Schaalbaarheid gepaard aan performance is het uiteindelijke doel

Open Source Analytical Databases

Jos van Dongen

De jaarlijkse BI Survey van Nigel Pendse laat al jarenlang zien dat er een sterke correlatie bestaat tussen query performance en het enthousiasme van gebruikers voor hun BI-omgeving. Misschien voor de meesten van u geen verrassing, maar toch: hoe mooi de rapporten ook zijn en hoe geavanceerd de analyses, zonder goede performance haken mensen af. In de woorden van Pense zelf: "The faster the query response, the more business benefits are reported and the more likely it is that business goals will be achieved."

Om deze snelle responstijden te bereiken kan er naast een goed ingerichte infrastructuur met snelle processoren, snelle (en liefst veel) schijven en veel geheugen nog een extra wapen in de strijd worden gegooid: een analytische database, speciaal ontwikkeld voor BI-doeleinden. Ongeveer een jaar geleden vroeg ik me in een blog af waar de open source (OS) analytische databases bleven. Dit naar aanleiding van diverse berichten waarin OS BI-leveranciers gebundelde oplossingen aankondigden met Vertica (Pentaho) en Paraccel (Jaspersoft), twee van de bekendere spelers op dit gebied. In DB/M 8, 2008 heeft u in het artikel 'Inside the New-Generation Analytic DBMS' kunnen lezen wat deze producten onderscheidt van de traditionele databases, en in DB/M 3, 2008 werden bovenstaande producten al besproken. Het is echter goed om nog even de belangrijkste verschillen op een rijtje te zetten:

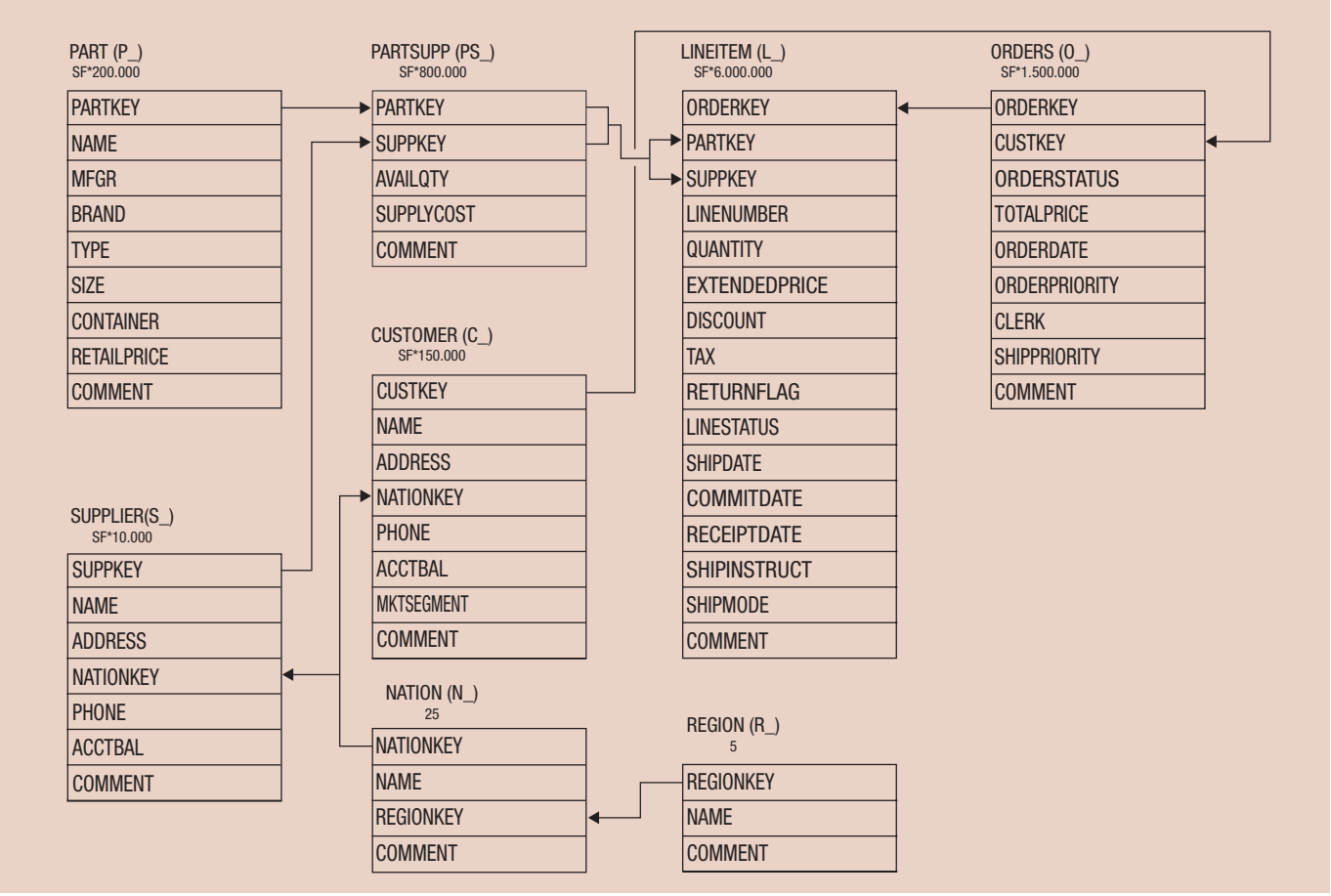
- Column based: data worden niet rij voor rij, maar kolom voor kolom opgeslagen;
- Compressie: gegevens worden gecomprimeerd opgeslagen, tot soms wel een factor 20;
- MPP: meerdere machines worden parallel geschakeld om query's te verwerken;
- Shared Nothing: elke machine in een cluster heeft zijn eigen CPU, geheugen en disks;
- Memory based: er wordt intensief gebruik gemaakt van RAM geheugen.

