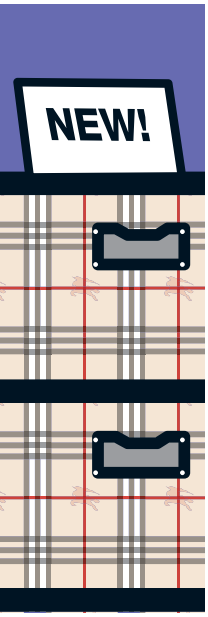

Volledige OS BI stack bereikbaar voor steeds meer organisaties

Trends in Open Source

Jos van Dongen

Open Source 'rocks', zoveel mag inmiddels wel duidelijk zijn. Nu zelfs partijen als Accenture en Gartner Group conferenties rondom dit onderwerp organiseren, kunnen we concluderen dat Open Source langzamerhand als een serieus alternatief wordt gezien en het hobby/ zolderkamer imago definitief achter zich heeft gelaten.

Een andere belangrijke indicator hiervoor is het geld dat door venture capitalists in Open Source ondernemingen wordt gestoken. Deze clubs zitten voor het geld in de wedstrijd en zijn nog maar zelden betrapt op liefdadige beweegredenen. Het is dan ook goed om de ontwikkelingen in het afgelopen jaar onder de loep te nemen en te kijken welke mogelijkheden dit voor het komende jaar te bieden heeft.

Follow the money

Kijkend naar de investeringen die het afgelopen jaar met name in de Verenigde Staten zijn gedaan vallen enkele zaken op. De overname en consolidatiegolf uit de 'gesloten' wereld begint ook door te dringen tot de Open Source projecten en bedrijven.

> Het grootste nieuws op BI-gebied is echter Maria

Een paar recente grote overnames: Linux leverancier Red Hat koopt applicatie-server JBoss en hardware-boer en Java-hoeder Sun koopt database-leverancier MySQL en is bezig met een enorm project om een tweede generatie Java applicatie-server (en nog veel meer) te bouwen. Het gaat echter niet alleen om overnames; er wordt ook veel geïnvesteerd in verschillende startups, en het gaat hierbij om serieuze bedragen. In 2008 ging het tot en met Q3 om 397,8 miljoen dollar, fors meer dan de 286,1 miljoen dollar in de eerste drie kwartalen van 2007 en zelfs meer dan de 381,4 miljoen dollar uit het recordjaar 2006 dat uitkwam op 546,3 miljoen dollar. Om een indruk te geven van de omvang van deze bedragen: het is ongeveer tien procent van het totale venture capital dat in software wordt geïnvesteerd. Toch gaat de economische neergang niet ongemerkt aan de Open Source wereld voorbij. De investeringen in Q3 zijn een stuk lager dan in hetzelfde kwartaal vorig jaar en de verwachtingen voor Q4 zijn niet al te hoog gespannen. De twee meest in het oog springende investeringen binnen het BI-domein waren wel de 10 miljoen dollar elk die zijn geïnvesteerd in Infobright en EnterpriseDB. Het ging hierbij in beide gevallen om een 'series-C' funding, wat zoveel wil zeggen dat de producten op orde zijn en dat nu de markt veroverd dient te worden. Interessant aan deze gevallen is dat Infobright tegelijkertijd de stap van proprietary naar Open Source heeft gezet en hiermee een duidelijk andere richting inslaat dan concurrenten Vertica en ParAccel. Ook database-leverancier EnterpriseDB heeft een vergelijkbare stap gemaakt door een deel van zijn voorheen proprietary oplossing als 'PostgreSQL Plus' Open Source te maken. Een andere BI-speler die van meet af aan als Open Source speler door het leven gaat is Pentaho die in februari van dit jaar nog 12 miljoen dollar durfkapitaal bij kon schrijven om verdere groei te financieren. Hiermee liep het bedrijf wel een paar maanden achter op grote concurrent Jaspersoft die hetzelfde bedrag al in oktober 2007 in de boeken kon noteren.

