Analytische databases versus transactie-databases

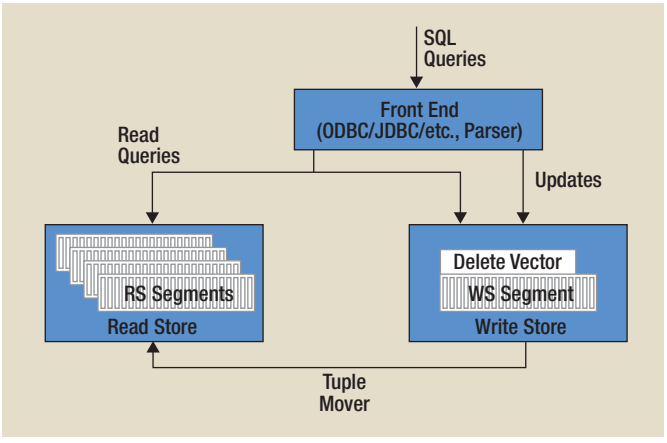
Hoge Snelheids (data) Lijn

Jos van Dongen

Er is een nieuwe generatie analytische databases, die in het spoor van de datawarehouse appliance trend de laatste twee jaar het levenslicht gezien heeft. Maak kennis met Infobright, ParAccel, Vertica, Exasol of DatAupia en hun claims om query's tot 400 keer sneller af te handelen dan de traditionele producten, tegen een fractie van de kosten van deze laatste.

Het hele datawarehouse-gebeuren is begin jaren negentig geboren vanuit de gedachte (en ervaring) dat een database die is geoptimaliseerd voor transactieverwerking niet geschikt is voor het verwerken van analytische vragen. Lijstjes met verschillen tussen OLTP en DWH zijn in elk inleidend boek te vinden dus dat gaan we hier niet herhalen. Hoewel deze verschillen in veel gevallen inderdaad geleid hebben tot het inrichten van een aparte database voor analysevragen, werd (en wordt) meestal nog steeds gebruik gemaakt van dezelfde 'good old' database-technologie die al dertig jaar gebruikt wordt voor OLTP-systemen.

En ondanks dat Sybase met hun IQ-product al jarenlang laat zien dat er betere oplossingen zijn voor analytische toepassingen dan het standaard RDBMS, heeft dit nog niet geleid tot een massale overstap naar alternatieven. Hierin komt echter rap verandering, met name als gevolg van het succes van datawarehouse appliances als Greenplum, DATAllegro en vooral Netezza. Deze producten hebben laten zien dat het loont om op een andere manier met data om te gaan en dat het inzetten van een combinatie van gespecialiseerde hardware en software tot spectaculaire verbeteringen leidt van performance èn tot een verlaging van kosten. Toch zijn deze producten nog steeds gebaseerd op de traditionele, rijgeoriënteerde wijze van data-opslag, hoewel er onder de motorkap behoorlijk wat is verbouwd aan de PostgreSQL- en Ingres-engines.

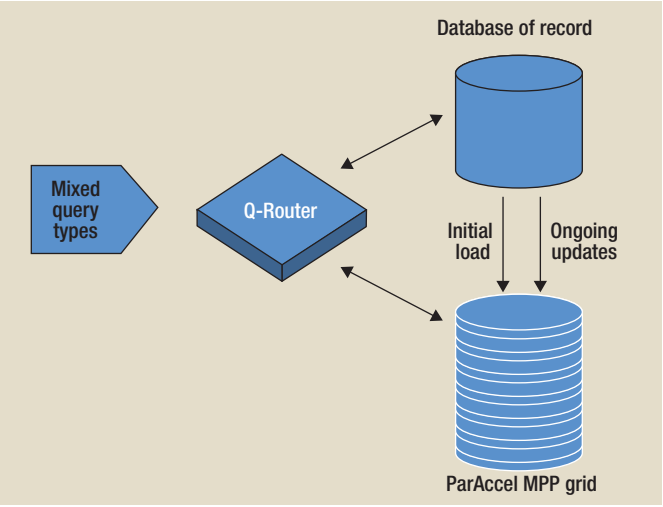


Afbeelding 1: Vertica met Read Only Store en Write Only Store.

