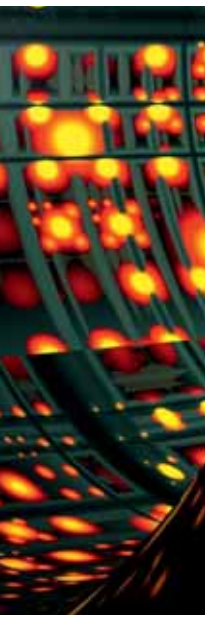

Het databaselandschap is diverser dan ooit

Baas boven databaas

Jos van Dongen

Jarenlang waren databases 'saai', althans voor het grote publiek. Niet voor dit blad gelukkig, waar altijd wel ruimte en aanleiding was om aandacht te besteden aan het fenomeen 'database'. Sterker nog, de naam Database Magazine reflecteert nog steeds aan de roots van de uitgave waar in eerste instantie bijvoorbeeld de SQL versus QUEL en relationele versus netwerk en hiërarchische databases veldslagen werden uitgevochten. Welnu: die tijd is weer helemaal terug!

De huidige discussies gaan nog steeds over SQL, maar dan versus Map/Reduce, of over columnar versus row-based, MPP versus SMP, Open versus Closed Source, Cloud versus On-Premise, in-memory versus disk based, en SSD's versus 'spinning' disks. Dit artikel neemt al deze onderwerpen onder de loep en zorgt er hopelijk voor dat u de bomen weer door het bos kunt zien.

De Analytics revolutie

De ontwikkelingen in de database-industrie die de afgelopen jaren hebben plaatsgevonden hebben natuurlijk een oorzaak. Het is moeilijk om er één specifiek aan te wijzen, maar de gemeenschappelijke noemer 'analytics' lijkt een mooi uitgangspunt. We willen domweg meer doen met beschikbare data; niet alleen rapportjes bakken, maar ook verbanden ontdekken, voorspellingen proberen te doen, bijvoorbeeld over klantgedrag, en we zijn ook steeds meer geïnteresseerd in wat er buiten de bedrijfsmuren gebeurt. Wat hebben consumenten over ons bedrijf te melden op Twitter, Facebook en LinkedIn? En hoe kunnen we zinvolle informatie halen uit alle gegevens die zijn opgeslagen buiten het datawarehouse? Ook willen we meer mensen toegang geven tot deze informatie en dankzij het 'Google effect' worden we met zijn allen steeds ongeduldiger als het gaat om responstijden.

Kortom: meer en meer verschillende soorten data, meer verschillende vragen en meer gebruikers met ook weer verschillende behoeften. Meer, meer, meer, maar ja, vaak ook: minder geld, minder resources en minder tijd. Het goede nieuws is dat er voor alle verschillende behoeften wel een bevredigend antwoord is, maar het is helaas net als met de schoenmaker die een bord op de deur had hangen met: "wij leveren kwaliteit, snelheid, en een lage prijs, maar u kunt maar twee van de drie opties tegelijk kiezen". Zo is het ook met de huidige generatie databases: er is niet één die aan alle behoeften tegelijk kan voldoen, en dat is ook precies de reden dat het databaselandschap diverser is dan ooit.

