Trends in datawarehousing 2007

Datawarehousing in de derde era

Paul van der Linden

Volgens IDC Research beweegt de BI-markt zich in cycli van 15 jaar. Tussen 1975 en 1990 lag de nadruk voornamelijk op rapporten maken op mainframes en was slechts een beperkt aantal statistische softwarepakketten voorhanden. In de tweede periode van 15 jaar (1990-2005), door IDC de 'modern era' genoemd, lag de nadruk op client/server based BI en uiteindelijk bij query, rapportage en analyse via een webbased architectuur.

Nu zouden we dus in de derde cyclus zitten (2005-2020). IDC is van mening dat er twee ontwikkelingen zijn die deze derde era kenmerken. Op de eerste plaats de verdere verspreiding van BI naar andere gebruikers binnen en buiten de organisatie. Maar ook het verschuiven van de traditionele focus op rapporteren naar het centraal stellen van de besluitvorming. Als dat zo is, wat zijn dan de uitdagingen in datawarehousing waarmee we in de derde era te maken krijgen?

Uitdagingen

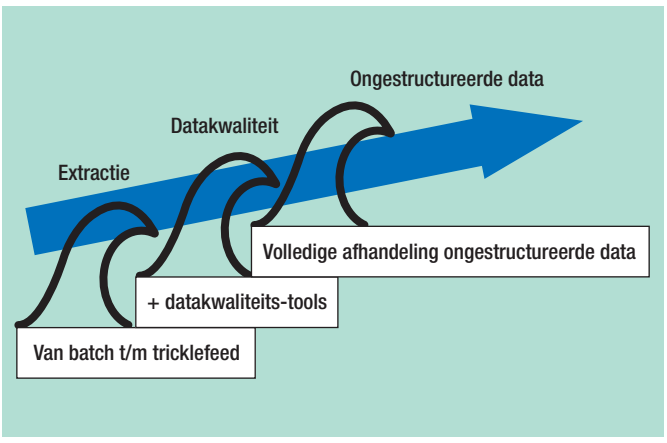
Om die derde cyclus succesvol in te kunnen vullen zal een hele waaier aan datawarehousing-uitdagingen gepareerd moeten worden:

- Mixed workloads: het nieuwe datawarehouse moet kleine, grote, urgente en minder urgente dataloads aankunnen. Aankunnen betekent hier meer dan technisch mee om kunnen gaan, het betekent ook in alle gevallen optimaal (in termen van tijd en geld) er mee omgaan;
- Gestructureerde en ongestructureerde data: het totale data-volume neemt nog steeds toe. Het snelst groeiende aandeel bestaat uit ongestructureerde data. Het datawarehouse moet dus met ongestructureerde data kunnen omgaan en deze data gecombineerd met gestructureerde data kunnen aanbieden;
- Meer data langer bewaren; het datawarehouse moet steeds grotere volumes data hanteren waarbij schaalbaarheid dus centraal staat;
- Datakwaliteit was uiteraard altijd al belangrijk. Maar als data verspreid worden naar steeds meer mensen (binnen en buiten de organisatie) is datakwaliteit nog belangrijker. Regel- en wetgeving (Sarbanes-Oxley) stellen eveneens hogere eisen aan de datakwaliteit.

