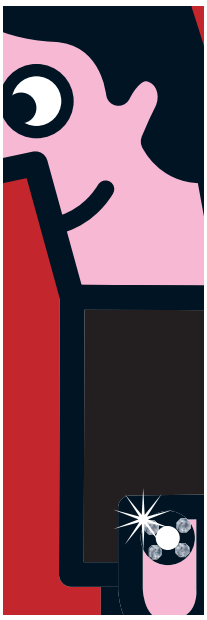


Robuuste oplossingen om in de gaten te houden

Trends in Open Source BI

Jos van Dongen

Het is u wellicht ontgaan, maar op 20 maart van dit jaar gebeurde er iets bijzonders in Nederland. Voor de eerste keer werd er tijdens het Database Systems congres plaats ingeruimd voor open source ontwikkelingen, in dit specifieke geval voor OS databases.

Ook binnen andere gremia begint langzaam het besef door te dringen dat open source BI software een zinvol alternatief kan zijn. Zo werd de ETL-matrix verrijkt met maar liefst twee OS ETL-producten, en worden commerciële leveranciers steeds vaker geconfronteerd met een OS 'tegenstander' in selectie-trajecten. Het afgelopen jaar is er in dit blad in vrijwel elk nummer aandacht besteed aan OS BI software. Het laatste nummer van dit jaar biedt dan ook een uitgelezen gelegenheid om eens terug èn vooruit te kijken en te beoordelen of al deze aandacht terecht is. De welbekende BI-piramide, die van onder naar boven uit de onderdelen ETL, datawarehouse, rapportage, analyse en dashboarding bestaat, levert hiervoor een mooie indeling.

ETL

In DB/M3 2007 is uitgebreid aandacht besteed aan de twee meest serieuze spelers in dit veld, K.E.T.T.L.E. en Talend. Deze oorspronkelijke namen zijn steeds meer op de achtergrond aan het raken aangezien Kettle als onderdeel van Pentaho is omgedoopt in Pentaho Data Integration, en Talend als onderdeel van het Jaspersoft platform in JasperETL. Het is interessant om te zien hoe beide producten zich ontwikkelen.

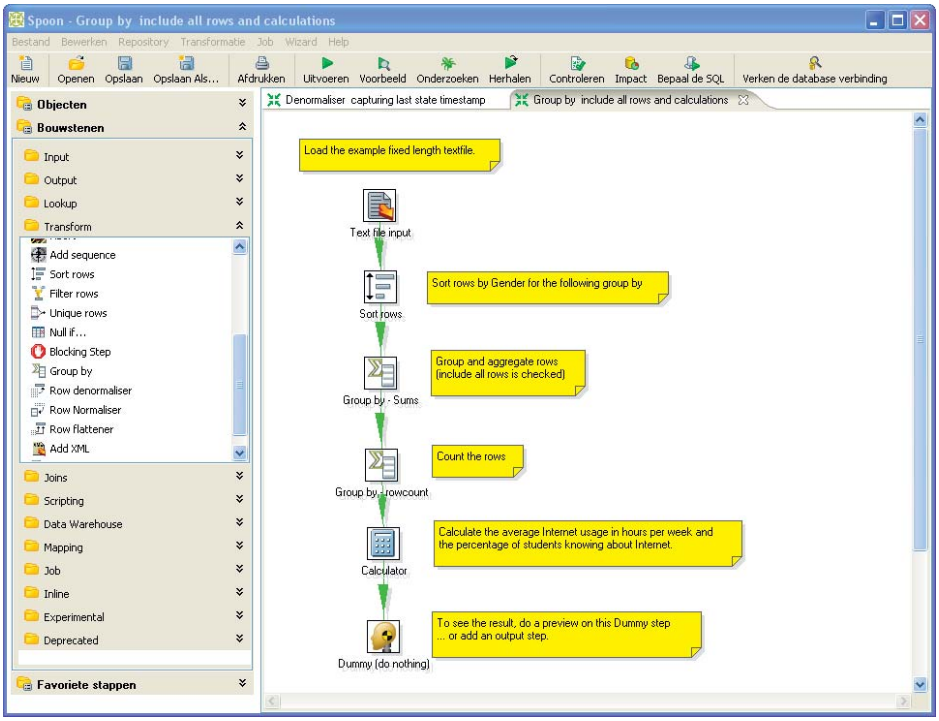
Talend positioneert zichzelf steeds meer als 'high-end' oplossing door een additionele 'Integration Suite' te bieden bovenop de standaard 'Open Studio'. De integration suite biedt voorzieningen voor team-ontwikkeling, parallelle en gedistribueerde verwerking van jobs en coördineren en monitoren van deze jobs via een webbased dashboard. De meest in het oog springende toe-voeging is echter Talend On Demand, een SaaS (Software as a Service) oplossing waarmee het mogelijk wordt om (gratis!) gebruik te maken van een door Talend gehoste centrale reposi-tory. Dit biedt team-ontwikkelmogelijkheden die over de grenzen van de eigen organisatie heen gaan, zodat bijvoorbeeld het ontwikkel- en testwerk door derden uitgevoerd kan worden. Acceptatie en uitvoering kunnen vervolgens binnen de eigen organisatie gebeuren zonder dat er iemand achter zijn of haar eigen PC vandaan hoeft te komen. Zo draagt Talend zelfs nog bij aan het terugdringen van de files.

