SQL Server 2008 is nieuwe stap op weg naar 'BI for the masses'

De democratisering van data mining

Robbert Hoeffnagel

Microsoft heeft sinds de overname van ProClarity flinke stappen vooruit gezet als het om Business Intelligence gaat. Onlangs verzorgde Rafal Luckawiecki van The Botticelli Project een presentatie over de Analysis Services die horen bij de huidige SQL Server 2005-omgeving en de nieuwe 2008-editie. Wat hem met name aanspreekt is de kwaliteit van de algoritmen die Microsoft standaard meelevert. Bovendien lijkt het concern er in te zijn geslaagd om BI toegankelijk te maken voor iedere medewerker die enigszins met Excel uit de voeten kan.

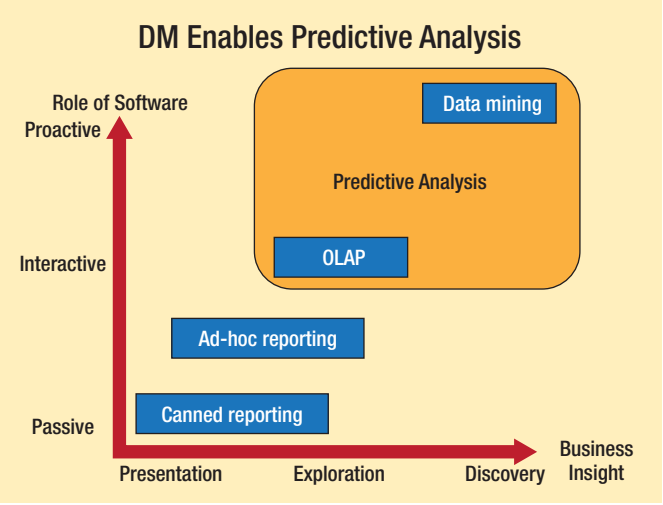
Wie Project Botticelli even niet kan plaatsen, hoeft zich geen zorgen te maken. Nee, het is niet de codenaam van een ontwikkelproject van Microsoft voor de opvolger van SQL Server 2008 en ook niet voor een nieuwe generatie BI-tools. Project Botticelli is daarentegen wel de naam van een kleine Engelse consultancyfirma, waar men zich bezig houdt met een breed scala aan op ICT gerichte adviespraktijken.
Neem Rafal Lukawiecki. Hij is inmiddels uitgegroeid tot de BI-specialist van de firma, een positie die hij op een opmerkelijke manier heeft weten te bereiken: via zijn kennis van cryptografie. Wie met encryptie aan de slag gaat, wil met behulp van onder andere kansberekening, statistiek en algebra data verbergen. De BI-specialist probeert met precies dezelfde hulpmiddelen juist data en de patronen daartussen zichtbaar te maken. "Opposites attract", wilde Lukawiecki maar zeggen.

Nader onderzoek

Lukawiecki was onlangs op uitnodiging van Microsoft in Nederland. Hij verzorgde voor het concern een TechNet Expressessie over BI. Tijdens de lezing ging hij enerzijds in op de basisprincipes van Business Intelligence, terwijl hij anderzijds de producten van Microsoft op dit gebied aan een nader onderzoek onderwierp.
Een met ruim driehonderd mensen gevulde zaal geeft wel aan dat Microsoft ook in de BI-wereld inmiddels flink van zich doet spreken. Of misschien is het beter om te stellen dat bedrijven die al grotendeels op Microsoft-technologie zijn gestandaardiseerd ook belangstelling hebben voor de producten die het concern voor de BI-markt ontwikkelt. Zie het concern dan ook niet zozeer als een partij die nu direct gespecialiseerde aanbieders als SPSS of SAS Institute het vuur na aan de schenen zal leggen; de gespecialeerde business analist die het liefst zijn eigen en vaak zeer complexe query's en modellen bouwt behoort niet tot de eerste doelgroep waar Microsoft aan denkt.
Het concern mikt met zijn BI-tools veel meer op een ander publiek: denk aan medewerkers die redelijk met Excel uit de voeten kunnen en die vanuit hun functie behoefte hebben aan een snelle analyse van grotere hoeveelheden gegevens. Veel meer dus business managers dan BI-specialisten. Vandaar ook dat Lukawiecki in zijn presentatie een duidelijk onderscheid maakte tussen data mining, knowledge discovery en Business Intelligence. De laatste vormt de basis van de piramide, met knowledge discovery in het midden en data mining als punt bovenaan de driehoek.