Het idee achter de nieuwe generatie analytische databases is dat data opgeslagen in kolommen in plaats van rijen veel beter aansluiten bij het soort vragen in een typische BI setting. Om een vraag als 'geef mij de gemiddelde omzet per klantgroep per jaar over de laatste drie jaar' te kunnen beantwoorden hoeven ten-slotte maar drie kolommen in de database benaderd te worden, de rest hoeft niet mee te doen. In een traditionele rijgeoriën-teerde database dienen hiervoor indexen en/of materialized views gemaakt (en beheerd) te worden om hetzelfde effect te bereiken. Er is echter meer. Doordat elke kolom afzonderlijk opgeslagen wordt kan deze in veel gevallen ook flink gecompri-meerd worden, wat leidt tot een veel lagere I/O-belasting. En als dan de werklast ook nog een keer verdeeld kan worden over een (groot) aantal machines levert dit nog meer performance-winst op. Deze drie 'trucs' zien we terug bij vrijwel alle nieuwe analytische databases: een MPP (Massively Parallel Processing) architectuur, kolomgebaseerde opslag en agressieve datacom-pressie. In dit artikel wordt omwille van de ruimte slechts een beperkt aantal producten besproken, dus wie onder andere Valentina, VectorBase, Tenbase, Sand, Calpont en Bigtable mist: wellicht een volgende keer.

Vertica

De database-wereld zou er heel anders uitzien zonder Michael Stonebraker. Hij wordt gezien als één van de belangrijkste ontwikkelaars van database management-systemen en heeft een



Afbeelding 2: ParAccel Amigo configuratie met 'Q-Router'.

grote bijdrage geleverd aan de wetenschappelijke fundamenten van database-technologie. Het unieke aan Stonebraker is zijn combinatie van wetenschappelijk en commercieel inzicht. Al sinds 1973, toen zijn werk aan de Ingres database begon, weet hij telkens zijn onderzoeksresultaten te vertalen in een commercieel product. Dit geldt ook voor zijn laatste geesteskind, Vertica. Dit product is gebaseerd op het onderzoeksproject C-Store, en wie de Vertica whitepapers vergelijkt met de wetenschappelijke C-Store papers ziet dat er inderdaad een 1-op-1 vertaling is gemaakt van een theoretische verkenning naar een praktische en verkoopbare oplossing.

De naam van Michael Stonebraker staat niet alleen garant voor innovatieve producten, maar ook voor (veel) geld. In september 2007 is er 25 miljoen dollar opgehaald in een tweede financieringsronde, dus blijkbaar is er bij Venture Capitalists veel vertrouwen in een succesvolle afloop.
Vertica beschikt uiteraard over de basiskenmerken kolommen, MPP-architectuur en datacompressie, maar voegt daaraan wat extra's toe. Het belangrijkste punt waarin Vertica zich onderscheidt van de overige aanbieders is het (strikte) onderscheid tussen lees- en schrijfgeoptimaliseerde opslag. Uiteraard moet er op enig moment naar een database geschreven kunnen worden om nieuwe informatie te verwerken, maar Vertica gaat er vanuit dat meer dan 95 procent van de data 'read only' is. Dat betekent dat je daar ook de engine en opslag naar kunt optimaliseren. Binnen C-Store/Vertica heet dit deel van de database de RoS (Read Only Store), en het complement hiervan heet inderdaad de WoS (Write Only Store). Wijzigingen worden geschreven naar de WoS, waarna een speciale 'Tuple Mover' ervoor zorgt dat op reguliere basis deze wijzigingen worden verwerkt in de RoS. Query's afgevuurd op de database benaderen zowel RoS als WoS, waardoor een zeer geringe data latency (vertraging tussen beschikbaar komen van broninformatie en verwerking hiervan in een datawarehouse) bereikt wordt, zie afbeelding 1.
Een andere belangrijke uitbreiding op het kolomprincipe is de opslag in zogenaamde 'projecties'. Elke dataprojectie bevat een deelverzameling van de beschikbare kolommen en is gericht op het beantwoorden van een specifieke vraag, bijvoorbeeld met betrekking tot klanten of verkoopprijzen. Door deze projecties vervolgens horizontaal te partitioneren in segmenten en deze vervolgens over verschillende machines te verdelen, wordt parallellisatie van query's mogelijk gemaakt. Door daarnaast ook redundantie toe te passen wordt de oplossing zowel snel als zeer robuust. Helaas heeft Vertica nog geen benchmarks aangemeld bij de Transaction Processing Council (TPC), dus we moeten het voor wat betreft mogelijke performance-verbeteringen voorlopig doen met de marketinggegevens van het bedrijf zelf. Hoewel deze cijfers altijd met een flinke korrel zout moeten worden genomen trekken uitspraken als een 214 keer betere query performance toch de aandacht, helemaal als de hardware-kosten 50 procent bedragen van de vergelijkingsconfiguratie en de data drie keer zo snel geladen worden als met een OLTP batch loader. Met name dit laatste feit is interessant om te vermelden, omdat één van de oorspronkelijke problemen met kolomgeoriënteerde databases juist de (te) lange laadtijden betrof.