No(n)-SQL databases & Map/Reduce

Elders in dit nummer vindt u op pagina 13 een artikel over NoSQL- en non-SQL databases, een nieuwe generatie producten die met name ontwikkeld is voor grootschalige webtoepassingen. Veel van deze databases vinden hun oorsprong bij een enkel bedrijf dat een oplossing ontwikkelde voor een specifiek probleem. Google was één van de eerste met Bigtable, en het nu algemeen toegepaste Map/Reduce algoritme voor het doorzoeken van grote hoeveelheden gedistribueerde data komt ook uit dezelfde koker. Een bedrijf als Facebook laat zich op dit terrein ook niet onbetuigd en heeft het intern ontwikkelde Cassandra in 2008 als open source project uitgebracht. Later werd Cassandra een volwaardig Apache project.
Zo zijn er nog vele voorbeelden van dergelijke producten die technisch gezien behoorlijk verschillend zijn maar toch allemaal een paar gemeenschappelijke kenmerken hebben: automatische datadistributie, 'oneindige' schaalbaarheid, een flexibeler datamodel dan de klassieke SQL databases en een API die meer opties dan SQL biedt voor het bevragen en analyseren van de data. Soms is er ook helemaal geen sprake van SQL ondersteuning en dient alles uitgeprogrammeerd te worden in C(++), Java, Python, Ruby of Perl. Vrijwel alle grootschalige websites en sociale netwerken worden tegenwoordig aangedreven door no(n)-SQL databases. Het is natuurlijk heel indrukwekkend om te vernemen dat de Facebook database de 15 Petabyte grens is gepasseerd, maar de vraag is natuurlijk wat dit voor de datawarehousewereld betekent. En dat valt nogal mee (of tegen natuurlijk): de meeste No(n)-SQL databases zijn totaal ongeschikt voor

de meeste BI-toepassingen zoals we die momenteel gebruiken. Een concept als 'eventual consistency' is prachtig voor sociale netwerken, maar onbruikbaar in een BI-omgeving waar we graag keer op keer hetzelfde antwoord op dezelfde vraag zien, los van het exacte tijdstip waarop de vraag wordt gesteld. De gemiddelde retailer, zorginstelling of verzekeraar komt ook niet in de buurt van de datavolumes die Yahoo, LinkedIn, eBay of Twitter genereren en ook de manier waarop de data worden gebruikt verschilt nogal. Betekent dit dat de No(n)-SQL databases voor het grote publiek maar meteen in het database-rariteitenkabinet bijgezet moeten worden naast bijvoorbeeld de Object Oriented en XML databases? Dat is wellicht iets te kort door de bocht. Ook voor traditionele bedrijven zal de analyse van sociale netwerken en sociale media steeds belangrijker worden, en de grote hoeveelheden on- en semigestructuurde data (e-mail!) die binnen de bedrijfsmuren aanwezig zijn kunnen veel beter in een no(n)-SQL database worden opgeslagen en doorzocht dan in een 'gewone' relationele database. Dat geeft ook meteen antwoord op de vraag 'negeren, experimenteren of adopteren?', dat minimaal de tweede optie zou moeten zijn.

Hardware-ontwikkelingen

Het zijn niet alleen de database-ontwikkelingen zelf die in een stroomversnelling zijn geraakt, ook de ontwikkelingen op hardwaregebied zorgen ervoor dat er steeds meer mogelijk wordt. Intel krijgt het voor elkaar om bij gelijkblijvende prijzen elke nieuwe generatie CPU's twee tot driemaal meer verwerkingscapaciteit te geven. De huidige generaties Xeon 5600 (mid-range) en 7500 (high-end) presteren significant beter dan de vorige 5500 en 7400 generatie, met daarbij ook nog eens een fors lager energieverbruik.
Concurrent AMD zit niet stil en richt zich met name op grote rekencentra met voordelige en energiezuinige multicore CPU's; op het moment van schrijven biedt AMD al een 12-core CPU aan, waarbij Intel niet verder komt dan 6 cores. Op het gebied van brute kracht moeten de AMD's het nog steeds afleggen tegen Intel, maar voor virtualisatiedoeleinden is een server met vier 12-core CPU's natuurlijk een uitermate geschikt platform.

De geheugenontwikkelingen gaan natuurlijk ook steeds verder, maar hier is iets vreemds aan de hand: jarenlang gingen de prijzen per GB gestaag omlaag, maar over de afgelopen 12 maanden zien we een omgekeerde trend, waarbij de prijzen gemiddeld genomen met zo'n 50 procent (!) gestegen zijn. Met 'alles in-memory' is het dus nog even wachten geblazen.