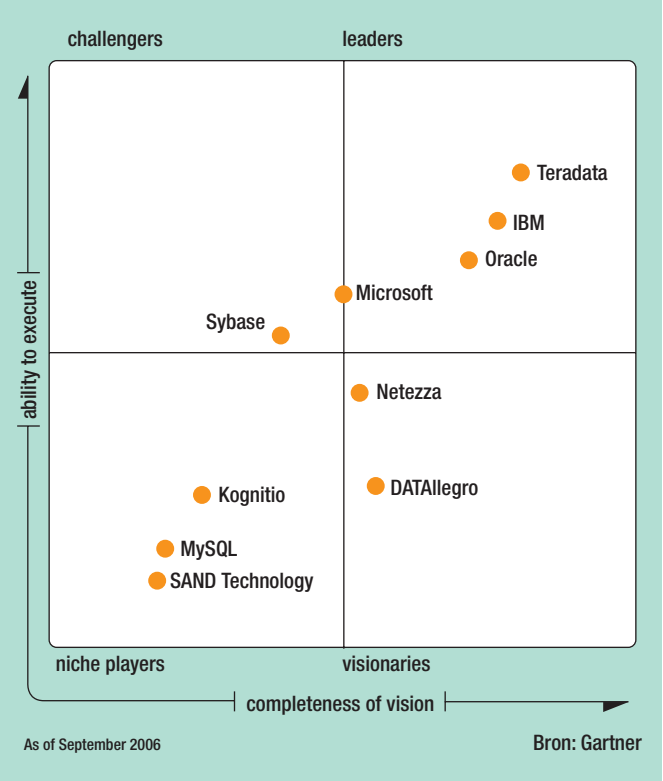
Vanuit de geschetste uitdagingen kan de vertaling worden gemaakt naar wat dit betekent voor de hoofdcomponenten van datawarehousing: de datawarehouse database en de data-integratie software.

Data-integratie

De ontwikkeling in data-integratie doet zich in drie golven voor. De eerste, reeds te signaleren trend heeft te maken met de snelheid waarmee data kunnen worden verwerkt en aangeboden. Hierbij hebben we het over het bieden van de hele range tussen batch-gewijze aanlevering en messaging (trickle feed). De tweede trend heeft te maken met de geboden validatiefunctionaliteit. Deze tweede golf is net ingezet. Data-integratie wordt uitgebreid met verdergaande datakwaliteit. Datakwaliteits-tools worden opgenomen in het data-integratieplatform. Ook een derde golf ligt voor de hand: de extractie van niet alleen gestructureerde data (in welk formaat dan ook opgeslagen), maar ook van ongestructureerde data. Hoe moeilijk dat gaat worden en hoe oplossingen eruit gaan zien hangt af van ontwikkelingen aan de database-kant (zie ook bij databases).



Afbeelding 1: Trends in data-integratie.



Afbeelding 2: Gartner Magic Quadrant DWH databases.

De huidige data-integratiemarkt is ontstaan door het samengaan van de markt voor ETL-tools, voor Enterprise Information Integration (EII) en alle vormen van datatransport tussen batchgewijze aanlevering en messaging in (de eerste golf). De nieuw ontstane data-integratiemarkt is daardoor nog sterk in beweging en zeer gevarieerd van samenstelling. Door de overname van Ascential, halverwege 2005, heeft IBM zich een vooraanstaande plek verworven in deze markt. In deze markt bewegen zich daarnaast ook iWay (Information Builders) dat de meest uitgebreide set van connectors biedt, maar ook SAS, SUN, Microsoft en de 'ouwe getrouwe' Embarcadero en ETI. Kortom: een zeer gemêleerd gezelschap van rijp en groen.