Vertica is ontwikkeld als software-oplossing en wordt ook als zodanig verkocht. Inmiddels is er echter samen met HP ook een complete appliance op de markt gebracht waardoor de implementatie-inspanningen beperkt blijven tot het aansluiten van de stekkers en kabels en het omhangen van de database-connectie van uw huidige naar het nieuwe datawarehouse. Interessant om te zien dat HP hiermee een concurrent wordt van zijn eigen NeoView product.

MPP versus SMP

In tegenstelling tot SMP (Symmetrical Multiple Processing) wordt in een MPP-architectuur gebruik gemaakt van de kracht van heel veel relatief eenvoudige computers. Deze ontwikkeling is vooral mogelijk gemaakt door de drastische prijsdalingen bij een tegelijkertijd logaritmisch toenemende performance van vooral processoren en geheugenchips. In de wereld van de supercomputing zijn de Cray en SGI SMP machines al enige tijd geleden van hun troon gestoten door de BlueGene machines van IBM die zijn opgebouwd volgens het MPP-principe. Ook Cray en SGI zijn trouwens overgestapt op deze architectuur, waarbij elke 'computing node' beschikt over zijn eigen OS, CPU en Memory. Wanneer vervolgens ook nog elke node zijn eigen disk(s) bestiert wordt gesproken van een Shared Nothing architectuur, en dat is precies waar de nieuwe generatie DWH databases gebruik van maakt. Het is echter een uitruil van voors en tegens. Hoewel een configuratie van 12 dual processor machines met elk 16 GB geheugen vaak veel voordeliger zal zijn in aanschaf dan één 24 CPU SMP machine met 192 GB RAM, dienen er in het eerste geval wèl 12 machines beheerd te worden in plaats van 1.

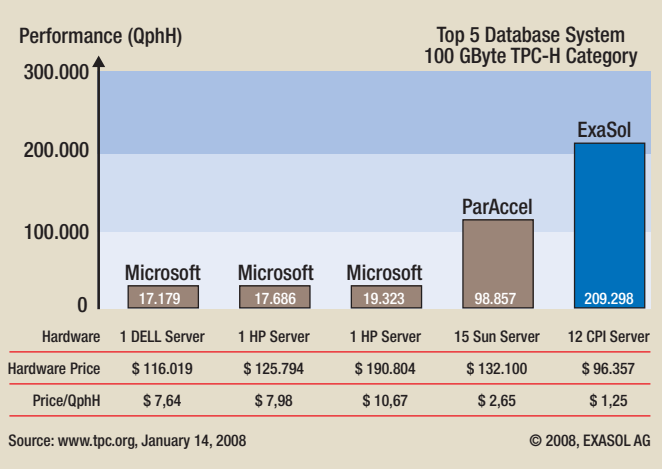
ParAccel

ParAccel (van Parallel Accelerator) is in 2005 opgericht door een aantal zeer ervaren krachten uit de database- en BI-wereld, wat blijkbaar voldoende vertrouwen wekte om eind 2007 20 miljoen dollar los te peuteren van durfkapitalisten. Het product lijkt in eerste instantie erg veel op Vertica en (vooral) ExaSol, maar voegt daar toch wat eigen onderdelen aan toe. Het gaat alweer om MPP, kolommen en compressie, maar ook om in-memory verwerking, een eigenschap die het deelt met ExaSol. Uniek aan Paraccel is echter de wijze waarop het product ingezet kan worden in een bestaande database/datawarehouse-omgeving. Er is namelijk een 'Amigo' variant beschikbaar (vooralsnog alleen voor SQL Server; een Oracle-versie is in ontwikkeling) die wèl alle voordelen (snellere responstijden tegen lagere kosten) maar niet de nadelen (vervangings-, implementatie- en trainingskosten) heeft van een complete vervanging.
Het aantrekkelijke deel van de Amigo configuratie is de zogenaamde 'Q-Router' die alle query's evalueert en afhankelijk van het type query deze doorstuurt naar ofwel de OLTP database ofwel het ParAccel MPP grid, zie afbeelding 2. De tweede implementatievorm wordt 'Maverick' genoemd en hier wordt de software als *stand alone* datawarehouse en datamart engine ingezet.