Van 'me too' naar 'me first'

Pentaho kondigde op 9 juli 2008 de beschikbaarheid van zijn BI-suite aan voor de Apple iPhone (zie afbeelding 1) en was hiermee de eerste die dit voor elkaar kreeg. In de maanden daarna verscheen het ene na het andere persbericht van de grote BI-leveranciers dat ze dit trucje nu ook konden, maar er kan er maar één de eerste zijn. Ook dit is een belangrijk keerpunt in de ontwikkeling van OS BI: tot nu toe was men vooral bezig om functionaliteit toe te voegen die al jaren beschikbaar was in de commerciële pakketten. Nog steeds geldt dat zeer kritisch gekeken dient te worden naar de match tussen de organisatie-behoefte en het aanbod van de OS-leveranciers. Dit geldt in

steeds mindere mate voor data-integratieproducten, vooral wanneer het gaat om de modulaire opbouw van de verschillende pakketten en de wijze waarop met data van het web omgegaan wordt. Commerciële leveranciers missen op dit moment volledig de boot als het op dit laatste aankomt. Data van het web lezen in formats als JSON en Atom of ondersteuning voor REST en rss zult u wel vinden in Open Source, maar vooralsnog niet bij de gesloten concurrenten. Maar goed, ze kunnen natuurlijk altijd leentjebuur spelen bij hun open voorbeelden, iets wat op steeds grotere schaal gebeurt. Kijk bijvoorbeeld eens wat Information Builders heeft gedaan om statistische analysemogelijkheden toe te voegen aan haar BI-suite: adopteren van Project R (zie DB/M 5, 2007), of Actuate dat BIRT als reporting tool inzet (en hier ook flink aan bijdraagt). Ook Business Objects volgt in deze trend: Espertech vormt de basis van een (nog uit te brengen) real-time oplossing, MySQL wordt aan alle kanten ondersteund (zelfs als repository database) en Apache Tomcat is al jarenlang de standaard applicatie-server die bij het pakket wordt meegeleverd. Dus niet alleen vinden steeds meer echte innovaties in de Open Source wereld plaats, er wordt ook dankbaar gebruik van gemaakt door de gevestigde orde. De 'verovering' van de BI-markt door Open Source producten zal dan ook steeds meer sluipenderwijs en van 'onderaf' plaatsvinden. Omdat producten direct beschikbaar zijn zonder ingewikkelde inkoopprocedures is het inzetten van OS software een snelle en flexibele manier om nieuwe functionaliteiten aan te bieden. Omdat steeds meer via browsers en portals kan worden aangeboden merken gebruikers er ook niets van dat een product het OS-stempeltje met zich meedraagt.

MySQL

Ongeveer een jaar geleden werd ik gevraagd om tijdens Database Systems een lezing te geven met als titel 'Open Source Databases, een jaar later'. Toen bedankte ik nog voor de eer omdat er niet zoveel nieuws te melden was. Dat is nu wel anders en helemaal als we naar de ontwikkelingen bij MySQL kijken. De overname door SUN heeft blijkbaar nieuwe krachten losgemaakt bij de Zweden, want er wordt voortvarend gewerkt aan een nieuw modellerings-tool (MySQL Workbench) en aan versie 6.0 van de database met daarin de nieuwe Falcon transactie-engine. Het grootste nieuws op BI-gebied is echter *Maria*, (hopelijk) een codenaam voor de nieuwe database engine die onder andere speciaal wordt ontwikkeld voor datawarehouse-toepassingen. Maria is bedoeld als vervanger voor de aloude MyIsam-basis en moet een ACID compliant en MVCC (multi-version concurrency control) transactie-engine bieden die zowel transactie- als niet-transactiegerichte databases kan ondersteunen. Voor de BI-wereld is het toevoegen van eigenschappen om datawarehouse-projecten met MySQL uit te kunnen voeren echter het meest interessant. Hiervoor moet nog wel even geduld worden uitgeoefend: de datawarehouse-eigenschappen (bitmap & xdb indexen) zijn pas voorzien voor het laatste Maria increment, versie 4.0. De alpha release hiervan staat niet eerder

Afbeelding 1: Pentaho op de Apple iPhone.

dan voor Q3 2009 op de planning als onderdeel van MySQL 6.1. Voordat er een GA (general availability) release beschikbaar komt zijn we dan alweer een tijdje verder. Kijkend naar het moeizame tempo waarin versie 5.1 dit stadium aan het bereiken is zijn de 6.0 en 6.1 vooruitzichten niet echt rooskleurig te noemen.