De grootste revolutie op hardwaregebied heeft zich echter op een ander vlak voltrokken: opslag! De term SSD (Solid State Drive) is inmiddels genoegzaam bekend en iedereen die zijn mechanische disk vervangen heeft door een solid state variant weet wat dat voor effect heeft. Zeker voor databasetoepassingen lijken SSD's het ei van Columbus: de toegangstijden zijn zo'n 50 keer korter dan bij de snelste mechanische schijven, de lees- en random I/O-snelheid is aanmerkelijk hoger, het energieverbruik significant lager en ze zijn ook nog eens veel kleiner in omvang. Dit is echter niet het hele verhaal: de opslagcapaciteit kan nog steeds niet tippen aan die van traditionele schijven, de prijs per GB is fors hoger (per GB van 2,- tot 9,- euro ten opzichte van 0,07 euro voor SATA disks), en het belangrijkste nadeel: minder duurzaamheid, veroorzaakt door de 'write fatigue'. SSD's hebben de vervelende eigenschap dat ze maar een beperkt aantal keer beschreven kunnen worden, en dat de performance in de loop van de tijd ook nog eens afneemt. De duurdere 'enterprise' SSD's hebben hier minder last van (en zijn derhalve ook veel duurder) maar het is wel iets waar u rekening mee dient te houden.

Wat betekent dit nu allemaal? Simpel: wanneer uw huidige database server een jaar of drie of vier oud is lijkt het raadzaam om eens naar een hardware upgrade uit te kijken. Gecombineerd met een upgrade van operating system en databaseversie moet u niet verbaasd zijn als dit al een drie tot vijf keer betere performance van uw datawarehouse oplevert, zonder dat u hoeft te kijken naar column stores, in-memory databases of MPP-oplossingen.

Classificatie

Al eeuwenlang proberen mensen de wereld om zich heen te classificeren; de biologie kent een uitgebreide taxonomie van alles wat er leeft, de scheikunde heeft ons verrijkt met de periodieke tabel der elementen en de geologie kent een indeling gebaseerd op tijdperken. Deze indelingen zijn echter niet zonder slag of stoot tot stand gekomen, en nog steeds vinden er verhitte discussies plaats over de diverse indelingen. Ook voor databases zijn er verschillende indelingen of classificaties mogelijk; een heldere taxonomie zou mooi zijn maar wellicht iets te ambitieus. Dus laten we eens kijken naar manieren waarop een indeling gebaseerd zou kunnen zijn:

- Opslagstructuur: bestanden, tabellen, kolommen, cubes, key/value pairs;
- Opslagarchitectuur: in-memory, diskbased, datacompressie;
- Schaalbaarheid: SMP, Clustered, MPP;
- Leveringsmodel: SaaS, Cloud, Appliance, On-Premise;
- Interface: SQL, Map/Reduce, C(++), Java, Python, Perl enzovoort;
- In-DB Analytics: UDF's, SAS, R, Proprietary functie-bibliotheken;
- Snelheid: TPC-H scores, laadsnelheid;
- Licentie/prijs: per CPU, node, datavolume, geheugengebruik, functionaliteit, open versus closed source.

Het lastige van al deze categorieën is het feit dat de indeling vaak niet éénduidig is, en dat er ook weer onderverdelingen bestaan. Het label 'SQL' bijvoorbeeld zegt op zich nog niet zoveel; pas als er wat verder gekeken wordt blijken sommige producten geen primary of foreign keys te ondersteunen, of niet in staat om correlated subquery's te herschrijven. Slechts weinige

van de nieuwe generatie analytische databases ondersteunen de volledige SQL2003 standaard inclusief Windowing functies, en ook iets essentieels als het kunnen uitvoeren van een online backup is in lang niet alle gevallen aanwezig. Hieronder zullen we al deze categorieën in meer detail behandelen.

