De evolutie van datamining

Mine your own business

Jaap Verhees

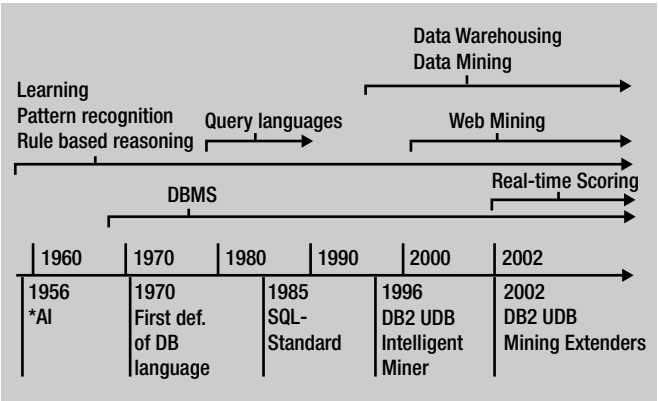
Datamining is in de laatste tien jaar van nut geworden in het bedrijfsleven om meer informatie op te halen, meer begrip te krijgen over – en daarmee meer grip te hebben op – de bedrijfsvoering, en om nieuwe wegen te vinden om naar andere markten te extrapoleren.
De kracht die datamining biedt op het gebied van analyse bij het ontdekken van nieuwe mogelijkheden hoeft niet langer bewezen te worden aan bedrijfsanalisten en datamining-analisten. De sceptici zijn inmiddels overtuigd.

Tegenwoordig wordt datamining niet langer beschouwd als een verzameling van stand-alone technieken die ver van de bedrijfsapplicaties afstaan. Vooral in de laatste drie jaar is de mogelijke integratie van datamining met kritische bedrijfsapplicaties een belangrijk aspect geworden voor e-commerce. Afbeelding 1 toont een historisch beeld van datamining.

Een populaire applicatie van datamining is inmiddels de voorspelling geworden van de toekomstige kooppatronen van e-commerce websitebezoekers, op basis van de informatie die wordt verzameld gedurende de interactiesessies met de websitebezoekers. Deze informatie kan demografische, sociografische, en technografische gegevens omvatten, die worden verkregen door middel van online-enquêtes en vragenlijsten. De website kan ook browser-informatie verzamelen door de bekeken objecten en de geklikte links vast te leggen.

Meta Group gaf ook aan dat datamining-tools en *workbenches* moeilijk te verkopen zijn als stand-alone tools, maar beter als onderdeel van een totaaloplossing. Bedrijven verlangen steeds vaker integratie van datamining-technologie met relationele databases en hun bedrijfsapplicaties. Om deze beweging te ondersteunen verplaatst het datamining-product zich van stand-alone technologie naar geïntegreerde technologie in de relationele database en verpakt in de applicaties. Zie afbeelding 2.

De gangbare definitie van datamining is: "The automated use of algorithms to sift through large amounts of data to find useful information, patterns, and trends that may have been previously unseen or unknown. The results gathered from mining data can then be used to devise marketing plans, better customer relations, improve sales, and more."



Afbeelding 1: Een historisch beeld van datamining.

In de traditionele datamining aanpak zal een expert een workbench gebruiken om de preprocessing-stappen, mining-algoritmes en visualisatie van de datamining-modellen uit te voeren. Het doel is om nieuwe en interessante patronen in de data te ontdekken en de gevonden inzichten en kennis op een of andere wijze te gebruiken in bedrijfsbeslissingen. De power users, zoals datamining-analisten en OLAP-analisten, verlangen separate en gespecialiseerde datamarts om de gespecialiseerde analyse op de voorbewerkte data uit te voeren. De integratie van datamining-resultaten in de operationele bedrijfsomgeving gebeurt in dat geval dan ook veelal ad hoc.

