Aan kop in data-integratie

Informatica vervult strategisch doel

Dick Schievels

Als je een themanummer maakt over data-integratie kun je bijna niet om Informatica heen. Het bedrijf domineerde jarenlang de ETL-markt maar formuleerde in 2005 de wens ook een leidende positie te willen veroveren in het opkomende marktsegment data-integratie. Dat strategische doel lijkt anno 2011 zo goed als bereikt.

November vorig jaar (DB/M 7, 2010) constateerde Norman Manley, de man achter onze ETL-matrix (zie www.dbm.nl) aan de hand van Google-statistieken betreffende in Google ingetikte zoekopdrachten dat gebruikers nog steeds praten over ETL, waar leveranciers vooral de term data-integratie hanteren. Hij tekende daarbij aan dat data-integratie weliswaar een breder gebied bestrijkt dan het oorspronkelijke ETL, maar in wezen dezelfde problematiek beschrijft.

Een bedrijf waar deze observatie perfect op van toepassing blijkt te zijn, is Informatica. Informatica heeft ETL bij wijze van spreken uitgevonden. Het werd opgericht in 1993, in de tijd dat de managementinformatiesystemen en datawarehouses in opkomst waren. Bedrijven schreven nog hun eigen programma's om gegevens uit mainframes en andere databases te halen en die vervolgens in een datarapportageomgeving te zetten: een datawarehouse.

"Dat was natuurlijk nogal arbeids- en tijdsintensief", vertelt Bert Oosterhof, Technisch Directeur bij Informatica Europe. "Informatica bedacht toen dat dat gemakkelijker moest kunnen door het een niveau hoger aan te pakken. Als je een op een repository gebaseerde omgeving creëert, waarbij je specificeert wat er moet gebeuren, dan heb je geen programmeurs meer nodig en kun je de code die je moet uitvoeren gewoon laten genereren door een engine, die dat veel sneller doet dan een legertje programmeurs."

Dat was het startpunt van de eerste generatie ETL-software, de codegeneratoren. Niet lang daarna gevolgd door de transformation engines die alle ETL-bewerkingen centraal op een server verrichten. Maar ETL werd op een gegeven ogenblik minder sexy en de term 'data-integratie' kwam in zwang. Waaraan zoiets precies ligt, is vaak moeilijk te traceren. Je merkt gewoon dat er in de markt opeens een terminologieverschuiving plaatsvindt en dat men het ene concept actief vervangt door het andere.

Real-time

"ETL, daar praten we eigenlijk zo min mogelijk over", zegt Bert Oosterhof spontaan, als ik de term tijdens het interview ter sprake breng. Waarom eigenlijk? "Het is natuurlijk wel een belangrijk fenomeen", legt hij uit, "maar ETL staat voor Extract, Transformation en Load. Extractie betekent dan meestal een actie om iets ergens vandaan te halen. Meestal gaat het dan om batchprocessen die je 's nachts draait om bijvoorbeeld je datawarehouse bij te werken. Maar die extractie is tegenwoordig niet meer alleen een 'pull' bestaande uit het ophalen van data uit je bronsystemen, maar ook een 'push'. Er gebeurt iets en je wilt dat meteen verwerken: databasereplicatie, message queueing, Microsoft MQ of Tibco bijvoorbeeld. Die messages die binnenkomen, daar wil je ook iets mee doen."