Opslagstructuur

Elke database hanteert zijn eigen wijze van dataopslag. Uiteindelijk zijn het natuurlijk allemaal bits en bytes, maar het gaat om de manier waarop deze zijn georganiseerd die het interessant maakt. De laatste jaren zijn vooral de column stores sterk in opkomst. Denk hierbij aan producten als Sybase IQ, Vertica, ParAccel, KickFire, Infobright en onze 'eigen' MonetDB en Ingres/VectorWise. In DB/M 8, 2008 heeft u kunnen lezen wat een column store onderscheidt van row-based producten als SQL Server en Oracle, dus dat kunnen we hier gevoeglijk achterwege laten. Het interessante aan column stores is dat steeds meer producten die een eigen in-memory engine meeleveren ook op basis van kolommen werken. Denk hierbij aan Microsoft PowerPivot, LyzaSoft en het Australische Yellowfin. Het is ook niet meer of/of, maar steeds vaker 'en'. Oracle Exadata v2 levert een hybride opslagstructuur, en ook Greenplum ondersteunt een hybride vorm van row/column based opslag. Nog flexibeler zijn de diverse 'NoSQL' producten zoals bijvoorbeeld Cassandra, dat in essentie een key/value pair structuur kent. Een speciale categorie wordt nog steeds gevormd door de OLAP databases zoals Microsoft Analysis Services, Hyperion Essbase en Jedox Palo die voorgeaggregeerde data in cubes opslaan om snelle analyses, write back (ten behoeve van 'what if'-vragen) en consolidatie te ondersteunen.

Opslagarchitectuur

Dit lijkt iets om nauwelijks over na te denken, wat dus helaas nog steeds op grote schaal gebeurt. Toch is het wel degelijk van belang om na te denken of en in hoeverre uw (geplande) database gebruik kan maken van een forse hoeveelheid RAM-geheugen, of dat het mogelijk is om een onderscheid te maken tussen 'hot', 'warm' en 'cold' data, waarbij elk type data een eigen storagetype toegewezen krijgt: de actuele, frequent geraadpleegde data in-memory; de gerelateerde grotere hoeveelheid data nodig voor trendanalyses op SSD's; en de data die af en toe nodig zijn op standaard SAS- of zelfs SATA-schijven. Het mooiste is uiteraard als een database op basis van gebruik zelf deze datadistributie voor zijn rekening neemt, iets waarmee Teradata ver voorloopt op de concurrentie. Oracle doet iets vergelijkbaars maar dan op basis van de leeftijd van de data, en Sybase IQ kan specifieke partities aan verschillende typen opslag toewijzen. Dit laatste kunt u bij veel andere producten ook wel op de een of andere manier voor elkaar krijgen, hoewel de beheerlasten hiervan navenant toe zullen nemen in verband met de benodigde DBA-inspanning. Overigens betekent 'in-memory' niet dat er geen persistentielaag aanwezig is; die is er altijd wel. De wijze waarop RAM wordt gebruikt kan ook nog verschillen; soms is het alleen een intelligente cachinglaag, soms worden ook daadwerkelijk delen van de database (of de gehele database) in-memory geladen.

Los van het medium waarop de data worden opgeslagen, kan er ook met de data zelf nog behoorlijk wat winst behaald worden als gevolg van compressie. Moderne CPU's zijn zo snel dat de kosten van compressie en decompressie ruimschoots opwegen tegen de te behalen I/O-winst. Column stores lenen zich van nature uitstekend voor compressie, wat u ook terug zult zien in het benodigde opslagvolume. In de meeste gevallen zal de database-omvang in dit type databases kleiner zijn dan de omvang van de geladen gegevens; een factor twee of drie is vrij normaal, terwijl een product als Infobright nog een stapje verder gaat en in veel gevallen in staat is om 1:10 en soms nog beter te bereiken.