Met de integratie van datamining is de focus nu meer gericht op implementatie van datamining in bedrijfsapplicaties (Afbeelding 2). Het publiek omvat nu ook de business eindgebruikers, die meer baat hebben bij een gebruiksvriendelijke interface gericht op datamining-resultaten en een strakkere integratie met de bestaande omgeving.

Een power user is nog steeds nodig om het ontwerp en de bouw van geoptimaliseerde en efficiënte datamining-modellen te bewerkstelligen. De power user beschouwt alle data van zowel de operationele databases en het datawarehouse als potentiële interessante kandidaten en op de traditionele niet bedrijfsgeïntegreerde wijze.

Zodra het datamining een onderdeel wordt van een totaaloplossing, is de echte gebruiker van de datamining-functies een ontwikkelaar die de resultaten van datamining verpakt voor de eindgebruikers. Deze persoon heeft de database-applicatie ontwikkelvaardigheden, en voedt de modellen en scores terug in het datawarehouse met geaggregeerde data, gespecialiseerde datamarts, en operationele databases die data bevatten op het niveau van individuele transacties.

Tegenwoordig maken veel applicaties, zoals Web-analyse, Web-personalisatie, e-commerce, en Customer Relationship Management (CRM), gebruik van geïntegreerde datamining-functies. Er zijn diverse business drivers om modellen en scores in een bedrijfsvoering uit te rollen. Bijvoorbeeld, de behoefte aan:

- Sneller time to market en 'closing the loop';
- Real-time analyse;
- Bestaande IT-skills verder benutten;
- Bouwen van herbruikbare processen en taken;
- Efficiëntie and effectiviteit;
- Kostenreductie van datamining-analyse.

Hoe wordt dit alles nu gefaciliteerd? Door ondersteuning met de nieuwste technologie die IBM aanbiedt door middel van drie database-extenders voor datamining, te weten: IBM DB2 Intelligent Miner Modeling, IBM DB2 Intelligent Miner Scoring en IBM DB2 Intelligent Miner Visualization.

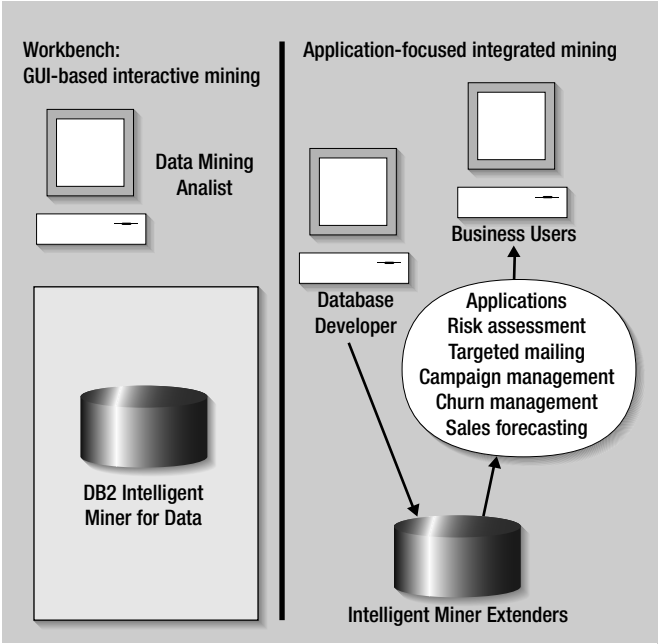
SCORING

Scoring is het gebruik maken van bestaande datamining-modellen die zijn gebaseerd op historische data en het simpelweg toepassen van deze modellen op nieuwe data. Bijvoorbeeld, je hebt een classificatiemodel dat regels bevat over het schatten van het "churn" risico voor klanten. Gegeven de profieldata van een bepaalde klant, berekent de scoringfunctie het churn-risico. Je kunt deze functie in real time toepassen op een individueel record, bijvoorbeeld de klant die momenteel contact heeft met iemand in het call center.