Dat dit inderdaad kan leiden tot een flinke performanceverbetering is onder meer terug te vinden in de TPC-H benchmark resultaten van de Transaction Processing Council (www.tpc.org/tpch/results/tpch_perf_results.asp), waarover later meer.

Closed versus Open Source

Kijkend naar de diverse analytische database (ADB) producten die er op de markt beschikbaar zijn valt een paar dingen op. Ten eerste het feit dat deze markt volop in beweging is, met bijna elke maand wel een aankondiging van een nieuwe startup die het ook gaat proberen. Dat ze het uiteindelijk niet allemaal gaan redden heeft Dataupia enkele maanden geleden op pijnlijke wijze duidelijk gemaakt, hoewel er inmiddels al wel een doorstart is gemaakt. Het tweede wat opvalt is dat de bedrijven die gestart zijn als softwarebedrijf steeds vaker de samenwerking met hardware leveranciers zoeken om een complete datawarehouse appliance aan te kunnen bieden. En de derde opvallende ontwikkeling is het uitbreiden van de 'kale' database functionaliteit met voorzieningen voor 'in database' analyses met behulp van statistische bibliotheken en het toevoegen van map/reduce functionaliteit. Deze laatste ontwikkeling is bijvoorbeeld goed te zien bij gevestigde partijen als TeraData en ook Greenplum blaast een aardig partijtje mee. We hebben het hier echter nog steeds over closed source oplossingen. Hoe staat het nu in de open source wereld? Heel simpel, daar is men wat minder ver. Geen kant en klare appliances, geen in-database analytics en ook andere kenmerken van ADB's als Paraccel, Exasol en Vertica zijn slechts ten dele beschikbaar.

