een persoon niet tot een team of is een persoon op meerdere afdelingen werkzaam.

Conformed dimensions

Dimensies worden vaak in meerdere kubussen hergebruikt. De product-dimensie is bijvoorbeeld te gebruiken in zowel de verkoop- als de voorraad-kubus. Dit type dimensie hoeft slechts eenmaal gedefinieerd te worden, wat hergebruik van de definitie in de andere kubussen mogelijk maakt.

Multi-pass query's

Het komt vaak voor dat een typische OLAP-vraag niet in één query valt te beantwoorden. Als meetwaarden uit twee feitentabellen met een gezamenlijke

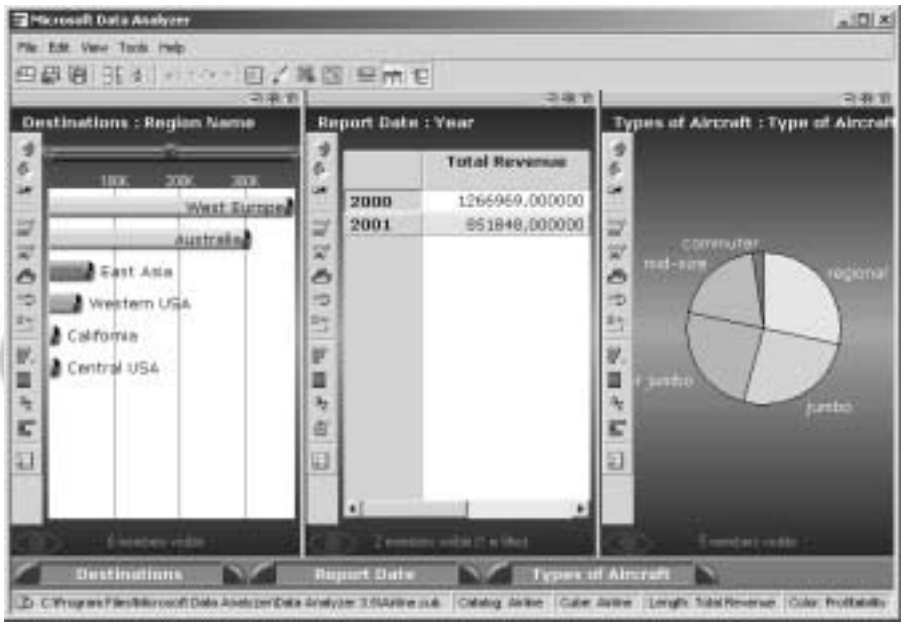
BO, Cognos en ProClarity

Analysis Services ondersteunt de OLAP-specifieke vereisten van OLE DB 2.0. Daarmee is het beschikbaar als bron van multidimensionale data voor leveranciers van OLAP-producten. De syntax om AS te benaderen, wordt gevormd door multidimensional expressions (MDX). MDX biedt een rijke bibliotheek aan analytische functionaliteit en kan worden aangevuld met gebruikersgedefinieerde functies.
Business Objects, een van de marktleiders in BI-tools, biedt via MDX Connect toegang tot AS' OLAP-functionaliteit. Met MDX Connect kan de OLAP-bron, met alle dimensie- en meetwaardedefinities, direct worden benaderd, zonder de noodzaak een aparte universe ervoor te definiëren. De belangrijkste OLAP-functionaliteit is vervolgens direct in de bekende BO-gebruikersinterface beschikbaar.
BO's grote concurrent Cognos Powerplay kan -naast uiteraard zijn eigen kubusopslagstructuur- AS als bron raadplegen. De gehele dimensiestructuur en meetwaardedefinities zijn direct beschikbaar, zonder de noodzaak ze opnieuw te definiëren in een Cognos-specifieke structuur. De rijke OLAP-functionaliteit van Powerplay is dan inzetbaar op de AS-data.
ProClarity is het best geïntegreerd met de functionaliteit van AS. Dit komt doordat Microsoft de ontwikkeling van dit product actief ondersteunt. Vrijwel alle in SQL Server ingebouwde OLAP-functionaliteit is daardoor beschikbaar in ProClarity's gebruikersinterface.

dimensie met elkaar worden gecombineerd, zijn daarvoor doorgaans twee query's nodig. Denk aan het opvragen van de voorraadstand en het totaal van de verkoopcijfers. AS ondersteunt dit probleem door de query automatisch -op de server- op te delen in meerdere query's en de resultaten van de deel-query's te combineren tot één antwoord.

Write-back

Ten slotte is de mogelijkheid tot het schrijven van dimensiewaarden in de kubus door eindgebruikers een bruikbare feature, hoewel nog weinig van de bekende OLAP-leveranciers deze optie ondersteunen. Analysis Services ondersteunt OLEDB for OLAP en is daarmee toegankelijk voor de meeste gangbare OLAP-frontends (zie ook kader *BO, Cognos en ProClarity*). Anderzijds kan het dienen als OLAP-server voor niet-SQL Server-databases, zoals Oracle en DB2.