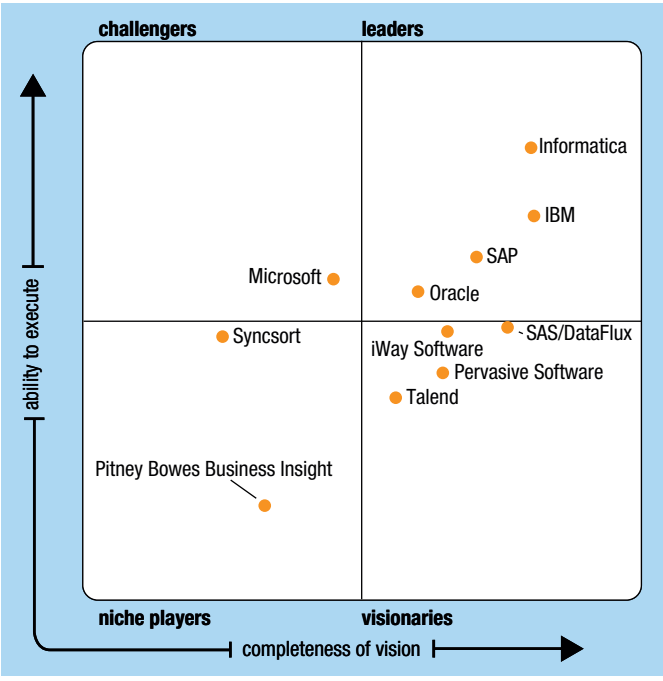
Oosterhof doelt op de real-time of near real-time verwerking, een eis die tegenwoordig op het vlak van data-integratie steeds belangrijker wordt. "Eigenlijk wil je daar dezelfde functies op los laten: kwaliteitscontroles, kwaliteitsverbetering, logica, transformaties enzovoort. Bij een bank komt bijvoorbeeld een Swift-transactie binnen die direct moet worden omgezet in een DB2- of Oracle-transactie. Dat zijn ook allemaal dingen die we met onze huidige technologie kunnen doen. Vandaar dat we uit die ETL-hoek vandaan willen."

Wat dat real-time aspect aangaat, merkt Oosterhof op dat als je je op dat vlak begeeft je steevast van klanten de vraag krijgt: jullie kunnen mijn message verwerken, maar wat als ik er nu één miljoen per seconde stuur? "Er zijn niet zoveel bedrijven die dat nodig hebben, maar in sectoren als de bankenwereld, energy trading en financial trading zijn de eisen wél zo hoog. Daarvoor hebben we anderhalf jaar geleden ook een stuk technologie overgenomen, ultra messaging, zodat we ook dat aspect kunnen oplossen. Daarmee hebben we nu een oplossing in huis waar we een paar miljoen messages per seconde

mee kunnen verwerken en bijna real-time kunnen analyseren."
De klassieke ETL-lijn van Informatica bestaat uit de PowerCenter-producten. PowerCenter is waar het bedrijf groot mee geworden is. Oosterhof: "Dat is van oorsprong een ETL-tool maar de laatste vijf jaar sterk veranderd en uitgebreid. Het is aan de achterkant herschreven. De gebruiker ziet daar niets van maar het werkt nu met een service oriënted architectuur en real-time; al dat soort dingen zijn erbij gekomen."

Vijf jaar geleden

Als we even in de archieven duiken van Database Magazine blijkt uit een artikel van Paul van der Linden uit 2005 dat de leiding van het bedrijf toen al voor het eerst de boodschap uitdroeg dat het vanuit zijn dominante positie op gebied van ETL en datawarehousing in de toekomst ook een leidende positie wilde veroveren op het terrein van data-integratie. Van der Linden bespreekt in zijn artikel onder meer een reeks componenten van data-integratie en eindigt zijn stuk met een reeks eisen, opgesteld door mensen van The Data Warehouse Institute, waaraan een leverancier van data-integratietechnologie uiteindelijk zal moeten kunnen voldoen, waaronder: grote datavolumes kunnen verwerken, diverse databronnen kunnen ontsluiten, krimpende batchwindows kunnen ondervangen, 24x7-beschikbaarheid, datakwaliteitfunctionaliteit en metadata management. Hij concludeert dat Informatica met zijn productaanbod van dat moment als basis een zeer gunstige uitgangspositie heeft om zijn strategische doel uiteindelijk te bereiken.
Nu, zes jaar later, blijkt inderdaad uit allerlei indices dat dit ook daadwerkelijk is gelukt. Een daarvan is de positie die het bedrijf in het leiderskwadrant van Gartner's Magic Quadrant (november 2010) voor data-integratietools inneemt (zie afbeelding 1).