Kettle heeft ondertussen release candidate 2 van versie 3 uitge-bracht, een flinke upgrade van de in mei besproken versie. Met name de interface is onder handen genomen (zie afbeelding 1) waardoor het werken met het pakket een stuk intuïtiever is geworden. Er is ook een flink aantal nieuwe mogelijkheden toe-gevoegd, en bestaande opties zijn uitgebreid. Een aantal features bevindt zich nog in het experimentele stadium (bijvoorbeeld de Oracle Bulk Loader en de Mondrian input), maar het geheel ziet er zeer compleet uit. Ook in de architectuur zijn er enkele verbeteringen aangebracht waardoor Kettle in de volgende release van de ETL-matrix extra punten gaat halen op het gebied van clustering, partitionering en parallellisatie.

Datawarehouse

De ontwikkelingen op het gebied van OS databases staan intussen ook niet stil, hoewel een en ander soms wat minder vlot verloopt dan men denkt. Zo is MySQL een aardig tijdje zoet geweest met versie 5.1, waarvan binnenkort de eerste 'release candidate' het levenslicht gaat zien. De grootste verbeteringen ten opzichte van de huidige versie 5.0 zijn partitioning en row-level replicatie. Met name de partitioning-mogelijkheden zullen een welkome aanvulling zijn in datawarehouse-land. MySQL 5.1 biedt range, list, hash èn key partitioning, waarbij de range en list partities weer subpartities van type hash of key kunnen bevatten. Ook Ingres begint zich de laatste tijd nadrukkelijk te manifesteren. Aan het eind van dit artikel leest u daar alles over.

De nationale trots MonetDB beleefde in juni van dit jaar de



Afbeelding 1: Kettle 3 interface.

officiële release van versie 5 die zoals al eerder aangehaald een onovertroffen performance haalt. PostgreSQL heeft zojuist beta 1 van versie 8.3 uitgebracht, met als belangrijkste verbeteringen de integratie van de full-text search engine in de kernel, en de ondersteuning van SQL/XML inclusief de toevoeging van een XML datatype.
Een relatief nieuw product is LucidDB (www.luciddb.org), dat net als MonetDB (zie ook pagina 32) een kolomgebaseerde opslag kent. Het unieke aan LucidDB is dat het tevens een ETL-oplossing biedt, niet als aparte oplossing, maar volledig geïntegreerd in het product zelf. Er is tevens een naadloze integratie met Mondrian bewerkstelligd, waardoor een high performance (R)OLAP-oplossing gerealiseerd kan worden.
De belangrijkste twee vragen voor toepassing in een datawarehouse omgeving zijn echter enerzijds de functionaliteit die de producten bieden, en anderzijds (en misschien nog wel belangrijker) de uiteindelijke performance. In afbeelding 2 is allereerst te zien dat de OS databases nog een lange weg te gaan hebben als het puur om de functionele volledigheid ten behoeve van een datawarehouse-omgeving gaat.