Het basisprincipe achter IM Scoring is de notie van datamining versus bedrijfsregels. De bedrijfsomgeving is in beweging, waarbij frequente updates nodig zijn, terwijl bedrijfsregels juist snel hun waarde in de tijd verliezen. Bedrijfsregels zijn daarbij vaak handmatig geïdentificeerd, hetgeen zowel een arbeidsintensieve als tijdverslindende bezigheid is.

IM Scoring identificeert onverwachte targets, kansen en problemen op een geautomatiseerde wijze. Bijvoorbeeld nieuwe data (wijziging in gedrag of karakteristieken van transacties) kan de IM Scoring-applicatie afvuren, om een score te produceren op basis van het onderliggende datamining-model. Dan vergelijkt het deze score tegen een bereik van andere scores, om automatisch te signaleren of dit een (on)verwacht resultaat oplevert.

IM Scoring biedt een gebruikersvriendelijke datamining-



AFBEELDING 2: SHIFT IN HET GEBRUIK VAN DATAMINING TECHNOLOGIE EN PUBLIEK.

faciliteit, dat:

- een database extensie is;
- geïmplementeerd kan worden in batch-modus of real-time-modus;
- de Predictive Model Markup Language (PMML) ondersteunt, en gebruik maakt van bestaande IT-vaardigheden binnen de organisatie.

Let op, IM Scoring genereert geen nieuwe modellen of nieuwe voorspelregels. De datamining-modellen die de rules bevatten zijn dus berekend via andere modelleerfuncties of zijn geïmporteerd vanuit externe datamining workbenches.

MODELING

IM Modeling biedt een set van functies zoals *add-on service* op DB2 Universal DataBase. De set bestaat uit een serie van user-defined functies (UDF), methoden, stored procedures en tabellen. We kunnen deze datamining-modelleerfuncties simpel gebruiken vanuit SQL-statements. IM Modeling zorgt daarmee voor:

- Interoperabiliteit.
  De modellen kunnen ontwikkeld zijn in andere applicaties en tools die interoperabiliteit ondersteunen via PMML-modellen. Of de modellen uit DB2 Intelligent Miner for Data (IM4D) zijn geëxporteerd als PMML-modellen. In de modelleerfase van een datamining-project zijn diverse modelleertechnieken voorhanden en toegepast voor eenzelfde datamining-probleem. Hun parameters zijn gekalibreerd tot optimale waarden.
- Bouwen en gebruiken van datamining-modellen die zijn opgeslagen in DB2 UDB.
  We voeren het modelleren uit via aanroepen naar de database,

PMML

PMML is een taal die is gebaseerd op XML, en biedt gebruikers een vlotte en simpele wijze om modellen uit te wisselen tussen de applicaties van vendors die aan deze standaard conformeren. Het is een vendor-onafhankelijke methode om modellen te definiëren, zodat proprietary kenmerken en incompatibiliteit niet langer een barrière vormen bij uitwisseling.

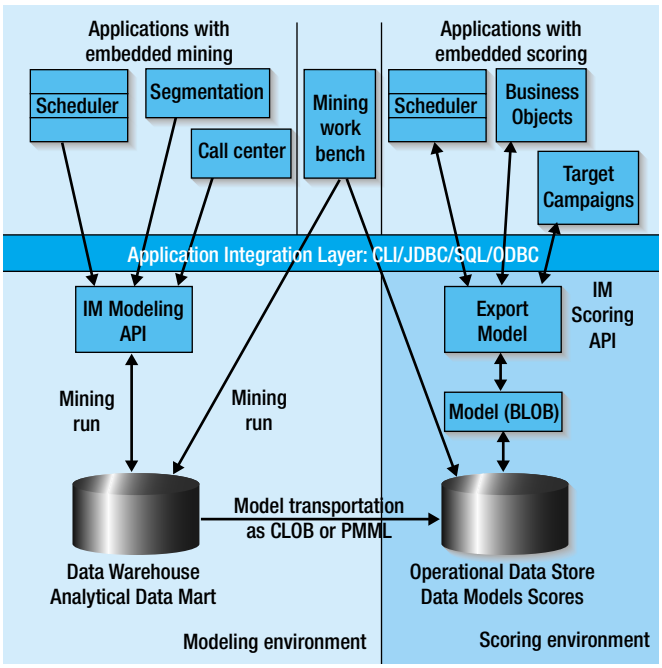
Een model dat in de ene applicatie is ontwikkeld kan in een andere applicatie gebruikt worden om de modellen te visualiseren, analyseren, evalueren of anderszins. Voorheen was dit praktisch onmogelijk, met PMML is dit simpel geworden.