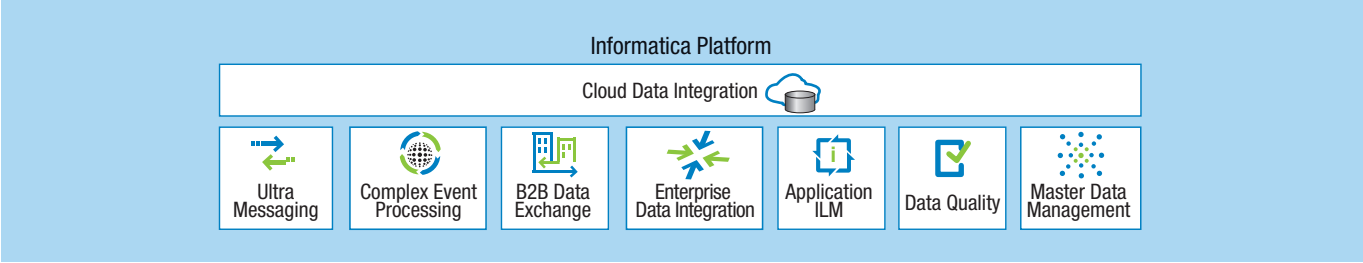
Afbeelding 1: Gartner Magic Quadrant data-integratietools (Nov. 2010).

Bert Oosterhof van Informatica: "ETL, daar praten we eigenlijk zo min mogelijk over."

Daaruit blijkt dat Informatica reuzen als IBM, SAP en Oracle afgetekend achter zich laat.

Productportfolio

Richten we onze blik wat nauwkeuriger op het huidige Informatica-platform (zie afbeelding 2) dan vinden we naast de centrale pijler Enterprise Data Integration, de aanvullende bouwstenen: ultra messaging, complex event processing, B2B data exchange, application ILM, data quality en masterdata management. Uit de afbeelding blijkt ook dat Informatica zijn data-integratieoplossingen tegenwoordig tevens via 'the cloud' aanbiedt.
Oosterhof over datakwaliteit: "Als je verder kijkt dan alleen de dataverplaatsing, zie je dat er nog een hoop aanvullende functionaliteit nodig is om van een volwaardige data-integratieoplossing te kunnen spreken. Als je data hier fout zijn en je verplaatst ze daar heen, dan heb je ze op twee plaatsen fout staan. Dus datakwaliteit is hierin steeds belangrijker geworden. Daarom is datakwaliteit ook een belangrijk onderdeel van ons productaanbod." Oosterhof is van mening dat bedrijven meer en meer inzien dat data een essentieel onderdeel vormen van de bedrijfsvoering. "Je hebt mensen, je hebt geld en je hebt data. Daar draait je hele bedrijf eigenlijk op."
In het verleden was het veel meer applicatiegericht. Men had een HR-systeem met personeelsinformatie, een voorraadsysteem, een inkoopsysteem, een klantinformatiesysteem, enzovoort.
En in die systemen zaten data. Die waren een onderdeel van de applicatie. Vandaar dat je op heel veel verschillende niveaus, in heel veel verschillende systemen dezelfde data had vastgelegd.



Afbeelding 2: Het Informatica-platform.

"Hier heb je klantinformatie nodig en daar heb je klantinformatie nodig. We zien nu steeds meer het belang van de abstractielaag, zodat je de functionaliteit scheidt van de data. Dat maakt dat je de data gemakkelijker kunt beheren: wie zijn de eigenaars, wat zijn de business rules, wat de datakwaliteitsregels en wat zijn de eisen die de business aan die data stelt."
Oosterhof betreedt daarmee, zo stellen we vast, de voor data-integratie onmisbare disciplines datagovernance en masterdata management. "Ja, masterdata management én metadata management", vult hij aan. "Metadata management omvat zeg maar de definities van wat er allemaal gebeurt en wat het allemaal betekent. En masterdata management gaat over 'the single point of view' en 'the single point of control', waarbij je zegt: hier heb ik één versie van de klant staan; welk systeem het ook nodig heeft, hier staat de mastercopy. Dat zie je ook steeds meer in zwang komen. En bij al dat soort projecten speelt Informatica een steeds belangrijkere rol. Kort samengevat bestaat onze product-stack dus uit data-integratie in allerlei soorten en maten, datakwaliteit, dataprofiling en masterdata management. Eigenlijk leveren we een totaal databaseplatform, alleen de persistency-laag, de database zelf, die laten we aan anderen over."