Het opvallendste punt is dat één van de grootste performance boosters, de materialized view (in combinatie met query rewrite-functionaliteit van de optimizer) nog door geen enkele OS-leverancier is geïmplementeerd, terwijl de commerciële leveranciers dit kunstje toch al een behoorlijk aantal jaren beheersen. Nu zijn trucs als bitmapped indexing, materialized views en partitioning natuurlijk prachtig, maar deze toevoegingen vereisen uiteraard wel een vaardige DBA èn de nodige tijd om een en ander te implementeren en te onderhouden. Hier komt een ander, relatief nieuw, verschijnsel om de hoek kijken: de datawarehouse appliance. Waarom wordt dat in dit kader genoemd? Welnu, de meest bekende spelers draaien nu eenmaal op een OS database: DATAllegro werkt met Ingres, en Netezza en Greenplum werken op basis van PostgreSQL. Uiteraard is er flink aan deze producten gesleuteld, maar als u ooit een Netezza-kast door 600 miljoen records hebt zien ploegen en binnen enkele seconden met het query resultaat op de proppen heeft zien komen, denkt u wellicht toch anders over de performance van OS databases.

Ook zonder appliance staat een aantal OS-producten zijn mannetje. Om dit inzichtelijk te maken is de TPC-H benchmark (zie www.tpc.org voor meer informatie) losgelaten op een zestal producten, waarvan 1 commercieel. Welke dat is mag hier overigens om licentietechnische redenen niet vermeld worden. Het gaat in alle gevallen om een 'plain vanilla' installatie, dus zonder extra optimalisaties, op een Windows 2003 server machine met een AMD Athlon64 3500+ processor met 2 GB intern geheugen en een enkele snelle 320 GB harddisk. De weergegeven tijd is de gemiddelde responstijd van de eerste tien query's uit de TPC-H benchmark. Er wordt gebruik gemaakt van een 'scale factor 2' (2 GB) database, zie afbeelding 3.
Kortom, klagen uw gebruikers over de performance van een datamart, kijk dan eens naar MonetDB. Nog een bijkomend voordeel: in afbeelding 2 is te zien dat MonetDB niet over een bulkloader beschikt. Het laden van data uit flat files met behulp van het eigen 'copy' commando gaat echter sneller dan met welke bulkloader in welke database dan ook. Features zeggen dus lang niet alles!

Rapportage

In DB/M2 2007 was de conclusie over OS reporting producten niet bepaald juichend. Met name het ontbreken van de SQL abstractielaag werd als groot gemis ervaren. Op dit vlak is er gelukkig goed nieuws: Pentaho beschikt sinds de laatste release (1.6) over een volwaardige metadata-laag die is gebaseerd op het OMG Common Warehouse Model (CWM). Dit laatste is belangrijk vanwege de uitwisselbaarheid met andere omgevingen die ook deze standaard ondersteunen, waardoor u enerzijds niet aan Pentaho vastzit, en anderzijds makkelijk vanuit een andere CWM-omgeving kunt migreren. De metadata-laag wordt in verschillende lagen opgebouwd en beschikt tevens over functionaliteit om meertaligheid te ondersteunen en uniforme opmaak van data-elementen te waarborgen, zie afbeelding 4.

Omdat Pentaho nu over een fatsoenlijke metadata-laag beschikt is het uiteraard ook mogelijk geworden om ad hoc query-functionaliteit toe te voegen aan de BI suite. Deze gaat gelukkig al een stuk verder dan het simpelweg data-elementen in een lijstje slepen (zoals bij voorheen bij JasperSoft) en biedt voor de niet te veeleisende gebruiker een goed startpunt voor het maken van rapporten. De met de ad hoc query builder gemaakte rapporten kunnen vervolgens verder worden ontwikkeld binnen de uitgebreidere report builder, die inmiddels ook over sub-reporting en multi-source capaciteiten beschikt. Deze laatste twee mogelijkheden zijn sinds de laatste release (2.1) ook beschikbaar in JasperReports. Jaspersoft heeft ook de ad hoc report-mogelijkheden behoorlijk aangepakt, blijkens het feit dat het nu mogelijk is om met *drag and drop* (AJAX) custom formules, grafieken, tabellen, kruistabellen en een combinatie hiervan een rapport in elkaar te kleien. Helaas gaat dit nog steeds niet op basis van een SQL abstractielaag, maar een ontwikkelaar kan wel standaard query's klaarzetten voor gebruikers, waarna deze (een selectie uit) de resultaatset kunnen gebruiken in de ad hoc interface. Ook heeft Jaspersoft werk gemaakt van de standaard visualisatiecomponenten, waarover later meer.