Schaalbaarheid

Er is geen enkel databaseproduct dat meer dan een machine vereist om te kunnen draaien, dus of SMP wordt ondersteund is niet relevant. Ook kan altijd een SAN gebruikt worden voor de opslag van grote hoeveelheden data, dus ook dat voegt weinig toe. Concurrency en responstijden zijn echter wél van belang, dus er zijn wel degelijk grenzen aan een 'scale up' (uitbreiden van 1 machine) benadering. 'Scale out' is een tweede strategie waarbij naar behoefte extra machines kunnen worden aangeschakeld voor zowel opslag als queryverwerking. Wanneer elke machine ook nog over een eigen, individuele dataopslag beschikt spreken we over een 'shared nothing' architectuur, zo niet dan gaat het om clustering. Dit laatste is bijvoorbeeld te zien bij Sybase IQ en Calpont InfiniDB die nog steeds een gedeeld dataopslagmodel kennen en dus geen zuivere MPP (Massive Parallel Processing) oplossingen zijn. Scale out is ook nog steeds het zwakke punt van de open source column stores als MonetDB, LucidDB en Ingres/VectorWise. De no(n)-SQL producten hebben hier echter over het algemeen geen enkele moeite mee en zijn juist ontworpen voor grootschalige gedistribueerde omgevingen. Een ander interessant ontwerpdetail is het 'design for failure' principe: het is niet erg als er tijdens het verwerken van een aanvraag een machine crasht, aangezien de automatische datadistributie en replicatie zorg dragen voor voldoende redundantie, zodat altijd een correct antwoord gegeven kan worden. Dit geldt trouwens niet alleen voor de No(n)-SQL databases: de meeste MPP-producten hanteren dit failover-principe.

Leveringsmodel

Cloud computing wordt steeds populairder, en sommige leveranciers bieden dan ook de optie om hun product als 'cloud' solution te gebruiken, waarbij de oplossing bij één van de vele cloud providers wordt ondergebracht. In andere gevallen wordt een SaaS-oplossing geboden, waarbij de leverancier zélf de hosting en support verzorgt op dedicated hardware. Kognitio is een mooi voorbeeld van deze laatste aanpak. Een andere vorm van levering is de door Netezza populair gemaakte Appliance, waarbij

	Structuur	Opslag	Compressie	Schaalbaarheid	Leveringsmodel	Interface	In-DB Analytics	Proprietary HW	Licentie
Aster Data	R	D	D	M	O, A, S	S, M	U, P		S
Dataupia	R	D		M	A	S			S
EXASOL	C	M	M	M	O	S			M
Greenplum	R	D		M	O	S, M	R		S
Greenplum SNE	R	D		S	O	S, M	R		F
GridSQL	R	D		M	O	S			F
HP Neoview	R	D		M	A	S	U	X	S
Illuminate	K	M		S	S	S, P	P		R
InfiniDB CE	C	D		S	O	S			F
InfiniDB EE	C	D		C	O	S			N
Infobright CE	C	D	D	S	O	S			F
Infobright EE	C	D	D	S	O	S			I
Ingres/VectorWise	C	D		S	O	S			C
Intersystems Caché	B	H		C	O, S	S, X, P	U, P		S, C, U, F
Jedox Palo	O	M		S	O	X			U
Kognitio	R	H		M	O, A, S	S			S
LucidDB	C	D		S	O	S			F
MonetDB	C	M		S	O	S			F
MS Analysis Services	O	H		C	O	X			S, C, U
MS PowerPivot	C	M	M	S	O	P			F
Netezza	R	D	D	M	A	S, M	U, R, P	X	S
Oracle Essbase	O	D		S	O	X			S, C, U
Oracle ExaData	H	D		C	A	S	U, P	X	C, S
ParAccel	C	H	D	M	O, A	S			I
SAP B/W	O	H		S	O	X			S, C, U
SQL 2008 R2 PDW	R	D		M	A	S	U, P		C
Sybase IQ	C	D	D	C	O	S	U		S, C
Teradata	R	H	D	M	A	S, M	S		S
Vertica	C	D	D	M	O, C	S			I
XtremeData	R	D		M	A	S	U	X	L

Legenda

Structuur
- B Object Oriented
- C Columns
- H Hybrid
- K Key/Value pairs
- O OLAP
- R Rows