FIGUUR 4: DATA ANALYZER.

DATA ANALYZER EN EXCEL 2000

Hoewel Microsoft tot begin vorig jaar verkondigde niet te verwachten een OLAP-tool uit te brengen, lanceerde zij afgelopen najaar Data Analyzer: een telg in de Office XP-familie, waarmee multidimensionale databases geëxploreerd kunnen worden. Gepresenteerd als Office-product, draait het pakket in feite standalone.

Excel werd gepresenteerd als hét BI-frontend-tool van Microsoft. De grote bekendheid van dit product, alsmede de mogelijkheden om eigen analyses en formules toe te passen in OLAP-overzichten maken Excel inderdaad tot een populaire frontend. Op het gebied van grafische presentatie en navigatie door een kubus laat Excel echter nog wel te wensen over. Het gat wordt voor een groot deel opgevuld door DA.

DA is een laagdrempelig analysetool, dat gegevens uit elke OLEDB for OLAP compatible bron (dus ook niet-MS bronnen!) kan ontsluiten. De presentatie van de informatie beperkt zich tot staafdiagrammen, taartdiagrammen en raster, waarbinnen eenvoudig genavigeerd kan worden in de dimensiehiërarchieën.

Enkele verrassende features zijn het BusinessCenter, waarmee met natuurlijke zinnen eindgebruikers door de meest gangbare vragen worden geloodst ("How did year-to-date unit sales compare to the same period last year?"), en de optie van *find-similar*, waarmee bij gevonden afwijkingen vergelijkbare voorkomens van de afwijking opgespoord kunnen worden.

Uiteraard is Data Analyzer verregaand geïntegreerd met andere Office-pakketten. Een greep uit de acties die met een overzicht kunnen worden uitgevoerd:

- verzenden via e-mail;
- publiceren naar html;
- exporteren naar Powerpoint;
- analyseren met Excel.

Verder biedt DA een ActiveX-control, om de overzichten te kunnen incorporeren in

Datamining

Naast de OLAP-functionaliteit biedt Analysis Services twee algoritmen voor datamining.

- *Beslissingsbomen*: een techniek gebaseerd op classificatie, waarbij een eigenschap (bijvoorbeeld respons op een mailing) wordt gevisualiseerd in een boomstructuur. De techniek wordt gebruikt om te achterhalen welke eigenschappen voorspellende waarde hebben voor een andere eigenschap. Elke tak van de boom representeert dan de gevonden classificatie, met het percentage dat daarbij hoort.
- *Clustering*: een techniek waarmee (verborgen) groepen in data gevonden kunnen worden. Een typisch voorbeeld van een toepassing van deze techniek is het vinden van segmenten in een klantdatabase. Bij het maken van een clustermodel kan worden opgegeven naar hoeveel groepen het algoritme moet zoeken. Als resultaat geeft het algoritme de eigenschappen van elke groep terug.

Verder bestaat de mogelijkheid aanvullende algoritmen toe te voegen door middel van OLEDB for datamining. De meegeleverde algoritmen zijn weliswaar niet erg geavanceerd, maar wel inzichtelijk en maken dataminingtechnieken eenvoudig beschikbaar voor veel applicaties.
Microsoft levert geen gebruikersinterface voor de datamining-algoritmen; daarvoor moet men te rade gaan bij andere leveranciers.

dashboard-applicaties, alsmede een API om maatwerk-applicaties te ondersteunen.

VERDERE ONTWIKKELING

In dit artikel is de invulling van een totale omgeving voor business intelligence en datawarehousing met alleen Microsoft-producten als uitgangspunt genomen. Dat blijkt in veel gevallen niet echt realistisch. De basis, namelijk de SQL Server-database met Analysis Services, biedt een krachtig platform, waarmee functioneel rijke, schaalbare applicaties gerealiseerd kunnen worden. Data Transformation Services moet gezien worden als een zeer eenvoudig tool, waarmee slechts recht toe recht aan transformaties uit bronsystemen gerealiseerd kunnen worden. Data Analyzer ten slotte biedt een goedkoop, eenvoudig tool om de gegevens te kunnen bekijken en applicaties mee te ontwikkelen.

Microsoft zelf zoekt overigens nadrukkelijk de samenwerking met toolleveranciers, en claimt geenszins aanbieder van een totaaloplossing te zijn. Maar als Big Green de lijn in de ontwikkeling van BI/DW-producten doorzet, maait hij wel steeds meer gras voor de voeten van de andere toolleveranciers weg.