De meest interessante ontwikkeling op het gebied van OS reporting vindt u echter niet in de BI suites, maar in OpenOffice! Op basis van jFreeReports (Pentaho) heeft Sun een report builder ontwikkeld die als gratis component beschikbaar is voor OpenOffice versie 2.3 en hoger. Deze beschikt over een Access-achtige query builder, ruime opmaakmogelijkheden en een standaard output naar keuze in Calc (Spreadsheet) of Writer (Tekstverwerker). Hiermee wordt OpenOffice in één klap een serieus alternatief voor de commerciële Office-varianten.

Analyse & dashboards

In DB/M5 2007 is al uitgebreid stilgestaan bij OS analytics en sinds die tijd is er niet veel schokkends gebeurd. Op het gebied van dashboarding zijn er wel enkele goede ontwikkelingen gaande. Jaspersoft lijkt in zijn nieuwste versie iets meer mogelijkheden te bieden dan Pentaho door de integratie en ondersteuning van JSR-168 Portlets en door (net als bijvoorbeeld Cognos) dashboard-achtige functionaliteit onder te brengen in de report builder. Met deze aanpak kunnen redelijk gemakkelijk interactieve dashboards worden ontwikkeld zonder dat er, zoals bij Pentaho, diepgaande JSP kennis voor nodig is. Toch kan het gebodene nog (lang) niet tippen aan hetgeen de leidende closed source vendors te bieden hebben.

Conclusie en vooruitblik

Eén conclusie die inmiddels is gerechtvaardigd is dat het zeer wel mogelijk is om op basis van open source software een volwaardige en volwassen BI-oplossing te realiseren. Ook wordt dit steeds gemakkelijker gemaakt door de 'all-in-one' installatie

	MySQL	PostgreSQL	Bizgres	MaxDB	Ingres	Firebird	MonetDB
Analytical fct (sql2003)	Alleen rollup	nee	nee	nee	nee	nee	Roadmap
Partitioning	5.1	Range/list	Range/list	nee	ja	nee	nee
Materialized Views	nee	nee	Greenplum	nee	nee	nee	nee
Bitmap index	nee	Aleen scan	Alleen scan	nee	nee	nee	nee
Full-text index/search	Aleen Mylsam	extentie	extentie	nee	nee	nee	Roadmap
Clustering	Aleen NDB	ja	ja	ja	ja	nee	nee
Replicatie	ja	Slony	Slony	ja	ja	nee	nee
Bulkload	ja	Copy	Enh copy	ja	ja	Via Ext.File	Copy

Afbeelding 2: BI/DWH functionaliteit OS databases.

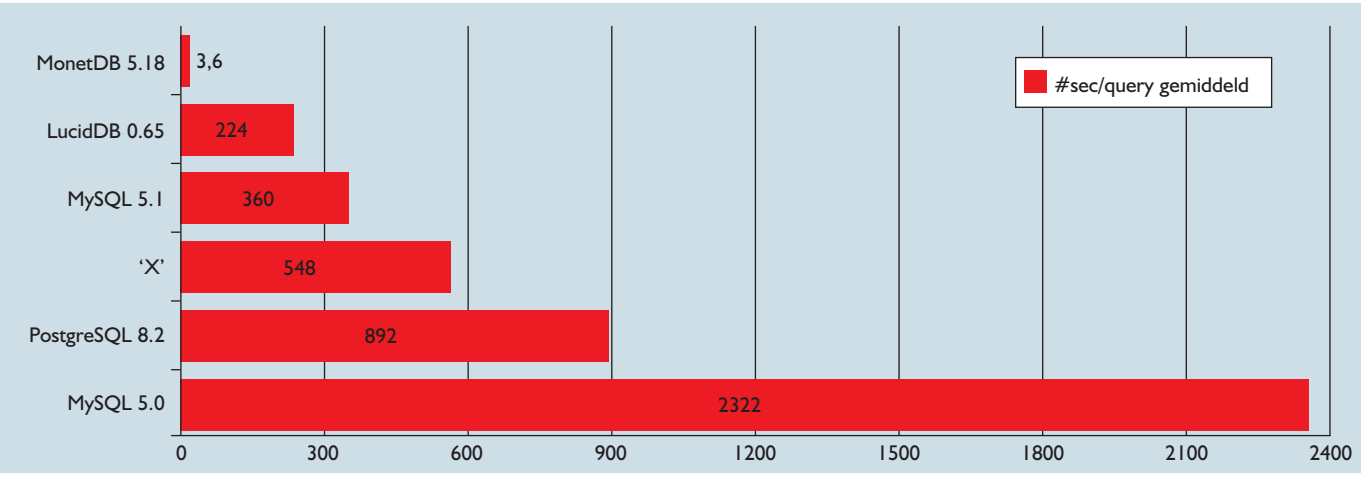
van Pentaho of de door Ingres geleverde BI appliance op basis van Jaspersoft. Dus voor wat betreft de standaard onderdelen in een BI suite moeten de traditionele leveranciers langzamerhand op hun tellen gaan passen. Dat is waarschijnlijk ook één van de redenen dat partijen als Cognos en Business Objects al langere tijd bezig zijn om de bakens te verzetten richting 'performance management', met voorzieningen voor scorecarding, planning, budgetting en consolidatie en het leveren van verticale oplossingen. Op al deze terreinen (en er zijn er nog wel meer te noemen, zoals bijvoorbeeld datakwaliteit) zijn er nog maar mondjesmaat OS alternatieven beschikbaar. Eén voorbeeld hiervan, en helaas nog een uitzondering, is Adaptive Planning (www.adaptiveplanning.com) dat een complete OS CPM suite biedt.