Ander nieuws van het MySQL front is Drizzle, een nieuwe 'fork'. Een fork is een afsplitsing van de software waarbij op basis van bestaande code een nieuwe weg wordt ingeslagen, bijvoorbeeld omdat een groep ontwikkelaars ideeën heeft die afwijken van de hoofdroute die is vastgesteld door een project of bedrijf. In dit geval gaat het om een groep die terug wil naar de basis van MySQL, namelijk een kleine, simpele, makkelijk te gebruiken database voor cloud- en webgebaseerde toepassingen. Er wordt dan ook verder gebouwd op de versie 4 codebase, en niet op versie 5 met zijn transactie management, stored procedures, triggers, views enzovoort. Doel van Drizzle is het leveren van 'massive concurrency', dus het gelijktijdig ondersteunen van zeer grote groepen gebruikers, bijvoorbeeld binnen online community's. Bij 'online community's' moet u niet alleen denken aan Hyves of MySpace, maar ook aan games als World of Warcraft met zijn miljoenen spelers.

Datawarehousing

Tot 15 september 2008 was het opzetten van een datawarehouse voorbij de 100 GB grens met Open Source middelen een heikel karwei. Er zijn inderdaad cases van grote (tot wel 25 TB) MySQL implementaties, maar dat zijn transactiesystemen, geen data-

warehouses. Met 500 GB houdt dit echt wel op, en dat geldt ook voor alternatieven als PostgreSQL. Wat is er dan op 15 september gebeurd dat zo belangrijk is voor de Open Source BI-wereld? Heel simpel, het tot die tijd gesloten Infobright is uitgebracht onder GPL en de broncode kan sindsdien vrijelijk gedownload worden. Infobright bestaat pas een paar jaar en is één van de nieuwe spelers op het *column based* database-terrein, zoals ook bijvoorbeeld ParAccel, Vertica en EXASOL (zie DB/M 3, 2008). Infobright brengt zijn product nu uit in een Enterprise en een Community Edition (IEE en ICE), waarbij de laatste gratis kan worden gedownload en geïnstalleerd. Het is echter geen zuiver dual license model, waarbij de functionaliteit van beide producten in essentie gelijk is maar alleen in de te leveren ondersteuning en support verschilt. Infobright heeft ervoor gekozen om de Enterprise Edition wél te voorzien van temp tables, materialized views, snelle laadprocedures én DML-mogelijkheden (insert, update, delete), waar dat in de CE niet voorhanden is. En dat laatste onderdeeltje maakt het nu net nauwelijks geschikt als serieus datawarehouse-alternatief. Wat moet je met een datawarehouse dat je alleen kunt laden maar waar je geen data kunt verwijderen of updaten? Bovendien zijn de snelle loaders wél voor de EE, maar niet voor de CE beschikbaar.
De EE is ook nog eens niet echt voordelig: voor het goedkoopste (silver support) abonnement wordt 9.950,- dollar per Terabyte per jaar in rekening gebracht, wilt u 24/7 support met 1 uur responstijd dan betaalt u 15.950,- dollar. Goed, vergeleken met producten als Vertica en ParAccel die een aanschafprijs van 100.000,- dollar per Terabyte rekenen valt het nog wel mee, maar deze laatste bieden een veel betere schaalbaarheid door de MPP shared nothing architectuur die Infobright niet heeft.

Ander nieuws op het datawarehouse- en MySQL-vlak is de introductie van Kickfire, een heuse appliance speciaal voor MySQL-omgevingen, die niet zozeer gebruik maakt van slimme software maar een speciale SQL chip aan boord heeft voor het afhandelen van analytische query's. Dat dit resultaten oplevert mag blijken uit de in april en mei gepubliceerde TPC-H benchmark-resultaten waarmee Kickfire meteen niet alleen de prijs/prestatie-ranglijst in de 100 en 300 GB range aanvoert, maar ook de toppositie op de overall performance-lijst inneemt als naar ongeclusterde systemen wordt gekeken. Geen Open Source, maar wel een interessante hardware-matige aanvulling als u al MySQL gebruikt voor andere doeleinden en deze omgeving eenvoudig wilt uitbreiden met datawarehouse-voorzieningen.