Verschillende leveranciers hebben inmiddels ook datakwaliteitstools toegevoegd aan hun data-integratieplatform (tweede golf). Voorbeelden hiervan zijn Business Objects, IBM en Informatica. Business Objects heeft hiervoor Firstlogic overgenomen. Informatica deed in januari vorig jaar hetzelfde met Similarity Systems. De eerder genoemde overname van Ascential door IBM heeft ProfileStage, QualityStage en DataStage aan WebSphere toegevoegd. Verwacht mag worden dat ook de overige leveranciers van data-integratieplatformen zullen volgen met het toevoegen van specifieke datakwaliteitssoftware aan hun data-integratieplatform.
Een data-integratieplatform dat niet is toegerust op ongestructureerde data heeft slechts een beperkt nut. Gezien het feit dat de datagroei nu juist in ongestructureerde data zit is het daarom essentieel dat ook deze bronnen ontsloten kunnen worden. Vanuit data-integratie geredeneerd is de meest eenvoudige oplossing dat ongestructureerde data worden aangeboden als gestructureerde data. Met andere woorden: dat voor de data-integratieslag al de omzetting van ongestructeerd naar gestructureerd heeft plaatsgevonden. Of dit ook de oplossingsrichting is die door de sector wordt gekozen is nog onduidelijk. Hiermee is ook nog niet helder hoe een data-integratieplatform in de derde golf uitziet. Maar dat er een oplossing voor ongestructureerde data voorhanden moet zijn moge duidelijk zijn. Zowel Informatica als IBM (data-integratieleiders in Gartner's Magic Quadrant) hebben een oplossing voor ongestructureerde data. Vanaf PowerCenter 7 kan gebruik worden gemaakt van Informatica's Unstructured Data Option. Met IBM's WebSphere DataStage kunnen eveneens tekstfiles worden afgehandeld. Marktanalist Gartner duidt het nieuwe activiteitenveld, dat betrekking heeft op het kunnen omgaan met gestructureerde, semigestructureerde en ongestructureerde data en het omzetten hiervan naar zinvolle en volledige infomatie, aan als Enterprise Information Management (EIM).

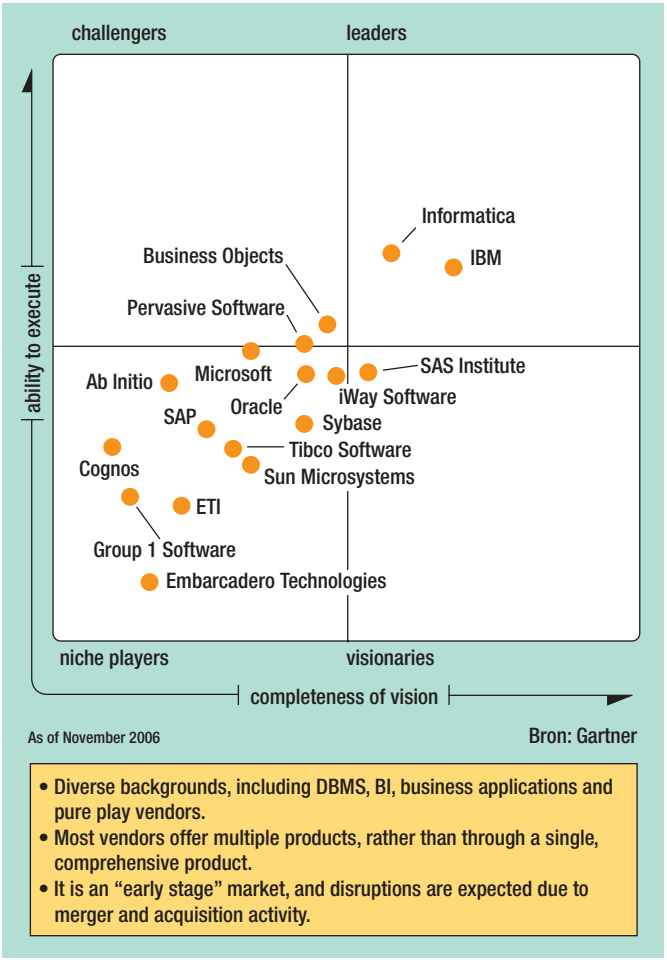
Databases

De database van tien jaar geleden en de huidige versie daarvan hebben weinig met elkaar gemeen. Oorspronkelijk zijn databases ontworpen met het vastleggen en afhandelen van transacties (records) in gedachten. Dat was ook de reden dat het relationele model zoals gepropageerd door Chris Date en Ted Codd opgeld deed. De implementaties van dat relationele model in echte softwareproducten was dan misschien niet helemaal zuiver, maar vergeleken met de huidige databases mogen we die achteraf toch wel als relationeel bestempelen. Sinds die tijd zijn een heleboel extensies toegevoegd die weinig te zoeken hebben in een relationeel model. Denk bijvoorbeeld aan de blob (binary large object) of meer recentelijk het verdwijnen van de kubus in de database. Voorbeelden van dit laatste zijn IBM's Cube Views, Oracle's Analytical Workspace en Microsoft's Analysis Services. De betreffende databases kun je niet meer met goed fatsoen relationeel noemen.