De software is beschikbaar in twee smaken (met verschillende prijskaartjes); in-memory of disk-based. Wanneer deze in appliance-vorm in combinatie met Sun hardware worden aangeschaft heten deze respectievelijk 'Phoenix' en 'Sedona'. En net als bij Vertica's HP partnership zien we hier een hardware-leverancier die met zichzelf concurreert, omdat ook de Greenplum appliance op Sun is gebaseerd.
ParAccel was trouwens de eerste nieuwe partij die het aandurfde om zichzelf voor een gestandaardiseerde datawarehouse benchmark aan te melden bij de Transaction Processing Council (TPC). Het resultaat van deze samen met Sun uitgevoerde exercitie was het verpletteren van alle tot dan toe gepubliceerde resultaten in het 100, 300 en 1000 GB segment, zowel op het gebied van performance als op het gebied van prijs/prestatie. Dat het overigens altijd sneller en goedkoper kan kunt u in de volgende paragraaf lezen.

ExaSol

Op het moment van schrijven is ExaSol de nieuwe koning in het 100 GB en 1 TB segment van de TPC-H benchmarks. En dit geldt niet alleen voor de performance, maar vooral voor de prijs/prestatie verhouding die spectaculair veel lager ligt dan van welke andere oplossing ook, zie afbeelding 3. ExaSol is een Duits bedrijf en is in 2000 opgericht door Michael Gutzmann die in de tien jaar daarvoor carrière heeft gemaakt in de wetenschappelijke wereld, met name op het gebied van parallel computing. Inmiddels heeft het bedrijf 50 medewerkers en is een paar maanden geleden de internationale weg ingeslagen, om te beginnen door gehakt te maken van alle bestaande TPC-H scores, inclusief



Afbeelding 3: ExaSol prijs/prestatie-verhouding.

die van concurrent en geestverwant ParAccel. ExaSol bereikt deze resultaten door het bekende rijtje MPP, kolomgebaseerde opslag en datacompressie in te zetten, en heeft daar nog het een en ander aan toegevoegd. Ten eerste worden zowel de (actieve) data als eventuele indexen gecomprimeerd in het geheugen. Ten tweede draait de database op een proprietary operating system genaamd ExaCluster OS. Dit is een aangepaste versie van de gratis open source Red Hat kloon CentOS, met voorzieningen die het met name geschikt maken voor het aansturen van grote hardware clusters. Tot slot is de ExaLoader ontwikkeld om zeer snel grote hoeveelheden data in het systeem te kunnen laden. Ook hierin verslaat ExaSol alle andere aanbieders, zowel traditioneel als kolomgebaseerd. De tuning van de database verloopt geheel automatisch en ook het fysieke ontwerp, inclusief de distributie van de data over verschillende nodes, is volledig transparant. ExaSol claimt naadloos samen te werken met alle bekende BI-leveranciers als Business Objects en Cognos en kan zodoende dienen als zeer snelle database back-end voor een veelheid aan analytische toepassingen.

Infobright

Dit bedrijf is opgericht door vier Polen, waarvan er drie gepromoveerd zijn aan de universiteit van Warschau. Het analytische database-product dat ze ontwikkeld hebben heet Brighthouse, en is in essentie een storage engine voor MySQL. Dat maakt Brighthouse ook meteen toegankelijk voor een grote groep gebruikers en betekent eveneens dat er ten eerste een wereld aan additionele software beschikbaar is en ten tweede dat het product naadloos in de meeste ICT-omgevingen ingepast kan worden. De software wijkt op een paar punten af van de hiervoor beschreven oplossingen. Allereerst is er nog geen sprake van MPP-ondersteuning, deze is voor het eind van 2008 aangekondigd. De tweede afwijking betreft de behaalde compressie die met meer dan 10:1 verder gaat dan alle vergelijkbare oplossingen. Het laatste, en wellicht belangrijkste, punt betreft de wijze van opslag en indexering. De data worden weliswaar kolomsgewijs opgeslagen maar verder onderverdeeld in 'data