Op het gebied van OLAP-servers is Microsoft daarin al geslaagd. Vrijwel alle bekende spelers in die markt hebben terrein moeten prijsgeven. Het zal geen verrassing zijn dat in Redmond wordt gewerkt aan de verdere ontwikkeling van zowel DA als DTS. Microsoft zal binnen een paar jaar voor alle schakels van een BI-architectuur een volwassen product aanbieden. ●

Noten en bronnen:

1. TPC is een onafhankelijke organisatie die performancevergelijkingen tussen databases mogelijk maakt. TPC-C is de benchmark voor transactieverwerking. SQL Server 2000 bezet de eerste drie plaatsen. TPC-H is de benchmark voor decision support-omgevingen en is onderverdeeld in 100, 300, 1000 en 3000 Gb omgevingen. Bij 100 en 300 Gb scoort Microsoft in de hoogste regionen, in de categorie 1000 en 3000 gigabyte zijn geen benchmarks voor Microsoft beschikbaar. Zie www.tpc.org/
2. Zie www.Microsoft.com/sql/techinfo/BI/terabytecube.asp
3. Voor het ontsluiten van legacy-systemen is overigens een ander product beschikbaar: Microsoft Host Integration Server.

SQL Server 2000: Ready for Prime Time? J. Rubin, Research note Gartner Group, 8 maart 2001.
Microsoft Corp. SQL Server 2000 Analysis Services. Datapro Research, 4 oktober 2000.
Microsoft SQL Server 2000 as a "Dimensionally Friendly System". Microsoft Corp.
Microsoft Data Analyzer product guide. Microsoft Corp., oktober 2001.
Diverse artikelen op www.microsoft.com
Diverse artikelen op www.sqlmag.com

Stefan van Duin (stefan.van.duin@cgey.nl) is managing consultant bij Cap Gemini Ernst & Young en gespecialiseerd in business intelligence en datawarehousing.

(vervolg van pagina 9)

"Gelukkig zijn er nog steeds bedrijven die het wel aandurven, en dat zijn niet de kleinsten. Wij opereren in een niche. Op een shortlist komen we dus al snel. Je hebt enerzijds de namen die op iedere shortlist prijken, omdat ze pretenderen alles te kunnen. Als je bij IBM, Microsoft of Oracle aanklopt, krijg je standaard de toezegging dat ze de gevraagde oplossing kunnen leveren. Daarnaast heb je de grote consultants en tot slot gespecialiseerde bedrijven, die net als wij, deeloplossingen leveren. Klanten die met ons in zee gaan, doen dat op grond van de technische kwaliteit die wij leveren."

De producten die Minerva ontwikkelt, kennen een lange draagtijd. Een bewuste keuze, legt De Vleeschauwer uit. "Wij ontwikkelen samen met onze klanten. Voor de realisatie van DBA hebben we een *advisory board* van twaalf bedrijven in het leven geroepen. We hebben een prototype gemaakt en aan de mensen uit die advisory board gepresenteerd. Met hun input zijn we verder gegaan. Op dit moment wordt het getest door de leden van de board. Wij hechten zeer aan die zorgvuldige aanpak. Het perfecte product bestaat niet. Net zo min als de perfecte vrouw. Daarom is die lange incubatietijd zo belangrijk. Technische hoogstandjes zonder praktisch nut worden er in de proefperiode uitgefilterd."

STRATEGISCHE BESLISSING

De aanpak van Minerva strookt niet altijd met de CRM-behoefte van de klant. "Datatransformatie is het meest onderschatte en delicate onderdeel van een CRM-traject. Je hebt het topje van de ijsberg, dat bestaat uit sexy grafieken en analyses met business objects. Dat spreekt tot de verbeelding. Wat je moet doen om die gegevens in een gestructureerd formaat te krijgen, interesseert eigenlijk niemand. Dat is zonde, want de kosten en baten van een CRM-project worden vooral door dat traject beïnvloed. Een verzekeringsmaatschappij heeft al zijn gegevens in een eigen systeem georganiseerd volgens eigen denkbeelden. Voordat je met een CRM-applicatie aan de slag kunt, moet je eerst die vertaalslag maken. In de praktijk zie je dat bedrijven dat met maatwerk proberen op te lossen. Voor een klant is de aanschaf van onze software een strategische beslissing. De invloed van die stap blijft vele jaren voelbaar. Het is een misvatting te denken dat een softwareproduct een korte levensduur heeft. Je moet een product implementeren, aanpassen aan je eigen omgeving en mensen opleiden. De vervangingskosten zijn hoog."

"Wat het voor ons extra lastig maakt", vervolgt de Minerva-CEO, "is dat de businessafdeling niet met onze tools aan de slag kan. De gemiddelde IT-manager vindt een onderwerp als datatransformatie maar saai. Het is in zijn ogen een noodzakelijk kwaad. Als je het puur bedrijfsmatig bekijkt, moet je het product kopen, want het verdient zich in no time terug. Maar het is heel moeilijk een potentiële klant van die boodschap te overtuigen. Ik denk dat veel mislukte implementaties zijn terug te voeren op dit probleem. Men doet het liever twee of drie keer dan in één keer goed, lijkt het wel." ●

Peter Steeman is freelance journalist.