De trend is het verdwijnen van de kubus

De trend die hier kan worden waargenomen is het verdwijnen van de kubus. Behalve dat de kubussen 'bij bosjes' in de database verdwijnen speelt ook de opkomst van de 64-bit architectuur hier een rol. Waar veel data, dankzij 64-bit architectuur, in geheugen kunnen worden gelezen verdwijnt de behoefte om fysieke kubussen aan te maken. Het voordeel van virtuele kubussen is immers dat je die niet hoeft te beheren en dat ze in een mum van tijd kunnen worden gemaakt, aangepast en naar behoefte ook weer verdwijnen. De recente overname van Hyperion door Oracle doet dan ook onder andere de vraag ontstaan wat het lot is van de befaamde Essbase-kubus.

De database wordt momenteel met twee uitdagingen geconfronteerd. De eerste uitdaging is hoe om te gaan met steeds meer data. In wezen is dit de vraag naar de schaalbaarheid van de huidige databases. De tweede vraag is enigszins ingewikkelder, want dit betreft de vraag hoe om te gaan met de exponentieel groeiende berg van ongestructureerde data.
Schaalbaarheid houdt in dat de prestaties van de database niet achteruitgaan als de hoeveelheid data of het aantal gelijktijdige gebruikers toeneemt. Voor gebruikers van het systeem zou dit ondoorzichtig moeten zijn. Dat betekent dat van de geleverde prestaties niet is af te leiden of het gaat om veel dan wel weinig data en/of gelijktijdige gebruikers. Kijkend naar de DWH DBMS servers is volgens marktanalist Gartner Teradata nog steeds de onbetwiste leider. Teradata maakt gebruik van X-86 technology en een MPP-topologie. Volgens CTO Stephen Brobst is het wijdverbreide verhaal dat Teradata alleen maar een plaats zou hebben in de meeste high-end datawarehouses een fabeltje. Juist door gebruik van X-86 technologie is het mogelijk om klein te beginnen en vervolgens extra CPU's bij te prikken. Hoe dan ook, gegeven de uitdijende datavolumes is het inmiddels zo dat organisaties voor wat hun behoefte aan dataopslag en -manipulatie betreft naar Teradata toegroeien. Teradata's grootste uitdaging ligt misschien niet zozeer bij concurrenten als HP en IBM, maar meer



Afbeelding 3: Gartner Magic Quadrant data-integratie tools.

of ze een uitstekend technisch product kunnen combineren met een even goede marketing ervan.
Hoe om te gaan met de steeds grotere berg ongestructureerde data wordt een vraag die steeds meer organisaties zich stellen. Vanuit verschillende invalshoeken (webcontent, document management, afbeeldingen etcetera) bestaan al oplossingen die in de meeste gevallen echter slechts een deel van de vereiste functionaliteit bieden. Het kunnen opslaan en terugvinden van ongestructureerde data is één ding, het extraheren van structuren en verbanden of kunnen bevragen ervan een hele andere.

> Van de geleverde prestaties is niet af te leiden of het gaat om veel dan wel weinig data

IBM's OmniFind software is gebaseerd op de Unstructured Information Management Architecture (UIMA) en biedt de mogelijkheid om gestructureerde en ongestructureerde data tegelijkertijd te analyseren. Concepten (entiteiten), feiten en relaties kunnen met behulp van OmniFind uit de ongestructureerde data worden 'ontdekt'. Denk bij ongestructureerde data aan alle informatie die is opgesloten in tekst, beeld, audio en video.
In Microsoft's SQL Server kan met het datatype ntext, text en image een enorme hoeveelheid data (tot en met 2 GB) per waarde worden opgeslagen. Ook wordt de mogelijkheid geboden voor text mining. Zo zit er een text mining-algoritme in SQL Server Integration Services waarmee de ongestructureerde data geanalyseerd kunnen worden.