Er bestaan op dit moment drie open source column stores: MonetDB, LucidDB en InfoBright. De belangrijkste ontbrekende feature van al deze producten is de MPP shared nothing architectuur van de high-end gesloten producten. Andere middelen om een hoge performance te realiseren worden echter niet geschuwd. MonetDB maakt graag gebruik van (veel) geheugen en Infobright biedt een erg goede compressie. LucidDB tenslotte biedt geen compressie en maakt minder optimaal gebruik van beschikbaar geheugen, maar heeft weer andere voorzieningen zoals bitwise indexing die zorgen voor een betere performance. Er is echter op dit moment nog geen open source database beschikbaar die alle trucjes van de gesloten concurrenten



Afbeelding 1: TPC-H Schema.

beheerst. Uiteraard is men zich bewust van deze beperkingen en is het de bedoeling om clusteringfaciliteiten toe te voegen. Wie in de MonetDB broncode naar repository's gaat snuffelen zal ontdekken dat er gewerkt wordt aan een versie met codenaam 'Octopus' en ook LucidDB heeft plannen om zogenaamde 'scale out' faciliteiten te ontwikkelen. Bij deze projecten is echter de beschikbare mankracht een belangrijke beperkende factor, en ook zakelijke ontwikkelingen zorgen ervoor dat de prioriteiten vaak elders liggen. LucidDB bijvoorbeeld was de databasemotor achter de SAAS BI vendor Lucidera die vorige maand ter ziele ging. Hiermee viel tevens de belangrijkste sponsor van het project weg. Voor echte high-end, multi-terabyte scale analytische databases bent u dus voorlopig aangewezen op de commerciële database- en applianceleveranciers.

Bouwen met PostgreSQL en MySQL

In de basis zijn zowel PostgreSQL en MySQL niet speciaal geschikt voor datawarehouse- en analytische toepassingen. Het open source karakter heeft toch heel veel bedrijven geïnspireerd om op de basis van beide producten alternatieven te ontwikkelen. In de meeste gevallen is hierbij de 'voorkant' van de database (het DBMS en de SQL parser) ongemoeid gelaten, maar is het juist de 'achterkant' (de storage en query engine) die is aangepakt. Dit heeft in het geval van PostgreSQL een aantal succesvolle commerciële producten als Netezza, Greenplum en Paraccel opgeleverd, maar ook nieuwkomers als Aster Data en het inmiddels failliete Dataupia hebben voortgeborduurd op de PostgreSQL code. Helemaal toevallig is dat niet: de BSD licentie laat eenieder vrij om met de broncode te doen wat men wil, inclusief het gebruik daarvan voor het ontwikkelen van closed source producten. Een bedrijf dat PostgreSQL wel als basis gebruikt voor producten die zelf ook weer (deels) open source zijn is EnterpriseDB. Het meest interessant in dit opzicht is EnterpriseDB's GridSQL, het voormalige ExtenDB. GridSQL (de naam zegt het al) maakt van PostgreSQL een MPP achtige oplossing waarbij de engine echter wel gewoon een standaard versie blijft. De performancewinst zal dan ook uit het clusteren van de database moeten komen waarbij query's parallel verwerkt kunnen worden. Greenplum is weer een ander verhaal: initieel was er een open source editie (BizGres) maar deze is meer en meer op de achtergrond geraakt en zelfs de bijbehorende website bizgres.org is overleden. Wel is er het Greenplum Network waar u na de gratis registratie de meest recente Greenplum versies kunt downloaden en eveneens licentievrij mag gebruiken voor ontwikkel- en testdoeleinden.

Ook MySQL is een dankbaar platform waarbij vooral de pluggable storage engine architectuur het erg gemakkelijk maakt om