Omdat PMML is gebaseerd op XML, komt de specificatie van PMML in de vorm van een XML Document Type Definition (DTD). Zie http://www.dmg.org/pmml-v2-0.htm

via zogenaamde calls. De SQL API biedt de mogelijkheid om deze calls te maken en biedt SQL-applicaties de voorzieningen om calls in de richting van associatie-, clustering- en classificatietechnieken te doen, om zodoende modellen te ontwikkelen die zijn gebaseerd op data die ook benaderd kunnen worden via dezelfde database.

Modellen zijn opgeslagen in DB2 UDB. Dit biedt op zich weer beheergemak, gecentraliseerde controle en beveiliging. Plus, het hebben van verschillende versies van modellen in een centrale database maakt het makkelijk om te switchen van model en biedt meer overzicht. Dit leidt ook tot kostenreductie van data- en modelbeheer.

- Ondersteuning voor de nieuwe PMML V2.0 standaard.
  De verschuiving van traditionele datamining workbenches (SAS Enterprise Miner, IM4D) naar meer modulaire tooling, om datamining uit te voeren en uit te rollen, wint terrein. Veel vendors van datamining-producten conformeren zich aan de PMML-standaard. Het resultaat is dat organisaties nu kunnen besluiten om het datamining-model van bedrijf A te kopen en de visualisatie- en applicatie-opties van tools van bedrijf B te kopen om het model te gebruiken.



Afbeelding 3: Componenten in business scenario's.

Een bank bijvoorbeeld, kan besluiten om een model met data ontwikkeld met SAS/EM aan te schaffen en het te laten exporteren in PMML-formaat, om daarna de IBM datamining-tools voor scoring en visualisatie te gebruiken om de resultaten uit te rollen naar de operationele bedrijfsomgeving.

Vice versa kunnen we IBM datamining tools gebruiken voor het modelleren van data en dan besluiten om het model op te nemen in de uitrolfase van de datamining-applicatie. We kunnen het prebuilt datamining-model via PMML-import in een CRM-applicatie zoals Siebel voegen. Op deze wijze is de organisatie dus niet afhankelijk van één leverancier.

VISUALISATIE

IM Visualization biedt Java-visualizers om de data-modelleerresultaten te presenteren voor verdere analyse. IM Visualization is gebaseerd op een aantal principes, als eerste interoperabiliteit. De modellen mogen ontwikkeld zijn in IM Modeling of andere applicaties en tools die interoperabiliteit verzorgen door gebruik te maken van PMML-modellen. De verschillende visualizers opereren met de modellen.

Ten tweede keuze in gebruik. Applicaties kunnen de IM Visualizers aanroepen om de model-resultaten te presenteren. We kunnen de visualizers als applets aanroepen en draaien in een web browser. Ten derde multi-platform mogelijkheden. Omdat de functionaliteit van IM Visualization geschreven is in Java, kunnen we de visualizers installeren op diverse hardware en software platformen.

En tenslottte ondersteuning voor de PMML V2.0 standaard. We kunnen de IM Visualizers gebruiken bij datamining-modellen die aan PMML geconformeerd zijn.