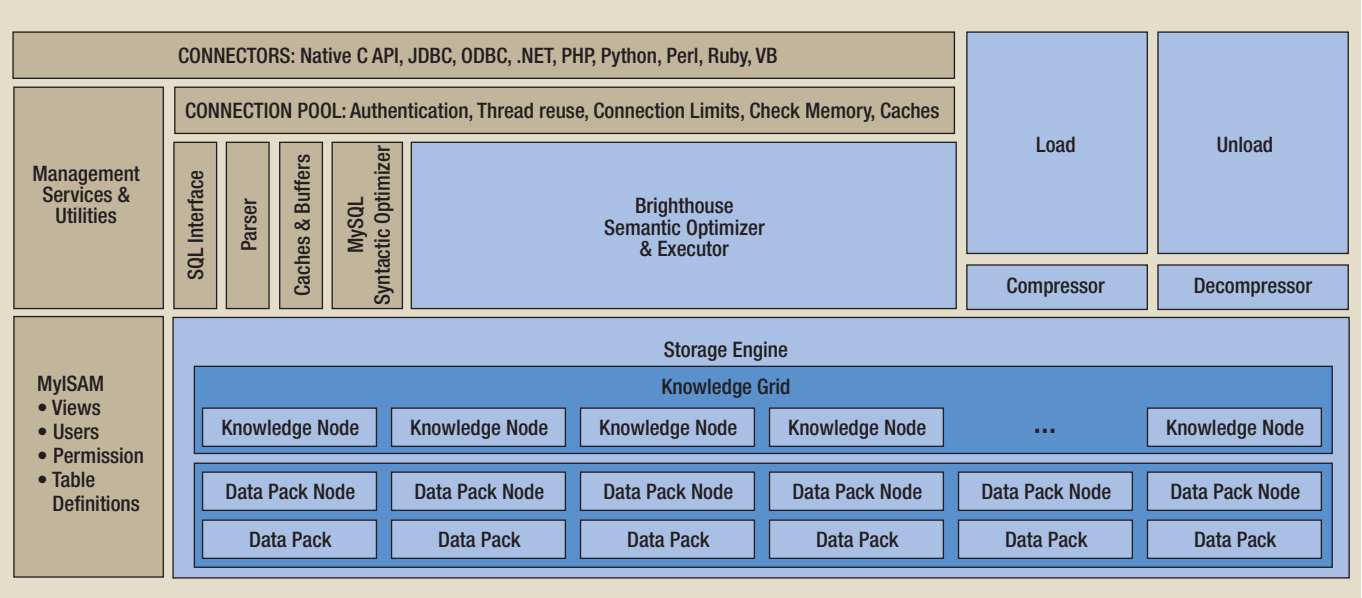
Kolommen versus rijen

Kolomgebaseerde opslag en bit-wise indexing is niets nieuws: het principe is in 1969 al toegepast in de 'TAXIR Accessioner' en is dus al bijna 40 jaar oud. De technologie wordt zoals gezegd al ruim 13 jaar succesvol toegepast binnen Sybase IQ. Overigens heeft Sybase deze technologie niet zelf ontwikkeld maar in oktober 1994 gekocht van Expressway Technologies. Ook de open source producten MonetDB en LucidDB zijn gebaseerd op opslag in kolommen. Simpelweg komt het er op neer dat records niet regel voor regel, maar kolom voor kolom worden weggeschreven, waardoor veel minder I/O vereist is dan de gangbare rijgebaseerde opslag. Een tweede voordeel betreft de compressiemogelijkheden. Omdat alle waarden in een kolom van hetzelfde type zijn kan er zeer efficiënt gecomprimeerd worden. Een andere manier om de benodigde opslag te verkleinen en de responstijd te verlagen is het vervangen van veel voorkomende waarden door ID's.

