Maak gebruik van big data om er waarde uit te halen

Data-integratie zit Teradata in de genen

Frans Godden

Dat de hoeveelheid data in de wereld schrikbarend groeit, weet u intussen al wel. Een nieuw probleem dat daaruit voortvloeit is dat die gegevens verspreid zitten over meerdere databases in allerlei vormen. Data-integratie staat dan ook hoog op het prioriteitenlijstje van vele IT-managers om uit die stortvloed aan data nog zinnige (en vooral betrouwbare) informatie te kunnen puren. "Maar daar zorgen wij voor", zegt Scott Gnau, president van Teradata Labs.

In elk bedrijf worden dagelijks ettelijke beslissingen genomen door meerdere mensen op verschillende niveaus, maar ongeacht rang of rol hebben die medewerkers allemaal één ding gemeen: voor hun beslissingen zijn ze afhankelijk van de informatie die ze aangereikt krijgen. Hoe beter en hoe consistenter die informatie is, hoe efficiënter ook de actie zal zijn die eruit voortvloeit. In de complexe wereld van vandaag betekent dit vaak informatie vergaren uit verschillende onderdelen van de business; verkoop, marketing, voorraadbeheer, financiën. Niet zelden zit hun informatie opgeslagen in van elkaar losstaande databases die afwijkende structuren gebruiken. Het is dus zaak om al die gegevens samen te voegen en te integreren tot één consistent geheel waarop een crossfunctionele analyse losgelaten kan worden. "Pure noodzaak", zegt Randy Lea, vice president Global Products & Services Marketing bij Teradata. "Als je niet met geïntegreerde data kan werken, moet je beslissingen nemen op basis van informatie uit slechts een gedeelte van alle data die in een organisatie aanwezig zijn en kan je dus ook maar op een beperkt aantal vragen een correct antwoord krijgen of geven. Maar hoe meer informatie je uit heel het bedrijf kan verzamelen, hoe beter en relevanter je beslissingen zullen worden – en hoe meer impact ze zullen hebben op de resultaten van je bedrijf".

Integratie is grootste uitdaging

Natuurlijk wordt het er niet eenvoudiger op met 'big data', het nieuwe buzzword in de IT-wereld dat verwijst naar de massa gegevens die nu gegenereerd worden, niet alleen door computers maar ook door sensoren, digitale camera's, GPS-toestellen enzovoort. Al die data komen terecht in databases die, zoals gezegd, niet zelden als silo's los van elkaar staan. Het is dan ook niet verwonderlijk dat een onderzoek van Unisphere Research bij ruim 400 IT-managers dit voorjaar uitwees dat het integreren van data die tussen verschillende databases uitgewisseld worden, een van de grootste uitdagingen is in een multi-database managementsysteem. Heel vaak ontwikkelen bedrijven zelf hiervoor oplossingen, die echter even vaak vroeg of laat tekort schieten. Immers, er doet zich nooit een status quo voor, er worden constant nieuwe bronnen en systemen aan het geheel toegevoegd en de complexiteit neemt alleen maar toe.

"Op die behoeften speelt Teradata nu perfect in met de introductie van het nieuwe Teradata Aster MapReduce Platform dat de opstap naar big data analytics aanzienlijk zal versnellen", legt Scott Gnau uit. "Tot nog toe hadden de kostprijs voor het ontginnen van grote volumes multi-structured data en het gebrek aan mensen met de nodige gespecialiseerde analytische kennis een stevige rem gezet op het gebruik van big data analytics. Het nieuwe platform verenigt echter MapReduce, de taal voor big data analytics, met SQL, de taal voor business analytics. Naast

Scott Gnau, President van Teradata Labs.

de Aster Database 5.0 zit er ook een nieuwe Aster MapReduce Appliance in – wat betekent dat de implementatie niet enkel meer in een SaaS- en cloudvorm hoeft te gebeuren – plus de Teradata Aster Adaptor voor supersnelle datatransfer tussen Teradata en Aster Data systemen. Wat vroeger een open source toepassing was bij Aster Data is nu een volwassen commercieel product geworden dat we tot een complete analytics oplossing gebundeld hebben".