Als we kijken naar de marktacceptatie van alle beschreven onderdelen is er wel een nuancering aan te brengen. Niemand kijkt meer vreemd op bij de inzet van PostgreSQL of MySQL als database, maar een OS ETL-tool wordt nog met de nodige argwaan bekeken. Op het gebied van de OS BI Suites is er wel het een en ander aan het veranderen. Zowel Pentaho als Jaspersoft timmeren stevig aan de weg, beide met recent geopende verkoopkantoren in Europa, diverse trainingen die ook in verschillende Europese steden gevolgd kunnen worden en een groeiend netwerk van partners en system integrators. Zelfs in Nederland vindt u al een gecertificeerde Pentaho partner. 2008 zou dan ook wel eens het jaar van de doorbraak kunnen worden voor OS BI, zeker bij organisaties die nog aan het begin van de BI-levenscyclus staan. Denk hierbij met name aan organisaties in de non-profit sector zoals gemeenten en zorginstellingen. Ook bedrijven die al een jaar of vijf (en soms al veel langer) bezig zijn en nu voor de keus staan om voor veel geld te upgraden naar een nieuwe release van de bestaande leverancier zouden zich wel eens kunnen bedenken en voor een voordeliger OS alternatief kunnen gaan. In elk geval komen er boeiende tijden aan: over een jaar zullen we zien wat hiervan terecht is gekomen.

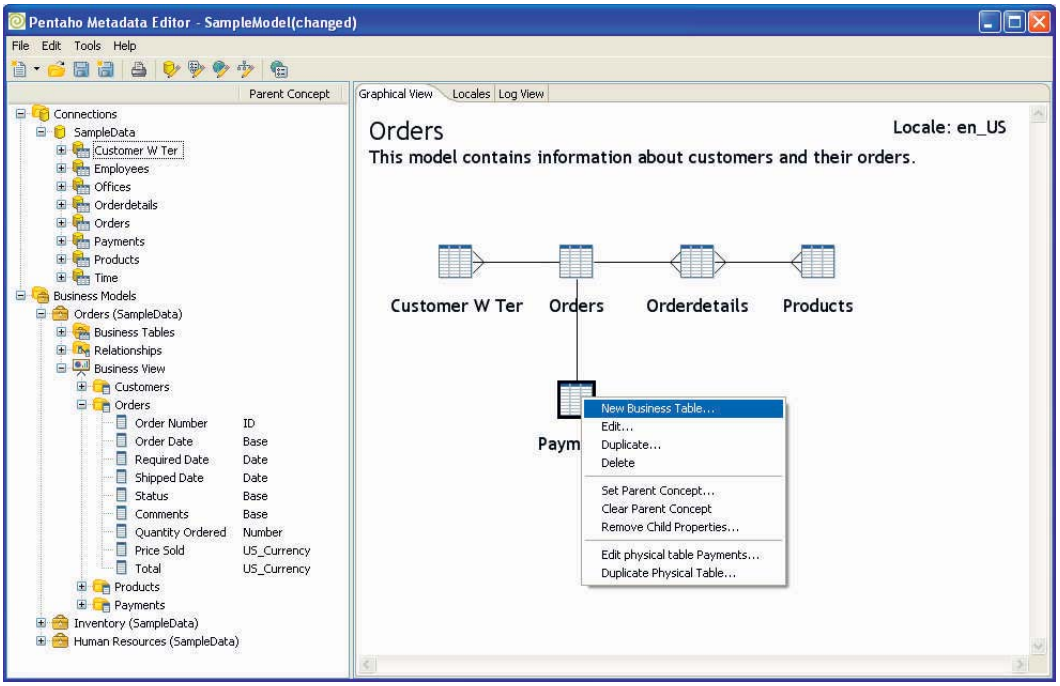
Ingres: Alive and Kicking?