speciale toepassingen te ontwikkelen. Eén van de meest aansprekende voorbeelden is Kickfire dat gestoeld is op een speciaal ontwikkelde SQL-chip. Kickfire is, mede door de aantrekkelijke prijsstelling, dan ook een logische keuze als een bestaande MySQL-omgeving van meer analytische 'PK's' voorzien dient te worden. Het is echter niet alleen de software die continu verbeterd wordt, ook hardware speelt een belangrijke rol bij het opkrikken van prestaties. Denk aan steeds goedkoper wordend geheugen, steeds snellere processoren maar vooral aan nieuwe harddisktechnologie. Solid State Disks (SSD's) zijn inmiddels betrouwbaar en betaalbaar genoeg voor bedrijfsgebruik en voor traditionele SAS- of SATA-schijven onverslaanbaar voor wat betreft toegangstijd (0,1 ms ten opzichte van 3,5 ms voor de snelste conventionele SAS disks) en transfer rates (255 MB/s ten opzichte van 120 MB/s voor SAS). Een trede hoger op de performanceranglijst staan producten als Fusion-IO die een nog hogere doorvoersnelheid kennen (500 MB/s schrijven, 800 MB/s lezen). De huidige prijzen (3000 dollar voor een 80 GB 'disk') staan voorlopig echter nog een grootschalige doorbraak in de weg, maar er valt met SSD's in combinatie met snelle Raid controllers al veel winst te bereiken (meer dan 3 GB/s throughput is hiermee op een standaard dual Xeon machine haalbaar). Al deze techniek moet echter wel gebruikt kunnen worden, en met name SSD heeft een paar karakteristieken die door de standaard storage engine van MySQL niet goed benut worden. Dit weerhoudt bedrijven als Schooner met hun MySQL Appliance en Virident met hun GreenCloud Server er overigens niet van om hardwarematige oplossingen te leveren gebaseerd op SATA en PCIe SSD's. RethinkDB denkt dat hardware alleen niet de oplossing is en is bezig om een speciale, nieuwe MySQL storage engine te ontwikkelen die is geoptimaliseerd voor het gebruik van SSD's. Momenteel bevindt het product zich nog in de pre-alpha fase maar is zonder meer een interessante ontwikkeling om te volgen. Een ander bedrijf dat zich nog in de embryonale fase bevindt is Calpont, waar men eveneens bezig is met de ontwikkeling van een nieuwe MySQL storage engine. In dit geval gaat het (in tegenstelling tot RethinkDB, Schooner en Virident die zich meer op hoge transactievolumes richten) om een datawarehouse database die van MySQL een modulair, parallel draaiend MPP cluster maakt.

TPCH scale factor 100, timing in hours:minutes:seconds

	MonetDB/SQL	PostgreSQL	MySQL
Q1	1:02:42	0:53:22	
Q2	0:01:20	0:15:17	
Q3	0:09:53	3:39:38	
Q4	0:05:06	0:02:29	
Q5	0:16:43	0:02:28	
Q6	0:06:15	0:10:14	
Q7	0:14:15	0:22:09	
Q8	0:08:04	6:15:27	
Q9	0:31:45		
Q10	0:41:49	0:03:18	
Q11	0:00:36	0:08:02	
Q12	0:08:39	0:10:13	
Q13	0:36:36	1:20:03	
Q14	0:09:31	0:10:53	
Q15	0:04:43	0:10:16	
Q16	0:37:14	0:29:30	
Q17	0:04:47		
Q18	0:37:37	0:34:34	
Q19	0:03:55	0:38:19	
Q20	0:05:35		
Q21	0:24:15	1:08:10	
Q22	0:01:25		
	load	147m38s + 285m41s	

MonetDB is somewhat slower
MonetDB is >10 x faster
Takes > 10 hr to run
Error, empty result

Afbeelding 2: MonetDB vs PostgreSQL.

De laatste op MySQL gebaseerde oplossing is al eerder in dit blad behandeld (DB/M 3, 2008 en DB/M 8, 2008), alleen in het eerste artikel nog onder de naam Brighthouse. Het product en het bedrijf voeren sinds de overstap naar de open source wereld dezelfde naam: InfoBright. Ook daar staan de ontwikkelingen niet stil. Versie 3.2 is onlangs gelanceerd, wat inhoudt dat steeds meer van de query's afgehandeld worden door de eigen Infobright engine en niet meer deels door MySQL. Infobright is er in twee versies: de Enterprise Editie (IEE) en de Community Editie (ICE). Er is veel energie gestopt in het zo gemakkelijk mogelijk maken voor nieuwe gebruikers om kennis te maken met de software, die dan ook voor de meest gebruikte Linux distributies en voor Windows in zowel een 32- als 64-bits vorm beschikbaar is. Ook kan een tweetal virtuele machines worden gedownload, één in combinatie met de Pentaho BI suite, de andere in combinatie met Jaspersoft, waardoor er met een paar muisklikken een compleet BI-platform kan draaien. Er is echter wel één grote 'maar' aan ICE: er is geen DML-functionaliteit beschikbaar, hiervoor zal overgestapt moeten worden naar de Enterprise versie die dit wel heeft. En een datawarehouse database die geen insert, update en delete ondersteunt, tja, ik blijf het een raar idee vinden. Toch zijn er bedrijven die hier prima mee uit de voeten kunnen. In één van mijn projecten is het central warehouse in PostgreSQL gebouwd dat 's nachts wordt bijgewerkt, en vervolgens worden de Infobright datamarts opnieuw gegenereerd en gevuld. In het tijdperk van de wegwerp datamarts (zie Harm van der Lek in DB/M 5, 2009) misschien helemaal nog niet zo'n gek idee.