Voor iedereen toegankelijk

In maart van dit jaar nam Teradata Aster Data Systems over voor 263 miljoen dollar, precies om te kunnen beantwoorden aan de groeiende vraag naar systemen voor het beheer van big data. Analist Colin White van BI Research spreekt van een uitgekiende zet omdat Teradata dankzij Aster analytics grote en complexe datasets meer toegankelijk maakt voor alle gebruikers via standaard SQL. Scott Gnau sluit zich daar bij aan: "Iedereen van de businesskant kan nu multi-structured data zien, onderzoeken en begrijpen, big data analytics is niet meer het exclusieve speelterrein van enkele data scientists of MapReduce specialisten. Steeds meer eindgebruikers krijgen daardoor inzage in voorheen onzichtbare correlaties tussen de gegevens zodat ze efficiënter kunnen werken en van data een toegevoegde waarde in plaats van een kostenpost kunnen maken".
Hij wijst daarbij op het belang van een totale kijk op de business – je moet over gegevens uit alle facetten van de business beschikken en ze relationeel integreren om zo alles vanuit een analytisch perspectief te kunnen begrijpen. "Als ik mijn marketingprogramma aanpas, wat is dan het effect op mijn balans? Of als ik mijn kapitaalstructuur verander, welke impact zal dat dan hebben op mijn klanten? In staat zijn de impact van een beslissing te zien op de verschillende afdelingen van een bedrijf is vaak cruciaal, en daar ligt precies de grote toegevoegde waarde van data-integratie", aldus Gnau.
Mike Koehler, President en CEO van Teradata, gaat zelfs nog een stapje verder. "Bedrijven hebben jarenlang al hun informatie opgeslagen in database-silo's, zonder te beseffen dat ze in feite bovenop een goudmijn zaten. Door hun data nu te gaan integreren, zullen ze, met behulp van de juiste data mining tools en data scientists, ontdekken dat er plots 'goudklompjes' in verscholen zitten die ze voordien nooit gezien hadden!"

Datakwaliteit

Natuurlijk verloopt zo'n data-integratie niet altijd van een leien dakje, een van de complicaties die vaak opduiken bij het integreren van gegevens is dat de kwaliteit van de data uit de onderscheiden bronnen erg kan verschillen. Vooral als het om complexe data gaat met gestructureerde en ongestructureerde gegevens. "Al wil ik hier toch terloops even iets opmerken", lacht Scott Gnau: "Als purist durf ik te stellen dat er geen ongestructureerde gegevens bestaan, alle gegevens hebben een structuur. Maar alle gekheid op een stokje: datakwaliteit kan een serieus struikelblok vormen bij het integreren van data. Bijna

Mike Koehler, President en CEO van Teradata.

zonder uitzondering komen er immers bij zo'n integratie problemen met de kwaliteit van de gegevens aan het licht. Bij het implementeren van een datawarehouse stellen heel wat van onze klanten vast dat er fouten in hun brongegevens zitten, al jaren, zonder dat iemand het ooit gemerkt heeft. Nu, precies omdat we die bronsystemen onaangeroerd laten en de gegevens enkel in ons datawarehouse gaan integreren, kunnen die klanten de fouten in hun eigen systemen nog corrigeren vooraleer ze geïntegreerd worden".

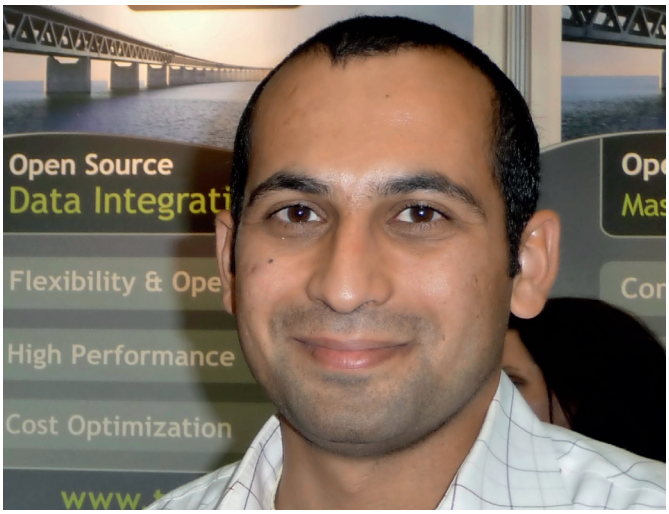
Overigens, zo voegt Gnau er nog aan toe, werkt Teradata voor het monitoren van die datakwaliteit ook samen met meerdere partners. Een daarvan, die onlangs zijn intrede deed in het Magic Quadrant for Data Integration Tools van Gartner, is Talend, een leverancier van open source software die in 2005 werd opgericht in Suresnes, Frankrijk, waar nog altijd het hoofdkwartier gevestigd is, samen met een tweede in Los Altos, Californië. Vandaag telt het bedrijf ruim 400 werknemers verspreid over een dozijn vestigingen over heel de wereld met grote klanten als Deutsche Post, Allianz, eBay en Sony Online Entertainment. "Maar het open source model maakt onze producten natuurlijk ook aantrekkelijk voor MKB's", zegt Parham Prvizi, Technical Principal bij Talend US. "De instapdrempel is immers erg laag, bedrijven kunnen onze Talend Open Profiler gratis downloaden, een open source data profiling tool die de inhoud, structuur en kwaliteit van complexe datastructuren ontleedt en vervolgens met behulp van Talend Data Quality alles opschoont. Met de overname vorig jaar van Sopera, een van de marktleiders in open source applicatie-integratie, kunnen wij nu ook een totaaloplossing voor open source middleware aanbieden waarin data-integratie, datakwaliteitbewaking en MDM, Master Data Management, samengebracht zijn. En daar is echt wel behoefte aan want haast elk IT-departement moet vandaag waken over de consistentie van zijn data en processen met behulp van modeling tools, workflow management en storage – zeg maar de basis van alle data governance".