Opslag
- D Disk
- H Hybride
- M Memory

Compressie
- D Disk
- M Memory

Schaalbaarheid
- C Clustered (shared data)
- M MPP
- S SMP (single node)

Leveringsmodel
- A Appliance
- C Cloud
- O Software only
- S SaaS

Interface
- M Map/Reduce
- P Proprietary
- S SQL
- X MDX, XML/A

In-DB Analytics
- P Proprietary
- R Cran/R
- S SAS
- U User defined functions

Licentie
- C CPU
- F Free
- I Input data volume
- L Loaded data volume
- M Memory used
- N Node
- R Records loaded
- S Server
- U User

een stekkerklare oplossing op basis van proprietary hardware geleverd wordt. De meeste appliances echter zijn tegenwoordig gebaseerd op standaard hardware, aangezien de snelheid in ontwikkeling van standaard componenten nauwelijks bijgehouden kan worden. U koopt daarbij nog steeds een volledig voorgeconfigureerde oplossing maar bent niet gebonden aan leverancierspecifieke hardware. Bij Oracle en HP ligt dit overigens net even anders; een Oracle Exadata appliance is uiteraard gebaseerd op SUN hardware, en HP's NeoView is vanzelfsprekend alleen maar op basis van HP hardware verkrijgbaar. De leveringsmodellen beginnen ook steeds meer diffuus te worden: er zijn diverse partijen, waaronder het Nederlandse Inergy, dat Netezza als BIaaS/DaaS-oplossing levert, maar het primaire model waarmee de leverancier werkt is in dit geval nog steeds een appliance.

Interface/programmeermodel

Dit is een interessante categorie, en ook eentje waarbij u zorgvuldig de kleine lettertjes, manuals en reference guides zult moeten doorspitten. Zoals al eerder gemeld noemen vrijwel alle leveranciers één of andere vorm van SQL support, maar vervolgens blijkt dat 'SQL2003 compliant' niet betekent dat deze standaard volledig is geïmplementeerd. Er zijn zelfs analytische databaseleveranciers die primary en foreign keys of zelfs simpele unique constraints maar onzin vinden. Gemakshalve wordt dit ook maar niet in de documentatie vermeldt, men zou eens lastige vragen kunnen gaan stellen. Gelukkig zijn er ook positieve uitzonderingen zoals EXASOL en Ingres/VectorWise die keurig een sectie 'unsupported features' opnemen. Het andere uiterste is Kognitio dat zo ongeveer álles kan wat ooit in de SQL-wereld bedacht is, en meer. Maar goed, het uitvoeren van een basis SELECT FROM statements lukt over het algemeen wel, wat niet gezegd kan worden van de No(n)-SQL databases die in veel gevallen helemaal geen SQL interface hebben. Dit betekent een aantal dingen. Ten eerste is er geen aansluiting met de meeste standaard BI-tools. Ten tweede dienen de query's geprogrammeerd te worden in één van de ondersteunde talen, waarbij er een aantal keuzes beschikbaar is (meestal Java, C++, Ruby, Perl en Python). Het gevolg is dat deze producten nog nauwelijks geschikt zijn voor bedrijven die niet over eigen software-ontwikkelaars beschikken, tenzij men deze wil inhuren of maatwerkoplossingen wil laten bouwen.

In-database analytics

Voorheen bestond er een strikte scheiding tussen datamining en statistische analyse aan de ene kant, en datawarehousing aan de andere. Dit is echter in rap tempo aan het veranderen. Diverse leveranciers hebben al functiebibliotheken ín de database gestopt, zodat alle voordelen van een parallele verwerking ook opgaan voor de datamining algoritmen. Teradata 'doet' dit met SAS waar Netezza en Greenplum de 'R' bibliotheek hebben geïntegreerd. Er vindt ook veel eigen ontwikkeling plaats op dit gebied, bijvoorbeeld door Aster Data dat een unieke SQL-MapReduce aanpak gebruikt en op basis daarvan kant en klare analytische functies beschikbaar stelt die in een normaal SQL statement zijn aan te roepen. Zoals in het overzicht van de analytische databases op pagina 19 te zien is, zijn het de gevestigde namen en een paar innovatieve partijen die méér doen dan SQL alleen, en ook user defined functies zijn nog lang geen gemeengoed bij de analytische databaseleveranciers.