Het is alweer drie jaar geleden (om precies te zijn: in DB/M5 2004) dat er in dit blad aandacht werd geschonken aan één van de oudste RDBMS'en op de markt; Ingres. De aanleiding voor dat artikel was de stap van Computer Associates om Ingres te 'open sourcen'. Feitelijk ging het om het terugkeren van Ingres naar zijn roots, aangezien het product in de jaren zeventig door Michael Stonebraker aan de universiteit van Berkeley is ontwikkeld en pas later is vercommercialiseerd. In november 2005 heeft deze stap daadwerkelijk zijn beslag gekregen en hebben tevens zo'n 100 voornamelijk technische mensen een overstap gemaakt van CA naar Ingres. Als we verder graven in het archief zien we nog een tweetal artikelen in DB/M5 en DB/M8 van 1999 waarin de nieuwe features van Ingres 2.5 worden besproken. En nu, na een aantal jaren van relatieve radiostilte, is Ingres naar eigen zeggen weer terug aan het front met een aantal nieuwe initiatieven, reden van het verschijnen van deze update.

Ingres is en blijft uiteraard het zeer solide database-platform waar nog steeds zo'n 10.000 organisaties wereldwijd gebruik van maken voor bedrijfskritische toepassingen. De huidige versie heet Ingres 2006 release 2 en is zoals gezegd een open source product. Dat men de afgelopen jaren niet heeft stilgezeten blijkt wel uit de nieuwe of bijgewerkte features, waaronder vanzelfsprekend de ondersteuning voor Linux en een groot aantal open standaarden en tools zoals JDBC, JBoss, Eclipse en PHP. Voor een compleet overzicht zie www.ingres.com/products/ingres-2006.php. Dit alleen echter is niet voldoende om onderscheidend te zijn in een inmiddels behoorlijk volle open source database-markt. Ook een andere eigenschap van Ingres zorgt ervoor dat het product enigszins in de vergetelheid is geraakt, namelijk het feit dat als het spul eenmaal draait er nauwelijks nog naar omgekeken hoeft te worden. Waar Ingres zich echter wel mee onderscheidt zijn twee kenmerken die het product ook bij uitstek geschikt maken voor BI-toepassingen, namelijk de ondersteuning voor het parallel optimaliseren



Afbeelding 3: TPC-H performance OS databases.



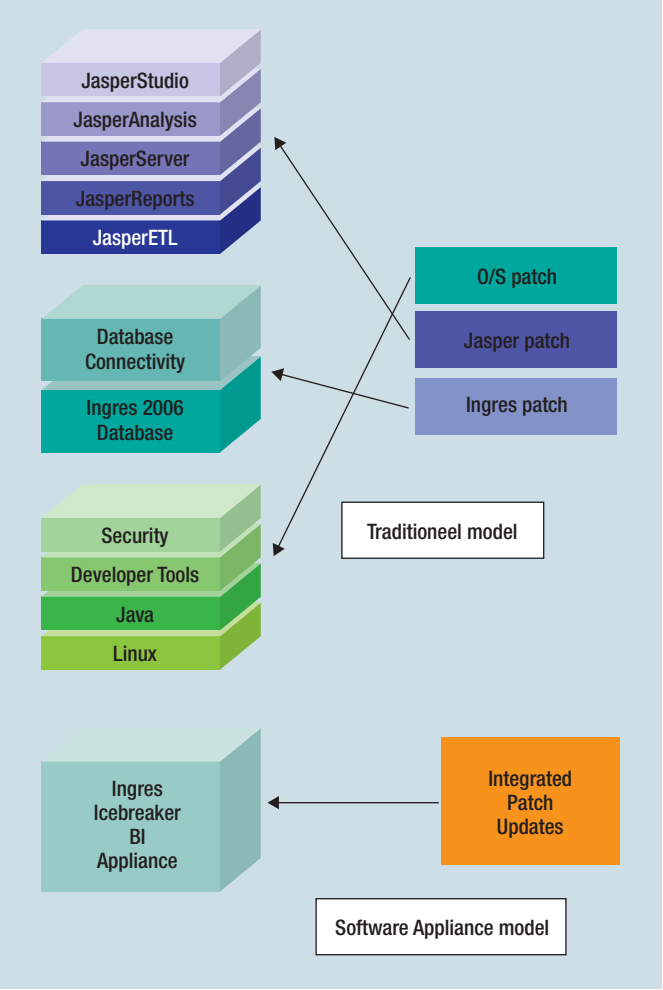
Afbeelding 4: Pentaho metadata-laag.