Performance benchmark: TPC-H

Allemaal leuk en aardig die verschillende oplossingen, maar hoe vergelijkt u nu de ene analytische database met de andere? Uiteindelijk natuurlijk door een proof of concept in uw eigen organisatie uit te (laten) voeren, met uw eigen data en uw eigen workload. Dat is de enige test die er echt toe doet, voor de rest bent u grotendeels aangewezen op whitepapers en mooie marketingverhalen. Tussen de glossy folders en de POC op locatie is er

echter nog een tussenoplossing die geboden wordt door de benchmarks van de Transaction Processing Council (TPC). Dit is een organisatie waarin vrijwel alle databaseleveranciers vertegenwoordigd zijn om zodoende in staat te zijn een industriebenchmark te ontwikkelen waarmee producten op onafhankelijke wijze vergeleken zouden kunnen worden. De council heeft verschillende benchmarks opgesteld, waarvan de TPC-H speciaal voor BI-toepassingen is bedoeld. De benchmark bestaat uit een set van 22 query's die op een (niet helemaal zuiver) stermodeldatabase worden afgevuurd, waarbij elke query een specifiek doel heeft. Afbeelding 1 laat zien hoe het databasemodel in elkaar zit.

> Performancewinst zal uit het clusteren van de database moeten komen waarbij query's parallel verwerkt kunnen worden

De data worden gegenereerd met een speciaal ontwikkeld programma (dbgen) dat wel eerst voor het eigen besturingssysteem gecompileerd dient te worden. Ook de query's en het schema worden gegenereerd, in dit geval met het tool qgen. Net als bijvoorbeeld in een sport als Judo wordt er gebruik gemaakt van verschillende 'gewichtsklassen' om de resultaten vergelijkbaar te maken, in dit geval gebaseerd op datavolume. Klassen worden aangeduid met hun 'Scale Factor' (SF). Het uitgangspunt is een SF1 data set, waarbij de 1 staat voor 1 Gigabyte. De feittabel (lineitem) bevat in een SF1 dataset 6 miljoen records en ongeveer 75 procent van de totale dataomvang. De officiële TPC-H ranking begint bij SF100, dus een 100 GB dataset met 600 miljoen rijen in de fact table. Vervolgens is er een SF300, SF1000, SF3000 en SF10000 ranking. De grootste gepubliceerde benchmark is een SF30000, inderdaad een 30 Terabyte dataset oftewel 180 miljard records in de feittabel, niet iets wat je thuis op de PC kwijt kunt. Kleinere datasets zijn echter prima te behappen, en in de open source wereld circuleren ook verschillende lijstjes waarop producten vergeleken worden met 1, 2, 5, 10, 20 en soms 100 GB datasets. Meestal betreft het vergelijkingen waarbij aan de ene kant een column store, zoals MonetDB of LucidDB, vergeleken wordt met MySQL en PostgreSQL, de standaard transactiedatabases die vaak ook voor (kleinere) datawarehouses of datamarts worden ingezet. Nu is de TPC-H test allesbehalve een benchmark om puur de prestaties van databases te vergelijken, maar zijn de resultaten zwaar afhankelijk van de gebruikte hardware. Veel van de query's forceren een full-table scan op de feittabel en zijn dus met name I/O gebonden wat weinig zegt over de software-, maar veel meer over de hardwareprestaties. Zie bijvoorbeeld query 1:

```
select
  l_returnflag,
  l_linestatus,
  sum(l_quantity) as sum_qty,
  sum(l_extendedprice) as sum_base_price,
  sum(l_extendedprice * (1 - l_discount))
                                  as sum_disc_price,
  sum(l_extendedprice * (1 - l_discount) *
                        (1 + l_tax)) as sum_charge,
  avg(l_quantity) as avg_qty,
  avg(l_extendedprice) as avg_price,
  avg(l_discount) as avg_disc,
  count(*) as count_order
from
  lineitem
where
  l_shipdate <= cast(date '1998-12-01' -
                        interval ':1 days' as date)
group by
  l_returnflag,
  l_linestatus
order by
  l_returnflag,
  l_linestatus;
```