Door vrije tekst te analyseren worden sleutelwoorden bepaald die vervolgens gebruikt worden om te clusteren en te categoriseren. Oracle claimt met Enterprise Content Management het eerste platform te bieden waarop document management, webcontent management, digital asset management en records management worden gecombineerd. Met de overname van Stellent in november 2006 heeft Oracle in ieder geval een serieuze boost gegeven aan zijn content management-oplossing.

Datawarehousing done differently

De boven beschreven trends hebben met name betrekking op datawarehousing 'as we know it'. Er zijn nog andere ontwikkelingen die hier genoemd moeten worden omdat ze impact kunnen hebben op de toekomstige ontwikkeling van de datawarehouse-markt.
Datawarehouse appliances. Netezza en DATAllegro zijn zogenaamde Datawarehouse (DWH) Appliances. Deze appliances bieden een beperktere datawarehousing-functionaliteit tegen een veel lagere prijs. Dat doen ze door gebruik te maken van commodity processoren, harde schijven en CPU's. In het geval van DATAllegro heb je te maken met de open source Ingres database onder Linux. Bij Netezza, ook een DWH appliance, wordt gebruik gemaakt van een eigen database onder Linux. De

combinatie van goedkoop, eenvoud en snelheid maakt de datawarehouse appliances bij uitstek een productcategorie om nauwlettend in de gaten te houden. De Netezza Snippet Server claimt een query performance die tussen 20 en 100 maal sneller is. Centraal hierbij staat de Snippet Processing Unit (SPU) die bestaat uit een FPGA (field programmable gate array) en een PowerPC-processor. FPGA's worden veel gebruikt voor het omgaan met streaming data (video, audio, graphics). Belangrijk is dat de processor dicht tegen de disk zit en dat de I/O via de FPGA loopt. De FPGA fungeert als zelfregulerende controller. Een disk heeft drie partities van 300 GB. De kleinste NSP heeft 28 SPU's. Netezza mag onder andere Wolters Kluwer, Orange, T-Mobile en Flora Holland tot haar klantenkring rekenen.

Datawarehouse management. Het gaat hier om de categorie leveranciers (Kalido, Dynalytical of het Nederlandse BIReady) die software biedt waarmee vanuit een model een datawarehouse wordt gegenereerd. Kalido lijkt met een prijspolitiek geënt op alleen de allergrootste bedrijven (wereldwijd) in Nederland de boot te hebben gemist. Het aantal installaties lijkt niet meer significant toe te nemen. Dat heeft het Britse bedrijf dan uitsluitend aan zichzelf te danken. Natuurlijk heeft de IT-afdeling in de meeste gevallen een ongezonde aversie getoond tegenover dit soort software. Een torenhoge prijs heeft niet geholpen om door deze weerstand heen te breken. Of BIReady hetzelfde lot is beschoren is niet zomaar gezegd. In tegenstelling tot Kalido heeft men hier wel in de gaten dat een lage prijs helpt bij adoptie. De tekenen dat BIReady ook internationaal interesse genereert zijn er in ieder geval. Wie nog niet over BIReady heeft nagedacht: nu is er de kans.

Übertrend

Achter de trends in data-integratie en databases zit dezelfde drijvende kracht die ook in Business Intelligence te zien is. Deze zogenaamde *übertrend* bestaat uit de opkomst van goedkopere, snellere en eenvoudiger (intuïtiever, gemakkelijkere) software, gevoed door nieuwe technologische toepassingen. Datawarehouse management software en datawarehouse appliances zijn dan de bepalende krachten. En net zoals in BI zal dit leiden tot een veel eenvoudiger IT-landschap met veel minder componenten (databases, servers, kubussen etcetera). Een ontwikkeling die voor het einde van de derde datawarehouse-cyclus zijn beslag zal krijgen.