en uitvoeren van (complexe) query's en de mogelijkheid om diverse vormen van (geneste) partitionering te gebruiken. Dit, in combinatie met de cluster-mogelijkheden, heeft er dan ook voor gezorgd dat Ingres gekozen is als database engine door DATAllegro, één van de leidende DWH appliance leveranciers. Ook een partij als Business Objects heeft (een deel van) zijn kaarten op Ingres gezet.

Niche

Als een goede database, open source of niet, niet meer voldoende is om aandacht van potentiële gebruikers te trekken moet er wat anders gebeuren. Men heeft zich dan ook laten inspireren door de virtualisatietrend die al enige tijd gaande is, en ook goed gekeken wat nu de grootste ergernis is bij het aan de praat krijgen van een open source stack. Dit laatste is het goed op elkaar aansluiten van alle losse onderdelen. Wanneer u zelf eens heeft getracht een werkende open source omgeving op te tuigen waarbij Linux, Apache, MySQL en PHP (LAMP) afzonderlijk zijn gedownload, en waarbij u ook nog een BI Suite als Pentaho of JasperSoft in combinatie met een applicatieserver als JBoss aan wilt sluiten, weet u waarschijnlijk wel wat hier bedoeld wordt. Met Ingres zag men hierin voor zichzelf een interessante niche die op twee manieren wordt ingevuld. In beide gevallen gaat het om wat men een 'software appliance' noemt. Nu zullen de puristen roepen dat een appliance toch minimaal over een aan/uit knop en een stekker dient te beschikken, maar het idee is zo gek nog niet. Met behulp van de software van rPath (www.rpath.com) levert Ingres allereerst het product 'Ingres Icebreaker', een kant en klare voorgeconfigureerde 'stack' met (Linux) operating system en database in één, die direct op een (virtuele) server kan draaien. De prijzen die Ingres rekent voor support van Icebreaker zijn hetzelfde als voor de 'kale' database en komen uit op ongeveer 6200,- euro per socket per jaar. Het staat u echter vrij om af te zien van deze support en de 'community' edition te gebruiken.
Een stapje verder gaat de tweede variant, de Ingres Icebreaker BI appliance. Dit is een complete BI stack bestaande uit operating system, database èn de JasperSoft BI suite, inclusief de open source ETL tool van Talend die inmiddels ook onderdeel uitmaakt van het JasperSoft platform. Het grote voordeel hiervan is dat maintenance en support via één kanaal beschikbaar zijn.

Alle updates komen van Ingres, waardoor het niet meer nodig is om de afzonderlijke onderdelen te patchen en vervolgens te testen. Ingres garandeert hierbij dan ook dat alle onderdelen op elkaar afgestemd blijven. Het verschil in complexiteit wordt geïllustreerd in afbeelding 5. Dit biedt vervolgens partners weer de gelegenheid om verticale oplossingen te ontwikkelen op dit platform, omdat de energie dan ook daadwerkelijk in de oplossing gestoken kan worden in plaats van het aan elkaar knutselen van alle losse onderdelen. Voorbeelden van deze partners zijn Wipro, Tata Consulting Services en Optwize. Een mogelijk issue is echter wel de prijsstelling: met 34.000,- US dollar per jaar voor maximaal twee sockets is de ondersteunde versie niet echt een koopje en zullen organisaties toch eerder geneigd zijn om bijvoorbeeld bij Microsoft te gaan winkelen. Een tweede issue is de 'single box' benadering. Over het algemeen zullen ETL, Datawarehouse en rapportage/analyse op verschillende servers draaien, voornamelijk om redenen van performance en schaalbaarheid. Door alles op één (virtuele) server te plaatsen vormen