INTEGRATIE VAN DE GENERIEKE COMPONENTEN

Er zijn generieke componenten die een end-to-end uitrolproces van het integreren van IM Modeling, IM Scoring en IM Visualization in bedrijfsapplicaties en oplossingen omvatten.

De diverse omgevingen en componenten die in verscheidende bedrijfsscenario's terugkomen, zoals geïllustreerd in afbeelding 3, bestaan uit:

- Klant profilering;
- Campagne management;
- Trigger-based marketing;
- Fraude detectie;
- Up-to-date promotie.

Voorbeeld van implementatiestappen

Klantprofilering is belangrijk om de commerciële doelen voor een *cross-sell campagne* te stellen. Maar alleen data zijn niet genoeg om actie te kunnen ondernemen. De integratie van de resultaten van datamining met rapportages is een kritische succesfactor om een succesvolle marketingstrategie te volgen.

Bij de promotie is *kannibalisatie* tegenwoordig een serieus probleem in Direct Marketing. Het gebruik van historische data is niet meer het enige aspect dat een rol speelt bij het voorspellen van gedrag, zoals de kansberekening dat een bepaald product wordt gekocht of de kans dat een klant vertrekt bij een bank of retailer. Ook het bepalen wat de kans is dat een bestaande klant ingaat op een promotie of juist overvoerd is en voorlopig niet meer benaderd moet worden met aanbiedingen, is een nieuw aspect.

Bij geïntegreerd campagnemanagement is het nodig om een combinatie van verschillende datamining-functies of -algoritmes te gebruiken. Klantprofilering is een prachtig tool om target lijstjes te bepalen, maar het is misschien niet genoeg. Voorspelalgoritmes kunnen nuttig zijn bij het bepalen van het gewenst promotieniveau van een klant op basis van:

- Het aantal contacten in het verleden;
- De tijd die ligt tussen contacten;
- Het type contactkanaal dat is gebruikt;
- Het type promotie;
- Het aantal producten in de klantportefeuille.

Een gerichte marketingcampagne kan best moeilijk te implementeren zijn. De grote datavolumes die nodig zijn, zijn vaak moeilijk te benaderen, laat staan te consolideren met conventionele tools binnen operationele systemen. Veel organisaties missen de expertise om complexe datamining en analytische of voorspellende taken uit te voeren, taken die juist noodzakelijk zijn om de effectiviteit van campagnes te verhogen.

Maar juist elke interactie met de externe wereld van een organisatie levert een kans om een bestaande klant te houden, om nieuwe klanten te winnen, of om cross-selling in een bepaalde productlijn voor mekaar te krijgen. Als het contact niet goed is gelegd kan een mooie kans ineens omslaan in een nachtmerrie. Het kost tijd en planning bij het faciliteren om met verschillende scenario's te kunnen omgaan, die optreden tijdens dergelijke interacties.

Voor een algemeen campagne managementproces geldt dat het uitrolproces de volgende vijf componenten moet behelzen:

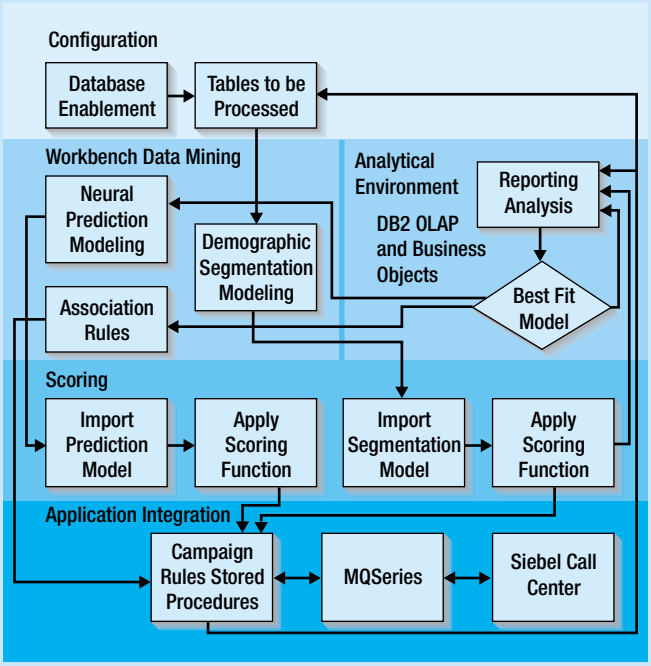
- Datastore;
- Data-analyse;
- Data-workflow;
- Applicatie-integratie;
- Kanalen.