Democratisering

Dat lijkt wellicht een wat academisch onderscheid, maar dat is het toch niet, meent Lukawiecki. Anders komen we namelijk in de knoop met het benoemen van Microsoft's BI-aanpak. Het is volgens hem een reeks van technologieën voor het analyseren van gegevens en het ontdekken van al of niet verborgen patronen daartussen en dan vooral bedoeld voor een breed publiek. 'BI for the masses', zeg maar. Dat stelt nogal wat eisen aan de gebruikte technologie en de manier waarop de brede doelgroep met de combinatie van statistiek, kansberekening en databasetechnologie moet kunnen omgaan.
Hoewel Lukawiecki die term niet gebruikte, maakt hij in zijn presentatie wel duidelijk dat BI in feite de democratisering van data mining is. Microsoft's aanpak is daar een goed voorbeeld van en geeft aan deze trend bovendien een stevige impuls. Microsoft's BI-tools zijn in de visie van Lukawiecki gericht op de eindgebruiker – een afdelingshoofd, een ondernemer – waar data mining eerst en vooral een IT-technologie is die de power user – Lukawiecki sprak liever van 'advanced user' – tot doelgroep heeft. Ook hanteert hij een term als 'knowledge keeper', om aan te geven dat Microsoft mikt op de kernfunctionarissen binnen een organisatie.
In zijn bespreking van de aanpak van Microsoft kwam dat verschil ook duidelijk naar voren. Het concern richt zich nadrukkelijk op de gebruiker die hooguit een bescheiden kennis van



Afbeelding 1: De positie van predictive analysis.

databases heeft. De tools hebben wel wat weg van T-SQL en Management Studio, waarbij opvalt dat Microsoft al sinds de komst van SQL Server 2005 een aantal DM-tools standaard beschikbaar heeft. Met de komst van SQL Server 2008 krijgt de BI-kant een stevige upgrade. Daarnaast gaat het concern nog een stap verder met wat Lukawiecki voor het gemak maar 'DM easy' noemde en dat gericht is op de gebruiker die redelijk goed in Excel is ingevoerd en die vanuit die omgeving BI-functionaliteit wil benutten.

'Predictive analysis'

Microsoft richt zich nadrukkelijk op het segment dat predictive analysis wordt genoemd (zie afbeelding 1). Hier gebeuren de laatste jaren 'de leuke dingen', stelt Lukawiecki vast, en aan die trend is nog lang geen einde gekomen, verwacht hij.
Zo heeft hij hoge verwachtingen van wat Google op dit gebied zal gaan betekenen. Hoewel hij geen hoge pet op heeft van de aanpak die dit concern volgt ten aanzien van security of privacy, ziet hij in Google wel dé partij die BI als een service gaat aanbieden. De functionaliteit die nu al her en der in de advertentieprogramma's, de zoekhistorie van individuele gebruikers en bijvoorbeeld de rss-readers wordt ingebouwd, vormen in zijn ogen slechts de eerste, voorzichtige stapjes die kant op.
Overigens verwacht Luckawiecki niet dat Amazon op dit gebied actief zal worden. Dat is opmerkelijk, aangezien dit concern inmiddels al wel volop bezig is met wat wel 'cloud computing' wordt genoemd evenals bijvoorbeeld database services. Verwerkingscapaciteit, opslagruimte en bijvoorbeeld functionaliteit op het gebied van datamanagement kunnen hierbij simpelweg op basis van 'pay as you use' worden gehuurd.

