VAN DER LANS

Nieuwe loot aan de analytics-boom

De gebruikers van Business Intelligence systemen kunnen momenteel kiezen uit een breed scala aan producten waarmee ze rapporten kunnen maken en analyses kunnen uitvoeren. Deze producten variëren van rechttoe rechtaan rapportageproducten via interactieve online analytical processing tools tot geavanceerde statistische en datamining producten. De bedoeling van al deze producten is dat ze het beslissingsproces van de gebruikers verbeteren. Ze helpen door gegevens onder andere te filteren, te sommeren, te groeperen en door te voorspellen en de resultaten grafisch te presenteren.

Maar er zijn zaken waartoe deze producten niet in staat zijn en dat is, onder andere, het analyseren van gegevens wanneer deze als een graaf of netwerk gestructureerd zijn en wanneer de analyse vereist dat die netwerkstructuur bewandeld moet worden. Stel eens voor dat een manager van een sociaal netwerk website, zoals Facebook of LinkedIn, wil weten wie de centrale leden van het totale netwerk zijn, waarbij de term centraal lid gedefinieerd is als een lid dat de meeste korte paden heeft naar de andere leden. Dit probleem kan niet opgelost worden door simpelweg gegevens bij elkaar op te tellen noch heeft het iets te doen met het voorspellen met behulp van statistische modellen. Nee, wat hier nodig is, is een product dat van lid naar lid door het netwerk kan wandelen. Maar dit is een eigenschap die de meeste bekende analytische en rapportageproducten niet ondersteunen.

We geven nog een ander voorbeeld. Met veel rapportageproducten zal een luchtvaartmaatschappij kunnen bepalen hoeveel vluchten per dag vanuit een specifiek vliegveld vertrekken. En als per vlucht de zogenaamde load factor (percentage stoelen verkocht) bekend is, kunnen ze ongetwijfeld de gemiddelde load factor voor vluchten van Amsterdam naar Londen berekenen. Beschikken ze over de geschikte statistische producten, dan kunnen ze zelfs laten voorspellen wat de load factor voor de komende maand zal zijn. Ze kunnen ook dashboards ontwikkelen die live de gemiddelde load factor voor alle vluchten tonen. Maar wat al deze producten niet kunnen is bepalen wat de twee goedkoopste of de vier kortste vluchten van Amsterdam naar New York zijn. En als we weer het voorbeeld van een sociaal netwerk nemen, deze producten kunnen niet bepalen welke andere leden een specifiek lid waarschijnlijk wel kent, maar nog niet mee verbonden is. Het zal ook lastig zijn voor een telefoonmaatschappij om te bepalen welke klanten mogelijkerwijs andere klanten beïnvloeden om bij de huidige provider te blijven of over te stappen.

De bovengenoemde problemen behoren tot het domein van *graph analytics*, ofwel het analyseren van grafen (netwerken). Graph analytics is een speciale vorm van analytics die al lang bestaat. In feite gaat de geschiedenis van graph analytics en de onderliggende grafentheorie terug tot aan de eerste helft van de achttiende eeuw. Tegenwoordig bestaan er krachtige producten en databaseservers speciaal ontwikkeld voor graph analytics. Het speciale aan deze producten is dat ze grafen bestaande uit honderden miljoenen nodes kunnen verwerken en ze snel kunnen analyseren. Ze ondersteunen de algoritmes om bepaalde karakteristieke graafproblemen op te lossen.

De BI-wereld is helaas nog niet zo bekend met het analyseren van grafen. Niet dat het onderwerp nieuw is, maar het wordt nog maar zelden toegepast in Business Intelligence systemen. En dat is jammer, want graph analytics heeft veel te bieden en de producten en databaseservers zijn er klaar voor. Graph analytics kan ook in veel omgevingen toegepast worden. Overheidsinstanties, financiële instellingen, distributie- en transportbedrijven, retailers, telefoonmaatschappijen en eigenaren van websites kunnen allemaal zeer nuttig van deze mogelijkheid gebruik maken.

Tamelijk recent is een nieuwe generatie databaseservers geïntroduceerd waar naar gerefereerd wordt met de intrigerende term NoSQL databaseservers. Let wel, dit is niet een homogene groep van producten, maar een groep databaseservers met zeer uiteenlopende mogelijkheden en toepassingsgebieden. Wat ze gemeen hebben is dat ze SQL niet als de primaire databasetaal zien. Sommige ondersteunen SQL geheel niet en andere ondersteunen het slechts als secundaire taal en dan soms slechts een subset van SQL. Enkele van deze NoSQL databaseservers kunnen geclassificeerd worden als graph databaseservers: producten die speciaal voor het ondersteunen van graph analytics ontwikkeld zijn. Voorbeelden hiervan zijn Objectivity's InfiniteGraph, AllegroGraph RDFStore, Neo4j en vertexdb.

Samenvattend; de huidige producten die ontwikkeld zijn voor graph analytics zijn klaar voor het grote werk. Vooral de op databaseservers gebaseerde producten zijn in staat om zeer grote grafen bestaande uit miljoenen nodes te analyseren. Deze vorm van analytics verdient daarom meer aandacht van alle BI-specialisten. De grote uitdaging is om te bepalen waar het binnen een organisatie nuttig ingezet kan worden. Waar kan graph analytics het huidige palet van BI-producten verrijken?

Rick van der Lans is zelfstandig IT-consultant.