Columnar database

In een recent onderzoek van het McKinsey Global Institute (MGI) wordt nogmaals beklemtoond dat bedrijven geen schrik moeten hebben van big data maar integendeel van de datastroom gebruik moeten maken om er waarde uit te halen. Zo kunnen organisaties die veel transactionele data genereren, er meer precieze en gedetailleerde performance informatie uit puren over alles en nog wat, van productvoorraden tot ziekteverzuim, om op die manier schommelingen op te sporen en de operaties te stroomlijnen. Grote bedrijven gebruiken ook die datamassa met analytics om gecontroleerde experimenten uit te voeren teneinde hun besluitvorming te verbeteren. "En uiteraard kan je met behulp van big data en analytics ook je klanten steeds nauwkeuriger segmenteren om hen meer persoonlijke producten en diensten aan te bieden", zegt Scott Gnau. "Een van de nieuwe tools die hen daarbij kunnen helpen, is Teradata Columnar, een uitbreiding voor de nieuwste versie van onze database, Teradata 14, die in december beschikbaar komt. In tegenstelling tot relationele database managementsystemen die hun data enkel in rijen opslaan, doet een columnar database dat in kolommen. Elk van die methoden heeft unieke voordelen afhankelijk van de applicatie en het soort van data, maar met onze nieuwe columnar-voorziening kunnen klanten voortaan beide vormen combineren voor hogere prestaties en flexibiliteit. Bovendien kan Teradata Columnar automatisch de meest geschikte compressiemethode kiezen en het compressiemechanisme dynamisch aanpassen aan de evoluerende data".
Niet dat columnar databases zo nieuw zijn, Sybase, IBM en Oracle, om maar enkele te noemen, hebben al langer columnar storage engines in hun producten zitten, maar het nieuwe met Teradata Columnar is dat nu voor het eerst columnar en row-based tabellen gecombineerd worden – en dat werd door de analisten die op de jongste Teradata Partners Conference in San Diego aanwezig waren als erg vernieuwend bestempeld.

En ook nieuwe hardware

Voor wie behoefte heeft aan nog meer performance voor de verwerking van zijn big data analytics introduceert Teradata begin volgend jaar de vijfde generatie van zijn Teradata Data Warehouse Appliance, een volledig geïntegreerd systeem dat tweemaal krachtiger is en drie keer meer datacapaciteit heeft dan zijn voorganger. Daarmee kunnen klanten een geïntegreerd analytics platform configureren dat van 2 tot 315 Terabytes uncompressed user data per cabinet aankan en dat data kan scannen tegen meer dan 38 Gigabytes per seconde. Ook dit systeem is uitgerust met een nieuwe block-level compressietechnologie die volledig automatisch in actie komt zonder iets aan de prestaties van het geheel af te doen, mede dankzij speciaal daartoe ingezette coprocessoren. En er zit ook een 'green' aspect aan, want de Teradata Data Warehouse Appliance 2690 verbruikt 60 procent minder energie en neemt 50 procent minder ruimte in dan zijn voorganger. Toch niet te verwaarlozen voor datacenters waar elke vierkante meter en het stroomverbruik cruciaal zijn.



Parham Prvizi, Technical Principal bij Talend US.

Twee nieuwe softwareproducten die in de rand van die hardware-aankondiging gelanceerd werden, zijn Teradata Unity en Data Mover. Het eerste, dat al een jaar of twee in ontwikkeling was, zorgt voor query routing en database-synchronisatie en updating over verschillende Teradata systemen. Scott Gnau hierover: "De bedoeling is vooral het beheer van het Teradata Analytical Ecosystem te vereenvoudigen door data en databases gesynchroniseerd te houden over alle systemen, gebruikers en query's op een intelligente manier te routen en zelfs zonder dat de gebruiker het merkt zijn query naar een ander systeem te sturen mocht het eerste offline gaan – high availability dus".
Het andere product, Data Mover, is een tool die bedoeld is om data en objecten zoals tabellen en statistieken vlotter van het ene Teradata systeem naar het andere over te zetten.

Veel migraties

Teradata is in elk geval in goeden doen, vorig jaar zag het zijn omzet met 13 procent toenemen en voor dit jaar wordt een stijging met 18 tot 20 procent verwacht. Volgens CEO Mike Koehler heeft het bedrijf de jongste tien jaar nooit zoveel nieuwe klanten binnengehaald als in 2011. Analisten schrijven dit vooral toe aan het feit dat veel organisaties naar de analytics-platformen van Teradata overstappen omdat ze problemen hebben om hun bestaande databases aan te passen voor BI- en analytics-opdrachten. Teradata maakt het die organisaties gemakkelijk met zijn Teradata Migration Accelerator die automatisch datastructuren, indexen, views, tabellen, code en bedrijfsgegevens overzet. Naar eigen zeggen zouden daarna de daling van het onderhoud en de vereenvoudigde architectuur de taak van een databasebeheerder met 70 procent verlichten en zou het aantal lijnen code met meer dan 50 procent dalen. Load processing time zou met 50 procent verminderen, terwijl ad hoc query performance tien keer beter wordt. Allemaal eigenschappen die beloofd zijn voor de komende versie Teradata Database 14.

Frans Godden is freelance journalist.