Winstgevende klanten

Predictive analysis vormt het segment waar BI en data mining van een vrij abstract en high-level hulpmiddel overgaan in voorzieningen die direct aan de business raken. Operationele tools die dagelijks bij het besturen van een organisatie kunnen worden gebruikt. Geen rapporten meer over hoe de situatie drie weken geleden was, maar real-time en – vaker waarschijnlijk – 'near real-time' verwerking en analyse van grotere hoeveelheden gegevens.
Lukawiecki liet zien hoe een ondernemer of afdelingshoofd zelf met BI-tools van Microsoft op relatief eenvoudige wijze kan vaststellen wie nu eigenlijk zijn meest winstgevende klanten zijn, met welk verloop hij onder zijn afnemers rekening moet houden of hoe hij beter inzicht kan verkrijgen in de verkopen of voorraadposities die voor de komende tijd verwacht mogen worden.
Neem die winstgevendheid van klanten. Het clustering-algoritme kan helpen met het groeperen van gegevens zodat klanten gesegmenteerd of geclassificeerd kunnen worden. Aan de hand van het door Microsoft meegeleverde decision tree-algoritme kunnen relaties tussen winst en klant worden gevonden. Aan de hand van association rules kunnen bovendien voorkeuren van afnemers worden achterhaald. En met een techniek als sequence clustering is het gedrag van klanten te bestuderen; alles bedoeld om uiteindelijk de winstgevendheid van potentiële nieuwe afnemers te kunnen voorspellen. Daarbij kan het helpen om alle data over bijvoorbeeld verkoop of voorraad met het time series-algoritme te structureren, waarna vervolgens met een algoritme als time series regression and prediction een voorspelling van toekomstige verkopen kan worden gemaakt.

Detecteren van fraude

Zo liet Lukawiecki meer voorbeelden zien. Wie met algoritmes als decision tree, naive Bayes (voor het classificeren van bijvoorbeeld teksten), clustering en neural network in de weer gaat, kan bijvoorbeeld onderzoeken waarom klanten op bepaalde marketingcampagnes wel reageren en op andere niet. Ook een algoritme als lift charts kan in dit kader zijn nut bewijzen.
Interessant was ook een onderwerp waar momenteel sowieso veel BI-aanbieders zich op richten: het detecteren van fraude. Met algoritmes als decision tree, clustering en neural network kan een business manager een model bouwen dat risico's in kaart brengt voor bestaande klanten en transacties. Aan de hand

Microsoft Performance Point Server

Is het business intelligence? Of toch maar performance management? Welke sticker we ook op Microsoft's Performance Point Server plakken, feit blijft dat het concern hiermee een opmerkelijk product in de markt heeft gezet. Met dit hulpmiddel kan een continue cyclus worden opgezet van het monitoren van relevante data, het analyseren daarvan en het vervolgens opzetten of aanpassen van bedrijfsplannen. Het product maakt deel uit van de Microsoft Office 2007-productreeks. Zie afbeelding 2.

van dit model kan ook een inschatting worden gemaakt van de risico's die het bedrijf loopt bij een nieuwe transactie.
Dat kan eventueel ook op een andere en wellicht ook wel slimmere manier worden gedaan, meent Lukawiecki. Hij vertelde dat de volgorde van transacties met behulp van een algoritme als sequence clustering kan worden gemodelleerd. Pas vervolgens algoritmes als neural network, decision tree en clustering toe om binnen zo'n groep ongebruikelijke transacties (de zogeheten *outliers*) op te sporen. Op basis van een dergelijke aanpak is het mogelijk om nieuwe events te analyseren op het moment dat zij daadwerkelijk optreden. Zo kan dus direct een inschatting van de kans op fraude worden gemaakt. Een vorm van real-time fraudedetectie dus.