Trends

Gevraagd naar de belangrijkste trends van dit moment wijst Oosterhof op de opkomst van allerlei nieuwe vormen van data, zoals ongestructureerde data, webpagina's en data afkomstig van social media als Twitter en Facebook. "Bedrijven willen daar iets mee, want als iemand iets goeds of slechts zegt over het bedrijf dan is dat van waarde en willen ze dat integreren met hun overige gegevens. In beeld krijgen wat dat voor positieve of negatieve gevolgen voor ze heeft. Dat vereist integratie van data van allerlei aard." Heeft Informatica ook nog iets te melden op gebied van Big Data? "Daar zit het nodige aan te komen", klinkt het geheimzinnig. "Maar we hebben nu al een oplossing voor Hadoop, het nieuwe platform dat gebruikt wordt voor Big Data."
De basis voor Hadoop is ontwikkeld door Google, leren we van Oosterhof. Zij zagen als eerste dat datavolumes in een enorm hoog tempo begonnen te groeien. Zo'n hoog tempo dat je als het ware geen tijd meer hebt om back-ups te maken. Ze hebben toen een gedistribueerd filesysteem ontwikkeld. Oosterhof legt uit: "Dat betekent dat als ik hier een blok informatie heb en ik moet dat naar schijf schrijven, dan schrijf ik dat ook nog naar een aantal andere schijven in een cluster. Dus ik heb altijd een aantal kopieën van mijn informatie. Je houdt bij waar die kopieën zich bevinden en dat kan ongelimiteerd schalen. Dus als je merkt dat deze schijf kapot is, en ik heb hier twee kopieën, dan herstel ik terwijl alles doorgaat in de achtergrond die andere kopie weer. Dus je hoeft nooit back-ups te maken, want alles is altijd beschikbaar."
Dat is HDFS, een distributed filesysteem, en onderdeel van Hadoop. Verder zit daar een programmeeromgeving op, genaamd MapReduce. "Stel je wilt weten wat de gemiddelde leeftijd van mensen in Nederland per gemeente is", poneert Oosterhof om uit te leggen wat je met MapReduce kunt doen. "Maar die data liggen allemaal verspreid over honderden nodes. MapReduce stuurt dan hetzelfde verzoek naar al die servers toe en krijgt dan van elk een antwoord terug. Dat is het mappen. Vervolgens komt het reduceren, het aggregeren van de resultaten. Dat moet heel snel gaan als je honderden of duizenden nodes hebt. Elke query is heel snel."
Voor Informatica is Hadoop een van de bronnen en doelen om naar toe te schrijven. Zo helpen zij klanten die data vanuit systemen naar Hadoop willen brengen of uit Hadoop willen halen om ze te integreren in hun datawarehouse. "Wat we ook aan het doen zijn, is dat we de logica die we hebben in ons transformatieproces niet zelf uitvoeren maar in MapReduce laten uitvoeren over al die nodes heen. Dat hoef je dan niet te programmeren. Je kunt de logica in de Informatica-repository zetten, waarna wij de code genereren om het uit te voeren op het Hadoop-platform. Dat is dus weer een volgende stap."
Gevraagd naar het integratieaspect in dit voorbeeld, schetst Oosterhof: "Het is zeker een vorm van integratie, want alle data die naar het Hadoop-platform gaan, moeten gecontroleerd worden, moeten geaggregeerd worden en moeten geïntegreerd worden met data in een MDM-systeem of in een DWH-systeem. Als ik een Twitter-bericht uitstuur van 'die service van ING klopt niet, ik heb dit en dit probleem', en ik ben bij ING een kleine klant, dan heeft dat voor hen misschien andere gevolgen dan als ik directeur ben van Shell en al mijn financiële zaken via ING afhandel. In dat geval zullen ze snel actie willen ondernemen. Dat heeft invloed, dus hebben ze die informatie nodig, want je moet het koppelen aan: wie is die persoon, wat is zijn netwerk, wat is de relatie tussen ons en die klant enzovoort. Dus is het een data-integratieprobleem en treden wij in het strijdperk."

Dick Schievels is hoofdredacteur van Database Magazine.