Dat is dan ook de reden dat de benchmark voor veel controverse zorgt tussen diverse analisten, met Curt Monash als meest uitgesproken (en bekende, zie www.dbms2.com) criticaster. Zonder er al te diep op in te gaan snijdt een deel van de kritiek wel hout en moet er zeer kritisch naar de resultaten gekeken worden. Ook zorgen de fabrikanten er wel voor dat de cijfers nauwelijks vergelijkbaar zijn doordat ze allemaal (bewust) gebruik maken van verschillende hardware- en softwareconfiguraties. Dat probleem heeft u echter niet als u in eigen huis producten vergelijkt. De TPC-H benchmark is, los van de eerder besproken minpunten, wel een prima middel om het kaf van het koren te scheiden als er een keuze gemaakt moet worden voor een nieuwe database.

Zodra de omvang van de dataset toeneemt zien we product na product afhaken, omdat query's oneindig lang (meer dan tien uur) lopen, verkeerde resultaten teruggeven of simpelweg met een error het proces afbreken. Op een relatief lichte omgeving (enkele quadcore processor met 8 GB geheugen) wordt direct duidelijk dat al vanaf een 2 GB dataset forse verschillen optreden tussen de transactie- en de analytische databases. Deze verschillen lopen exponentieel op, en al snel blijkt dat MySQL bij 100 GB helemaal afhaakt zoals in afbeelding 2 is te zien (database kon niet geladen worden). De tabel toont een experimenteel resultaat van een TPC-H test uitgevoerd bij het Centrum voor Wiskunde en Informatica (CWI), de kraamkamer van MonetDB en Vectorwise.
In de tabel is te zien dat query 1 op deze hardware meer dan een uur nodig heeft om de resultaten terug te geven; wie de TPC-H

publicaties op www.tpc.org doorkijkt zal zien dat het met meer 'ijzer' allemaal véél sneller kan. De snelste tijd die ik recentelijk met een commerciëel product op een enkele (dus ongeclusterde) machine geklokt heb met query 1 en een 100 GB dataset bedraagt overigens 5,55 seconden. Daar komt momenteel geen open source database bij in de buurt. Sterker nog, als u de gepubliceerde top 10 non-clustered TPC-H resultaten gesorteerd op performance bekijkt, zult u zien dat dit meer dan tien keer beter is dan nummer 1, SQL Server 2008, en ook nog altijd zes keer sneller dan de met een SQL-chip gemotoriseerde Kickfire appliance. Het ziet er dan ook naar uit dat niet alle marketingverhalen per definitie overdreven zijn.