De informatie dient vlekkeloos door elk van deze componenten te stromen. Het proces neemt de data uit het datastore-gebied, voert verschillende analyses uit, en afhankelijk van het resultaat van de analyses, neemt het een beslissing over hoe de klant dient te worden benaderd, op basis van vooraf gedefinieerde campagneregels. Nadat de beslissingen zijn gemaakt, neemt de datastroom-controller de resultaten van de analyse mee naar de communicatiekanalen die in de campagneregels zijn bepaald.

In een dergelijke implementatie dienen we dus allereerst een klantgerichte centrale databank aan te leggen, bijvoorbeeld een organisatiewijd datawarehouse. Vervolgens dienen we de data-analysecomponenten te implementeren. Om belangrijke indicatoren te analyseren kunnen we allereerst gebruik maken van rapporten en OLAP-analyses. We kunnen daarnaast campagneresultaten en datamining gebruiken om additionele data te genereren, die we vervolgens gebruiken in onze analyses. We kunnen de datamining-modellen hanteren in een dergelijk scenario:

- Demografische clustering voor klantprofilering;
- Demografische clustering voor productprofilering;
- Associatie voor product bundelen;
- Neurale voorspelling van kannibalisatie bij promotie-inspanningen.

Het OLAP-datamodel heeft traditionele dimensies (tijd, product, regio, kanaal, contacttype) en meetwaarden zoals winst, kosten, omzet, gemiddeld aantal producten. Met datamining hebben we de mogelijkheid om nieuwe dimensies aan het model toe te voegen, het segmentatietype en het segmentnummer.



Afbeelding 4: Cross-selling outbound campaign implementation flow.

Om de mogelijkheden voor cross-selling binnen Direct Marketing te bepalen, worden associatieregels afgeleid uit de historische data, gebaseerd op de doelsegmenten voor de promotie. Deze regels, in combinatie met de analyse van de OLAP-data bij het bepalen welke producten met welke meest winstgevende segmenten samengaan, helpen de marketingafdeling bij het bepalen van het aanbod.

Nadat de modellen die we willen gebruiken zijn bepaald, worden de campagneregels gedefinieerd door middel van *stored procedures*. Deze procedures zijn verantwoordelijk voor de creatie van XML-documenten die naar Siebel worden gezonden via WebSphere MQ (MQSeries) transport. Siebel workflow is verantwoordelijk voor het versturen van de reacties terug naar het datawarehouse.

Als de reacties zijn verzameld in het call center-systeem en weer opgeslagen in een datastore, word Business Objects gebruikt om rapporten met campagneresultaten te genereren en de effectiviteit van de campagne te tonen. IM Modeling kan gebruikt worden om het model te tunen, met name het voorspellingsmodel, ook weer gebaseerd op basis van de effectiviteit van klantcontacten. Voor het meten van de effectiviteit en het mogelijk tunen van het model zijn de typen reacties belangrijk.

De volgende typen reacties zijn in dat geval een voorbeeld van nuttige informatie om te verzamelen:

- Klant hing op;
- Klant was niet geïnteresseerd in het horen van voordelen;
- Klant vroeg naar geprint materiaal over het aanbod;
- Klant was geïnteresseerd in individuele producten maar niet in een productbundel;
- Klant accepteerde het aanbod;
- Klant vroeg om additionele producten.