Afbeelding 5: tradioneel model versus software appliance.

juist deze onderdelen de bottleneck en zullen de grenzen bij grotere gebruikersgroepen snel in zicht komen. Daar lijkt dan ook een beetje de schoen te wringen: men ziet voor zichzelf vooral een rol als niche speler bij organisaties waar BI op eenvoudige wijze voor iedereen beschikbaar dient te zijn ('pervasive BI') en die open source en open standaarden hoog in het vaandel hebben staan (denk hierbij bijvoorbeeld aan (semi)overheid en gezondheidszorg). Aan de ene kant dus een aanbod dat technisch gezien uitgaat van een beperkte schaalbaarheid, aan de andere kant streven naar 'pervasive BI' voor honderden of duizenden gebruikers. Ingres ziet dan ook als doelgroep voor de appliance organisaties of afdelingen met 50-300 gebruikers en volumes tot ongeveer 1 Terabyte. Voor grootschaliger implementaties (> 1-2 TB, > 250 gebruikers) wordt verwezen naar de DATAllegro oplossing in combinatie met een 'tradionelere' opstelling waarbij de workload over meerdere servers verdeeld wordt. Dit laatste kan trouwens al wel voor het ETL-deel van de appliance: JasperETL (Talend) is een code-generator die naar keuze Perl, Java en SQL genereert. Deze code kan uiteraard op een willekeurige machine worden uitgevoerd, waarbij de appliance desgewenst als scheduler kan optreden.

De toekomst

Als u de bladen een beetje volgt bent u vast wel ergens het persbericht tegengekomen met de aankondiging van de nieuwe Icebreaker appliances. Ingres geeft weer regelmatig acte de présence op beurzen en seminars zoals onlangs bij de BI kring, en men is druk doende de Europese sales force weer op sterkte te krijgen. De vraag uit de titel of Ingres 'alive en kicking' is kan dan ook bevestigend worden beantwoord. In de back-office zien we daarnaast nog een flinke groep mensen die al meer dan 15 jaar aan het product werken, wat weer enig vertrouwen biedt als het om de continuïteit van de ontwikkeling en (technische) ondersteuning gaat. Ook het 'appliance' model biedt goede mogelijkheden om naast een BI stack andere database-intensieve oplossingen aan te gaan bieden. Dit laatste betekent echter het loslaten van de huidige BI focus en maakt daardoor de marketing boodschap wat lastiger over te brengen. Met de Icebreaker BI appliance en de intensieve samenwerking met DATAllegro alleen is men er echter nog niet om een serieuze speler in de BI-markt te worden, ook al beschikt men dan over een goede database. Om een paar voorbeelden te noemen: Ingres beschikt niet over full-text indexing en search mogelijkheden, wat een handicap vormt als men naast gestructureerde ook semi- en ongestructureerde data in de database wil opslaan en terugvinden. XML support is niet aanwezig, wat om dezelfde reden een minpunt genoemd zou kunnen worden.
Hét tovermiddel tegen lange wachttijden, de materialized view, ontbreekt eveneens en naar bitmapped indexen en de in de SQL 2003 standaard gedefinieerde analytische functies zult u ook vergeefs zoeken. Waarschijnlijk gaan full-text en bitmapped indexing de roadmap voor 2008 wel halen (hoewel dit nog niet definitief is), dus er wordt wel degelijk serieus naar deze aanvullingen gekeken.

Ingres kan zich op dit moment op BI-vlak dan ook nog niet meten met de grote commerciële leveranciers. De vraag is echter of dat wel nodig is: Ingres biedt als enige OS database-leverancier nested partitioning en parallel query processing, twee functies die in de commerciële wereld alleen in de 'Enterprise' editions beschikbaar zijn, met het daarbij horende prijskaartje. Als er daarnaast serieus werk wordt gemaakt van bitmapped en full-text indexing levert het bedrijf een product dat gelijk alle open source RDBMS'en op BI-gebied functioneel gezien achter zich laat. Kortom, het is een bedrijf om in de gaten te houden en mocht u op zoek zijn naar een robuuste database voor een BI-oplossing binnen uw organisatie, verdient het aanbeveling om Ingres minimaal op te nemen in de evaluatie.

Jos van Dongen
Jos van Dongen (jvdongen@tholis.com) is Senior Consultant bij Tholis Consulting.