Snelheid

Over 'snelheid' kunnen we kort zijn: ja, er zijn standaard benchmarks zoals TPC-H en ja, leveranciers schermen graag met 'wij versus product X' vergelijkingen. Maar hoe leuk spelen met databases en verschillende benchmarks ook mag zijn: het zegt helemaal niets over hoe een product zich gedraagt met úw workload en úw data. Er is daarom maar één zinnig advies in deze: doe *altijd* een Proof of Concept met uw eigen data, uw eigen query's en uw eigen BI-tools. Dit geldt overigens voor zowel query- als laadsnelheid. Het is leuk dat ParAccel 9 TB per uur kan laden, maar als u niet over het daarvoor gebruikte 48-node cluster beschikt is zo'n cijfer niet echt bruikbaar.

Licentievorm

Vroeger was misschien niet alles beter maar in elk geval wel overzichtelijker, zeker als het gaat om databaselicenties. Er waren enkele knopjes om aan te draaien zoals aantal gebruikers, aantal servers en aantal CPU's maar daar hield het dan ook wel mee op. Die tijd is voorbij; vergelijken van databaselicentievormen en bijbehorende prijzen is bijna onbegonnen werk. De meeste nieuwe analytische databases kennen een licentie op basis van datavolume waarbij niet gekeken wordt naar opslagvolume maar naar de hoeveelheid input data, en hoe meet je zoiets? Een product als EXASOL baseert de prijs op de hoeveelheid toegewezen RAM-geheugen, terwijl het uit de as herrezen Dataupia per 2 TB node beprijst. XtremeData is voor zover mij bekend de enige die rekent op basis van user volume (hoeveelheid opgeslagen data in de database), wat weer gevolgen kan hebben voor hun inspanningen op het gebied van compressie. Uiteraard zijn er dan ook nog de open source leveranciers, die veelal een splitsing maken tussen 'Community' en 'Enterprise' editions. De CE is het 'echte' open source product, vaak met beperkingen, terwijl de omzet behaald wordt door de verkoop van de EE lijn die bijvoorbeeld wél schaalbaar is of wél een parallel loader of DML support heeft. In het overzicht zijn deze producten dan ook als twee afzonderlijke entiteiten zichtbaar.

Tot slot

Een overzicht als dit kan niet meer zijn dan een momentopname. Toch geeft het wel aan waar het heen gaat: 'big data' en meer analytische mogelijkheden dan SQL alleen. De markt die in eerste instantie is ontgonnen door Teradata en Sybase IQ en later met name door Netezza in beweging is gezet, is blijkbaar ook voor een partij als Oracle interessant geworden getuige het Exadata offensief. Microsoft heeft met Excel PowerPivot, Analysis Services en hun Parallel Data Warehouse meerdere ijzers in het vuur en kan op alle niveaus van zowel datavolume als analytisch vermogen een stevig partijtje meeblazen. Voor vernieuwende benaderingen op het gebied van architectuur en data-analyse is het zaak om partijen als Greenplum, Aster Data en XtremeData in de gaten te houden. Zoekt u performance in een klassieke BI/SQL-omgeving, dan zijn de verschillende column stores zoals Vertica, ParAccel en Ingres/VectorWise nog steeds onverslaanbaar; reden waarom er ook door bijvoorbeeld Oracle dankbaar gebruik gemaakt wordt van column based technieken. Wilt u daarentegen een product dat zo ongeveer alles kan doen, inclusief OLAP, data- en text mining, kijk dan nog vooral eens naar IBM InfoSphere Warehouse.

Jos van Dongen (jos.van.dongen@deltiqgroup.nl) is Associate en Principal bij DeltIQ Group.