Conclusie

Volgens IDC ontwikkelt de BI-markt zich in cycli van 15 jaar. Inmiddels zijn we begonnen aan de derde cyclus, die tussen 2005 en 2020 loopt. De kenmerken van BI in deze derde cyclus zijn: meer mensen aansluiten op BI (intern en extern); en de omslag maken van rapportage-centrisch naar besluitvorming centraal. Om dit voor elkaar te krijgen is de "combinatie van data-integratie, datakwaliteit, masterdata management en datawarehousing nodig om een schaalbaar en flexibel platform te krijgen dat de verschillende eindgebruikers-tools en applicaties ondersteunt" (quote van IDC).

Datakwaliteit

De markt voor datakwaliteits-tools is met 300.000,- USD aan jaarlijkse licenties (gegevens Gartner, april 2006) een relatief beperkte markt. Aangescherpte wet- en regelgeving als Sarbanes-Oxley en Basel-II hebben de nadruk op datakwaliteit verder versterkt. De markt is echter nog sterk gefragmenteerd en nog zeker niet uitgekristalliseerd. In het verleden heeft datakwaliteit vooral betrekking gehad op klantdata. Oplossingen voor andere soorten data (bijvoorbeeld productdata) worden nu ook steeds meer geboden. Datakwaliteit heeft minimaal betrekking op standaardisatie, parsing, profiling, matching en cleansing. Hier blijkt al dat er vele raakvlakken zijn met ETL-tools en in een breder verband: data-integratie. Tot de bekendste partijen op het gebied van datakwaliteits-tools behoren Trillium, DataFlux (SAS), Firstlogic (Business Objects) en IBM (door de overname van Ascential; nu opgenomen in de WebSphere-familie). Het in Arnhem gevestigde Human Inference mag uiteraard ook niet onvermeld blijven, hoewel Gartner dit bedrijf (net als Informatica) bij de visionaire bedrijven plaatst.

Op het vlak van data-integratie is de ontwikkeling gemaakt van ETL-tool naar data-integratieplatform (alle vormen van data-integratie). De tweede golf die nu wordt gezet is om datakwaliteitstools toe te voegen aan het data-integratieplatform. Waarna ook de stroom van ongestructureerde data verwerkt moet kunnen worden (derde golf). In termen van Gartner's Enterprise Information Management (EIM) zijn we echter nog niet zo ver. Voorlopig hebben leveranciers de handen vol aan de eerste twee golven.
Op database-gebied gaat het om schaalbaarheid (groeiende datavolumes) en wederom het omgaan met ongestructureerde data. Het raakvlak met data-integratie ligt ook precies hier. Als de oplossing is om ongestructureerde data om te zetten in gestructureerde data (en dat is nog de vraag), zal bepaald moeten worden waar dit dan zal plaatsvinden. Dat kan in de aanleverende databron, in het datawarehouse, maar ook elders.
Tenslotte ontstaan ook nieuwe oplossingen zoals datawarehouse appliances. Deze leveren tegen veel lagere kosten en bij geringere complexiteit een beperkter deel van de functionaliteit. DWH management tools zoals van Kalido en BIReady kunnen het datawarehouse-proces ook verder versterken, doordat ze vanuit een modelmatige aanpak het datawarehouse genereren. Beide ontwikkelingen (datawarehouse appliances en DWH management tools) zijn onderdeel van de 'übertrend': de trend achter de trend. Goedkopere, snellere en eenvoudigere software leidt tot een veel eenvoudiger IT-landschap met veel minder componenten (databases, servers, kubussen etcetera). Een ontwikkeling die nog wel wordt voltooid voor het einde van de derde datawarehouse-cyclus.

Paul van der Linden (Paul.PFH.vanderLinden@AtosOrigin.com) is senior consultant Data Warehousing/BI bij Atos Origin en geeft leiding aan Data Warehousing Cost & Lifecycle Management (CLM).