Afbeelding 4 geeft het implementatieproces weer.

Zoals we zien kan het resultaat van diverse individuele datamining-processen gecombineerd worden om een meer complexe beslissingsstroom voor de campagne te ontwerpen. Merk op dat een van de typische eigenschappen van deze flow haar 'closed loop' is, die door de campagneregels wordt geboden. Resultaten van de marketing acties worden teruggevoed in de rapporten en brontabellen voor (daarop volgende) aanpassingen.

We kunnen drie verschillende omgevingen onderscheiden:

1) De bedrijfsapplicatie-omgeving.
   Bedrijfsapplicaties kunnen alle applicaties zijn die zich lenen voor analyse. Het kunnen real-time- of batch-applicaties zijn die een fraudedetectie-applicatie omvatten. Maar bijvoorbeeld ook een call center-applicatie, die de kans op churn van een klant berekent, ook weer gebaseerd op de meest recente wijzigingen. Of een call center-applicatie die een "next best offer" promoot, gebaseerd op de meest recente klantinformatie. Of een wekelijkse batch-applicatie, een zogenaamde scheduler, die klantenleads voor gerichte verkoopcampagnes genereert. Meer voorbeelden; een traditionele workbench omgeving zoals IM4D of SAS/EM of een credit scoring-applicatie op de desktops van kredietadviseurs in bankkantoren.
2) De modellering-omgeving.
   De modellering omgeving is een DB2 database die "enabled" is voor datamining. De database heeft alle additionele database-objecten die nodig zijn voor het uitvoeren van de IM Modeling API. Normaliter is de bron-database een deel van het datawarehouse, waarbij de data die worden gemodelleerd al wel zijn opgeschoond, verbonden, voorbewerkt en verrijkt.
3) De scoring-omgeving.
   Scoring is het toepassen van een model dat is gebouwd in een modelleeromgeving. Het model moet dus bestaan alvorens het kan worden gebruikt voor scoring van nieuwe klantgegevens. Meestal zal de scoring-omgeving dan 1 tabel behelzen in de Operational Data Store (ODS). We beschouwen een ODS als een huis dat halverwege een datawarehouse en OLTP-systemen staat. Modellen zijn ontwikkeld in een datawarehouse en uitgerold naar de ODS. Uiteindelijk worden de scores als resultaat van de scoring-run dan hoogstwaarschijnlijk geïntegreerd in de operationele OLTP-applicaties, waar de scores juist nodig zijn voor verschillende verkoopcampagnes, het call center of CRM-applicaties.

ANDERE INTEGRATIE-MOGELIJKHEDEN

Database-geïntegreerde mining maakt datamining simpeler en biedt daarmee meer mogelijkheden bij de integratie met diverse andere applicaties. Een recente trend voor DB2 datamining is dan ook integratie in andere analytische oplossingen. Zo zijn er al web analytical-oplossingen voorhanden, die real time-scoring bieden op basis van website-verkeersgedrag. En de IBM IM4D-technologie is al geïntegreerd in SAP Business Warehouse en CRM. Zo is bijvoorbeeld klantinteractie opgedirkt met productaanbevelingen, die zijn berekend met behulp van datamining. Tevens bevat de WebSphere Commerce Analyzer inmiddels *predefined mining configuraties*.

CONCLUSIE

Tegenwoordig is datamining is niet langer een set stand-alone technieken, ver van de bedrijfsapplicaties gesitueerd en alleen in gebruik bij datamining-specialisten of statistici. Integratie van datamining met mainstream-applicaties wordt momenteel een belangrijk punt voor e-businessapplicaties. Om deze beweging te faciliteren, is datamining nu een extensie van de relationele databases die de database-administrators of IT-ontwikkelaars gebruiken. Zij gebruiken datamining nu net zoals zij elke andere standaard relationele functie hanteren. ●

Dr. Jaap Verhees (jaap_verhees@hotmail.com) is adviseur.