Nieuwe ontwikkelingen

Tot nu toe ging het in dit artikel voornamelijk om ofwel speciale DWH producten, ofwel uitbreidingen op basis van MySQL en PostgreSQL. Maar er zijn toch wel meer open source databases, hebben die niet zoiets? Nou, nu misschien nog niet, maar bij Ingres is het datawarehouse-kwartje ook gevallen. Hoewel Ingres één van de oudste RDBMS'en ter wereld is wordt het voor datawarehouses niet echt veel ingezet, althans niet meer sinds het op Ingres gebaseerde DATAllegro is overgenomen door Microsoft. Dat zou komend jaar wel eens drastisch kunnen veranderen door de toevoeging van VectorWise. Hoewel VectorWise net als MonetDB bij het CWI is ontstaan en veel van de achterliggende ideëen met MonetDB gemeen heeft, betreft het een compleet nieuw ontwikkeld product. Wie meer over het onderzoek wil weten dat ten grondslag ligt aan VectorWise kan op de CWI site op zoek gaan naar het MonetDB/X100 project.

De combinatie Ingres/Vectorwise heeft pas onlangs zijn geboortekaartje de wereld ingestuurd

De combinatie Ingres/Vectorwise heeft pas enige weken geleden zijn geboortekaartje de wereld ingestuurd, en de daaropvolgende kraamvisite die ik op uitnodiging aflegde liet zien dat dit een product is dat het komend jaar voor aardig wat opschudding zal gaan zorgen in de databasewereld. Ook bij MonetDB zelf zit men niet stil. Momenteel is men met een tweetal veelbelovende onderzoeken bezig, één met de al eerder genoemde codenaam Octopus en een ander onder de vlag Cyclotron. Octopus is er op gericht om MonetDB op multi-node clusters te kunnen draaien (dat is niet zo nieuw) waarbij de capaciteit dynamisch wordt aangepast aan de beschikbare hoeveelheid machines (en dat is weer wél nieuw). Cyclotron is wat lastiger in een paar zinnen uit te leggen, maar komt neer op een cluster van machines die in een ringvorm verbonden worden waarbij data- en resultaatsets in cache continu als in een centrifuge rondgeslingerd worden door de ring van machines. Vervolgens maakt het niet uit welke machine benaderd wordt; de gemiddelde query-resultaten zullen in alle gevallen gelijk zijn. Ook hier weer is schaalbaarheid gepaard aan performance het uiteindelijke doel. Wanneer er producten op de markt komen op basis van deze technieken is moeilijk te zeggen (misschien wel nooit), maar wat wel duidelijk is, is dat het CWI op het gebied van databaseonderzoek wereldwijd één van de toonaangevende instituten is. Onlangs heeft het CWI-team bijvoorbeeld tijdens VLDB 2009 (International Conference on Very Large Data Bases) de '10 year best paper award' in ontvangst mogen nemen. Niet alleen voor het beste paper over de afgelopen tien jaar maar tevens voor het pionierswerk op het gebied van column stores. En daar mogen we best een beetje trots op zijn!

Conclusie

Het moge duidelijk zijn: als u de open source route kiest en op zoek gaat naar een snelle analytische database is de spoeling vooralsnog erg dun. Toch zou u in elk geval eens moeten kijken naar LucidDB, MonetDB en InfoBright. Downloaden en installeren is in de meeste gevallen een fluitje van een cent. Wilt u een bestaande MySQL-omgeving een flinke boost geven, kijk dan eens naar Kickfire. Voor het echt zware werk bent u vooralsnog aangewezen op commerciële producten, die (met uitzondering van Sybase IQ) ook nog eens grotendeels niet vertegenwoordigd zijn op de Nederlandse markt. En zoals gezegd, met Calpont en Ingres/Vectorwise in aantocht zou het er over een jaartje wel eens heel anders uit kunnen zien.

Jos van Dongen (jos@tholis.com) is onafhankelijk adviseur, auteur en spreker.
Met dank aan: Martin Kersten en Peter Boncz (MonetDB), Marcin Zukowski (Vectorwise B.V.).

Meer informatie

RethinkDB: www.rethinkdb.com/
Schooner: www.schoonerinfotech.com/
Virident: www.virident.com
Infobright: www.infobright.org
LucidDB: www.luciddb.org
MonetDB: www.monetdb.com
Vectorwise: www.vectorwise.com
Calpont: www.calpont.com
GridSQL: www.enterprisedb.com/community/projects/gridsql.do
Greenplum Network: http://bgn.greenplum.com