Schreef ik trouwens dat ook voor PostgreSQL geldt dat schalen voorbij de 100 GB grens niet meevalt? Dat is niet helemaal waar. Ook binnen de PostgreSQL wereld zijn er namelijk positieve ontwikkelingen te melden, en wel vanuit de hoek van EnterpriseDB. Dit bedrijf schermde voorheen met zijn 'Open Source based' Oracle compliant oplossing, maar dat product was zélf weer geen Open Source. EnterpriseDB heeft echter ook het licht gezien en levert nu drie producten die alle met verschillende supportcontracten geleverd worden. Het vlaggenschip is de EnterpriseDB Postgres Plus Advanced Server die nog steeds gesloten is, met name vanwege de Oracle compatibiliteit van dit product, waardoor voor Oracle geschreven database-programmatuur nagenoeg ongewijzigd op een EnterpriseDB database kan draaien. EnterpriseDB levert ook de 'kale' PostgreSQL database met eventueel betaalde ondersteuning, maar het meest interessante product is PostgreSQL Plus, de Open Source versie van de Advanced Server waaruit een aantal onderdelen is gehaald. Wat echter behouden is gebleven is GridSQL, en dat is voor DWH-doeleinden ook meteen het meest spectaculaire nieuws (zie het schema van GridSQL op pagina 15 in dit blad). Hiermee kan namelijk een MPP shared nothing architectuur worden opgezet, waarbij de GridSQL software de centrale coördinatie verzorgt voor het partitioneren van de databases en het parallelliseren van query's. Als u dit dan ook nog op een slimme en dynamische manier opzet binnen Amazons EC2 platform, kunt u voor enkele euro's per dag een multi-Terabyte datawarehouse optuigen dat ook nog een zeer goede performance levert. Een soort Greenplum (dat op dezelfde manier werkt) maar dan met gratis software en zonder de hoge investeringen in (SUN) hardware.

De grote spelers

De namen Jaspersoft en Pentaho mogen inmiddels als bekend zijnd verondersteld worden. Het zijn de twee grote namen die een complete Open Source BI-oplossing leveren, soms zelfs completer dan veel commerciële leveranciers. Beide bedrijven zijn gebouwd rondom een verzameling projecten en bieden een platform waarbinnen deze verschillende producten kunnen draaien, maar daar houdt de vergelijking wel een beetje op. Jaspersoft heeft het afgelopen jaar versie 3.0 van zijn serverproduct uitgebracht met als belangrijkste highlights de mogelijkheid om op een visuele 'drag and drop' manier dashboards te bouwen, en de metadatalaag waarmee ad hoc reporting (eindelijk) ook binnen de Jaspersoft omgeving mogelijk is. Er is echter een kleine 'maar': het zijn geen Open Source producten! Jasper is een weg ingeslagen waarbij de 'community' edition niet meer dient als innovatieplatform voor de Enterprise Editions, maar nog slechts een uitgeklede variant is van deze laatste. Pentaho is juist bezig met een omgekeerde beweging: de nieuwe versie 2 van het (tot nu toe) commercieel gelicenseerde BI-platform wordt onder de GPL uitgebracht. Het zal dan ook interessant zijn om te zien welke strategie op termijn het beste gaat uitwerken. Overigens heeft Pentaho nog steeds een Enterprise Edition waarin een aantal closed source componenten zit, zoals bijvoorbeeld de single sign-on mogelijkheid.

Het grote Pentaho nieuws is de beschikbaarheid van versie 2.0 van zijn platform, wat een aanzienlijke verbetering inhoudt, zowel voor (eind)gebruikers als voor beheerders. Alles zit een stuk logischer en eenduidiger in elkaar, waardoor de suite beter gepositioneerd is om de concurrentie met de gevestigde orde aan

te gaan. Dat dit ook al in Europa lukt blijkt uit een (niet bij naam genoemde) retailer waarmee een contract boven de 1 miljoen dollar is afgesloten. Ook op andere plekken kunt u Pentaho tegenkomen. De Report Builder die als add-on voor OpenOffice beschikbaar is komt van Pentaho en er zijn verschillende OEM-overeenkomsten afgesloten. De deal met ERP-leverancier OpenBravo is hiervan wel de meest in het oog springende.

Het overige productnieuws dat beide spelers het afgelopen jaar over de community's hebben uitgestort is aanzienlijk; web 2.0 is uiteraard een 'hot item', evenals integratie met verschillende andere producten. Denk hierbij aan Excel, maar ook aan diverse portal-producten. Beide platformen beschikken inmiddels over portlets die aan de JSR-168 portlet specification voldoen, waardoor componenten van zowel Jaspersoft als Pentaho gebruikt kunnen worden in elk portal dat met deze portlets overweg kan, en dat zijn ze nagenoeg allemaal. Verder zijn er natuurlijk nieuwe versies van de ETL-producten, Flash-generatoren, Excel-integratie, extra reporting-mogelijkheden, een nieuwe versie van OLAP engine Mondrian (die zowel door Pentaho als Jaspersoft wordt gebruikt, terwijl Mondrian toch echt deel uitmaakt van de Pentaho stack, over 'open' gesproken!) en extra zaken als row level security. Helaas is de ruimte te beperkt om overal diep op in te gaan, maar bent u geïnteresseerd kijk dan vooral op de respectievelijke websites, onderaan vermeld bij de referenties.