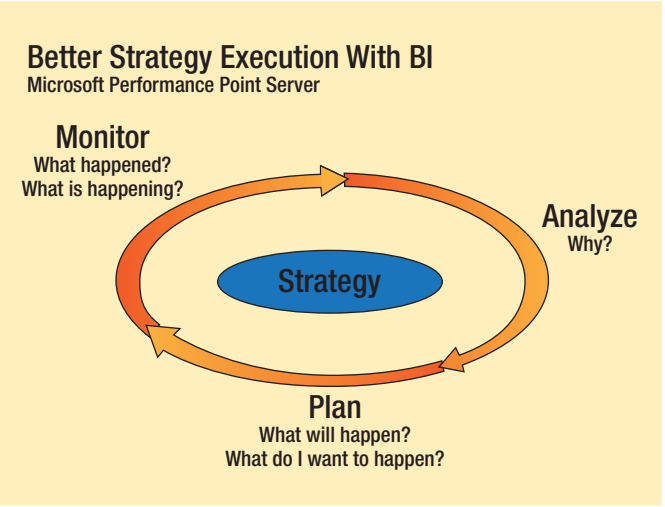
Concurrenten

Hoe zit het productaanbod van Microsoft op het gebied van BI er nu precies uit? Microsoft biedt zogeheten Analysis Services sinds SQL Server 2005. Officieel heeft het bedrijf een aantal producten voor dit doel ontwikkeld: SQL Server Integration Services, SQL Server Analysis Services, SQL Server Reporting Services, SQL Server Data Warehousing en SQL Server Data Mining. De geboden functionaliteit ligt gezien de naamstelling voor de hand.
Met deze producten – de 2008-editie is bovendien inmiddels in aantocht – concurreert Microsoft volgens Lukawiecki met nogal wat aanbieders. Opmerkelijk genoeg toch ook weer met traditionele spelers als SAS en SPSS. SAS omdat het nu eenmaal de grootste partij is, ook al richt dit bedrijf zich voornamelijk op traditionele BI-experts. SPSS is met Clementine sterk op het gebied van statische analyses en probeert nu zijn stempel te drukken op text mining.

Lukawiecki plaatst in het rijtje concurrenten ook IBM's Intelligent Miner, dat hij een uitstekend product noemt dat weliswaar nauw verbonden is met DB2, maar dat via PMML (Predictive Modeling Markup Language) toch ook prima in staat is samen te werken met een Microsoft-omgeving. Ook in Oracle ziet Lukawiecki een concurrent, aangezien deze firma in 10g Java API's ondersteunt. Kijk daarnaast ook eens naar Angoss of KXEN, adviseert hij. De eerste biedt vooral tools voor het visualiseren van resultaten die uit SQL Server worden gehaald, terwijl KXEN ondersteuning kent voor OLAP en Excel.

ProClarity

Lukawiecki is behoorlijk enthousiast over de al eerder genoemde algoritmen waarin hij de hand ziet van Microsoft Research, de R&D-poot van het concern. De interoperabiliteit met andere BI-omgevingen is verder redelijk goed geregeld, meent hij, en leunt zwaar op PMML. Daarmee is koppeling mogelijk met bijvoorbeeld SAS, SPSS, Oracle en IBM.
Bovendien biedt SQL Server 2005 een aantal interessante tools. BIDS bijvoorbeeld, wat staat voor Business Intelligence Development Studio. Hiermee kunnen modellen worden gebouwd. Verder is voorzien in de zogeheten Data Mining extensions for Excel (DMX), waarmee vanuit Excel de BI-hulp van SQL Server kan worden ingeroepen. Met de OLE DB kunnen aandere databronnen dan SQL Server worden gebruikt, terwijl ook is voorzien in ondersteuning van XML for Analysis (XMLA) waarmee analytische softwarepakketten toegang tot elkaars data kunnen krijgen.
Inmiddels is SQL Server 2008 alweer in aantocht. Van alle hiervoor genoemde BI-tools komen nu ook 2008-versies. Die bieden op een aantal vlakken verbeteringen waarbij Lukawiecki de nodige invloed meent te bespeuren van het eerder overgenomen ProClarity. Het wordt daardoor bijvoorbeeld eenvoudiger, zo stelt hij, om modellen te ontwikkelen en te testen, er is voorzien in een mogelijkheid voor het cross-valideren van modellen, terwijl ook filtering van data mogelijk is.



Afbeelding 2: Microsoft Performance Point Server maakt een continue cirkel van meten, analyseren en plannen mogelijk.

Office 2007

Wat algoritmen betreft ziet Lukawiecki met name mogelijkheden in de verbeteringen die ten aanzien van time series zijn aangebracht. Voor de fijnproevers: dit algoritme verenigt nu de kenmerken van zowel de ARIMA- als ARTXP-modellen in zich.
Daarnaast introduceert Microsoft een eigen data mining framework, dat gebaseerd is op het bekende CRISP-DB dat als neutraal standaardmodel van het data mining-proces geldt. Microsoft heeft dit model echter op een aantal punten uitgebreid.
Een sterk punt van Microsoft, zo stelt Lukawiecki ten slotte, is de integratie met Office 2007. In de demo's die hij tijdens zijn presentatie verzorgde, liet hij zien hoe de eerder genoemde BI-extensies voor Excel 2007 het mogelijk maken om op een relatief eenvoudige wijze gegevens in een spreadsheet door de BI-voorzieningen van SQL Server te laten analyseren. Vrijwel iedere technische complexiteit is hierbij verstopt achter wizards en knoppen, waardoor de gebruiker zich vrijwel uitsluitend behoeft te bekommeren om de te onderzoeken gegevens zelf.

Robbert Hoeffnagel is freelance journalist.