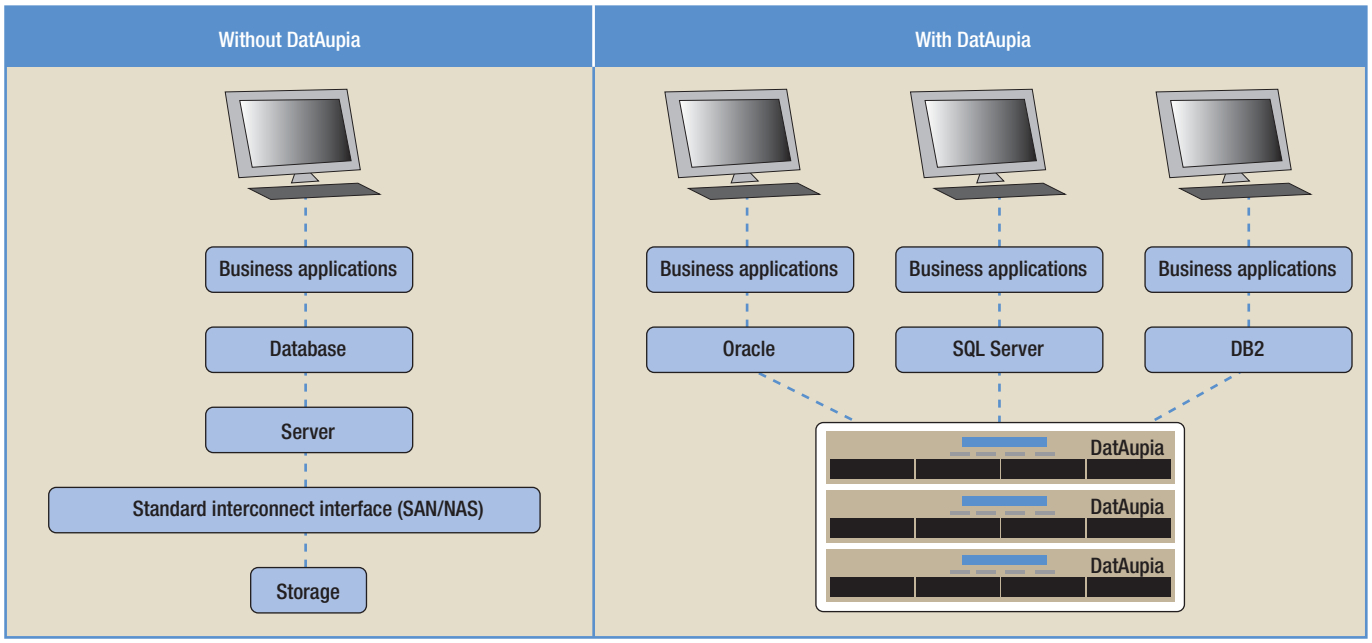
packs' van 64 KB. Elk data pack heeft een corresponderende data pack 'node' waarin de metadata van een data pack worden opgeslagen. Denk hierbij bijvoorbeeld aan minimum- en maximumwaarde en data pack totaal in het geval van numerieke waarden. Vervolgens worden er dynamisch zogenaamde 'knowledge nodes' aangemaakt waarin bijvoorbeeld wordt opgeslagen welke combinatie van data packs voor welke joins een resultaat gaan opleveren. Een en ander wordt uitgebreid beschreven in de white papers op de site. In afbeelding 4 is schematisch goed te zien hoe de opbouw van de database en overige software-componenten in elkaar steekt. Let daarbij vooral op de term 'semantic optimizer'. Er wordt dus getracht om op basis van de beschikbare kennis over de opgeslagen gegevens een snelle responstijd te realiseren, in tegenstelling tot de overige oplossingen die een meer 'brute force' aanpak hebben; een wellicht subtiel, maar toch niet onbelangrijk verschil.

DatAupia

In 2005 verlieten twee (mede)oprichters van Netezza hun bedrijf en gingen elk een eigen weg. Barry Zane besloot zich meer te gaan richten op datawarehouse-software en richtte het hiervoor besproken ParAccel op, Foster Hinshaw besloot het appliance-trucje nog een keer uit te halen en begon DatAupia. Over de achtergrond van de wat vreemd klinkende naam doen verschillende verhalen de ronde, maar dat het iets met 'Data' en 'Utopia' te maken heeft staat in elk geval vast. DatAupia levert de Satori 12000 server die, in tegenstelling tot bestaande appliances als Netezza, Greenplum en DATAllegro, bedoeld is om een bestaande SQL Server-, Oracle- of DB/2-omgeving een flinke performance boost te geven in plaats van een complete vervanging van deze oplossingen na te streven. Volgens recent onderzoek is namelijk 75 procent van de bedrijven niet van plan om een 'rip and replace' actie uit te voeren op een bestaand datawarehouse, ook al zou dit nog zo veel voordelen bieden. DatAupia biedt dus een hardware-oplossing die alleen de opslag- en executielaag van de database vervangt, niet de database software zelf. Hèt grote voordeel van deze benadering is dan ook dat deze oplossing volledig onzichtbaar is en geen enkele impact heeft op bestaande applicaties. Zie afbeelding 5 voor een schematische weergave van dit fenomeen.
De Satori 12000 server is opgebouwd uit blades met elk twee dual core Opteron processoren en 8 SATA disks in een RAID-configuratie, zodat 2,2 TB opslag per blade beschikbaar is. Om te profiteren van de MPP-architectuur dienen wel meerdere blades aan het werk gezet te worden. Het gebruikte operating system is Linux, waaraan DatAupia een stuk middleware heeft toegevoegd