Poppenspel

Dat het de Open Source leveranciers menens is bij het veroveren van de markt, blijkt ook uit de personeelswisselingen en uitbreidingen van het afgelopen jaar. Bij de meeste grote spelers halen de oprichters een ervaren CEO binnen om verdere groei mogelijk te maken, terwijl ze zelf de rol van VP business of product development op zich nemen. Zo heeft Andy Astor van EnterpriseDB Ed Boyajian weggekaapt bij Red Hat, en heeft Jaspersoft's Paul Doscher plaatsgemaakt voor Brian Gentile die afkomstig is van Informatica. Ook bij Pentaho zit men niet stil: in juli werd de versterking van de board of directors met Zack Urlocker aangekondigd, en in oktober werd Lars Nordwall gestrikt als manager business development. Zack is VP Products in SUN's database group (en had dezelfde functie al bij MySQL), terwijl Lars de Open Source Salesforce.com concurrent SugarCRM van niets naar 3000 klanten wereldwijd hielp groeien. Tel hierbij op dat het Pentaho management team al bestond uit ervaren BI-spelers, waaronder de voormalige Business Objects marketing VP Lance Walter.
Er is overigens niet alleen maar goed nieuws over personeelswisselingen. Uit het MySQL kamp komen wat andere berichten. Verschillende analisten fronsten al hun wenkbrauwen toen Monty Widenius Sun verliet, maar toen ook mede-oprichter David Axmark het voor gezien hield werd even gevreesd voor een complete uittocht van de MySQL top. Maar goed, zelfs als iedereen Sun verlaat en zelfs als Sun MySQL bij het grof vuil zet zal MySQL als product wel overleven, daar zorgt uiteindelijk de community wel voor. Kijk bijvoorbeeld naar PostgreSQL, daar zit geen enkel commercieel bedrijf achter en is toch één van de meest geavanceerde en robuuste databases ter wereld.

Conclusie

Op basis van bovenstaand relaas kunt u zelf wel opmaken dat de ontwikkelingen in de Open Source wereld net zo snel gaan als bij de proprietary concurrenten, zo niet sneller. Met nieuwe Open Source producten waarmee ook grotere datawarehouses zijn in te richten komt het optuigen van een volledige OS BI stack voor steeds meer organisaties binnen bereik. Wat ook helpt is dat het niet alleen de broncode is die open is. De openheid zit met name ook in de vergaande interoperabiliteit van de verschillende componenten waardoor u nooit afhankelijk bent van één leverancier voor het complete platform. Dit blijkt ook uit onderzoeken waarbij men in eerste instantie ging voor Open Source als middel om kosten te besparen, maar als belangrijkste eindresultaat de flexibiliteit van de oplossing roemde.
Voorspellen blijft een uitdaging, maar vooruit, een poging kan natuurlijk altijd gewaagd worden. Ten eerste is er de economische tegenwind die wel eens voor een onverwachte duw in de rug van Open Source producten zou kunnen zorgen. Ten tweede zal ook de consolidatiegolf nog wel even doorgaan, en dan kijk ik met name naar SUN. Na de overname van MySQL en de ontwikkeling van GlassFish met daarbinnen het datamanagement-platform Mural zou een overname van Infobright en/of Kickfire een logische vervolgstap kunnen zijn. Overname van Pentaho is ook een interessante optie, waarmee SUN in één klap een compleet Open Source alternatief vormt voor de gesloten MISO clubs. De ontbrekende ERP-suite kan vervolgens bij OpenBravo of Compiere worden ingekocht. Goed, ik heb geen kristallen bol dus het blijft koffiedik kijken, maar wie weet.

Referenties

OS Venture Capital: http://blogs.the451group.com/opensource/2008/10/16/vc-funding-for-open-source-down-12-in-q3/
MySQL Maria: http://en.oreilly.com/oscon2008/public/schedule/detail/2619
Drizzle: https://launchpad.net/drizzle
EnterpriseDB: www.enterprisedb.com
Kickfire: www.kickfire.com
Pentaho: www.pentaho.com
Jaspersoft: www.jaspersoft.com

Jos van Dongen

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

In het artikel over Open Source Data Quality in DB/M 7, 2008 ontbrak de vermelding van twee tools: Power*MatchMaker, een Open Source data cleansing tool, en Power*Architect, OS datamodellering en cube design. Beide producten zijn als commercieel pakket begonnen, maar sinds een jaar Open Source. Kijk op www.sqlpower.ca voor meer informatie.