Afbeelding 4: Opbouw Brighthouse.



Afbeelding 5: DatAupia.

dat voor de onderlinge communicatie tussen de blades zorgt. Intern zijn er testen uitgevoerd met een 50-blade systeem, dus met een capaciteit van 110 TB. DatAupia is zeer concurrerend geprijsd, waardoor het online beschikbaar houden van grote datavolumes voor veel organisaties bereikbaar wordt.

Licentiekosten

Eén van de vervelendste aspecten aan de (BI) software-markt is wel de totale afwezigheid van transparantie waar het gaat om licentiekosten. Ook in het geval van deze nieuwe generatie datawarehouse-oplossingen valt het niet altijd mee om inzicht te krijgen in de prijzen. Een beetje lastig als je enerzijds claimt een veel lagere TCO te hebben dan traditionele database-leveranciers, maar vervolgens geen inzicht in de prijsstelling geeft. Gelukkig hebben de meeste dit ook ingezien en zijn sommige leveranciers, zoals ParAccel en DatAupia, erg open over hun licentiekosten. Alle producten zijn gewoon leverbaar en alle leveranciers zien u graag verschijnen voor een *proof of value* in hun eigen demo/testruimte. Voor wat betreft ondersteunde operating systems dient u nog even de kleine lettertjes goed te lezen: Windows Server staat vrijwel nooit in het rijtje, Red Hat Enterprise Linux vrijwel altijd en afhankelijk van het product komt u Suse Linux, Solaris en Fedora ook nog tegen.

Conclusie

De conclusie ligt voor de hand: beschikt u over een datawarehouse en loopt u tegen performance-problemen aan, kijk dan eens naar een nieuwe generatie software die speciaal is ontwikkeld voor het verwerken van analytische workloads. Dit is echter een beetje kort door de bocht. De meeste producten die in dit artikel besproken zijn zult u niet op de prijslijst van een Nederlandse distributeur terugvinden. Ook voor support en (implementatie)ondersteuning zult u op dit moment vergeefs aankloppen bij Nederlandse bedrijven. En wat ook geldt: de meeste leveranciers hebben in de tweede helft van 2007 pas de tweede financieringsronde achter de rug, wat nog absoluut geen garantie is voor een langdurig en succesvol bestaan.
Toch zijn er wel mogelijkheden om met deze technologie aan de slag te gaan: Sybase IQ beschikt wèl over een Nederlandse verkoop- en support-organisatie èn heeft zich met meer dan 1200 klanten wereldwijd ook al bewezen. Wie meer avontuurlijk is ingesteld zou eens bij de Oosterburen op de koffie kunnen gaan om ExaSol te evalueren. Ook dit bedrijf bestaat al enige tijd, heeft de support op orde en heeft een groeiend aantal klanten in binnen- en buitenland. Wie eerst eens wil snuffelen aan kolomgebaseerde analytische databases kan trouwens ook terecht in de open source wereld. Zowel MonetDB als LucidDB bieden hiervoor een alternatief, maar ook geen of weinig ondersteuning. En heeft u lak aan alle hiervoor genoemde bezwaren en bent niet bang voor kinderziektes, dan houdt niemand u tegen om als 'early adopter' Vertica, ParAccel, Brighthouse of DatAupia in te zetten en hiermee goede sier te maken op één van de nog steeds in aantal toenemende BI-congressen.

Internet-adressen

B-Eye Network: www.b-eye-network.com; DatAupia: www.dataupia.com; ExaSol: www.exasol.com; Infobright: www.infobright.com; LucidDB: www.luciddb.org; MonetDB: www.monetdb.com; ParAccel: www.paraccel.com; Sybase: www.sybase.nl; TPC: www.tpc.org; Vertica: www.vertica.com

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