VAN DER LANS

# Zijn datawarehouses werkelijk zo groot?

Hoe vaak lezen we wel niet dat datawarehouses extreem groot zijn en dat ze de komende jaren zelfs exponentieel zullen groeien? Bijvoorbeeld, al in 2007 voorspelde Gartner dat 50 procent van de datawarehouses groter zou zijn dan vijftig Terabytes. Als voorbeelden worden meestal organisaties genoemd als Walmart, eBay en Yahoo. Dit zijn uiteraard organisaties die gigantische hoeveelheden gegevens verzamelen en dus ook zeer grote datawarehouses bezitten. Maar zijn datawarehouses echt zo groot? Of zijn de cijfers enigszins misleidend?

Wat ik zelf erg interessant vond was een onderzoeksresultaat van de Engelse analist Nigel Pendse dat hij tijdens een Array evenement presenteerde. Zijn organisatie voert elk jaar een onderzoek uit naar het gebruik van datawarehouses; de BI Survey. Daarvoor stuurt zijn organisatie honderden enquêteformulieren op. Voor zijn presentatie had hij onderzocht hoeveel gegevens een BI-applicatie nodig heeft om te functioneren.

De conclusie was dat dit rond de vijf Gigabytes ligt en dat die grootte door de jaren heen niet echt veranderd is. Om precies te zijn, honderden organisaties gaven aan wat hun BI-applicaties nodig hadden en de mediaan lag rond de vijf Gigabytes aan gegevens. Uiteraard waren er organisaties met BI-applicaties die heel veel meer gegevens nodig hadden, maar er waren er ook die minder nodig hadden.

Dit getal van vijf Gigabytes staat uiteraard in schril contrast met een getal als vijftig Terabytes. Ze horen eigenlijk niet bij elkaar. Want stel dat een organisatie een datawarehouse van vijftig Terabytes heeft opgebouwd, dan zijn er dus ongeveer tienduizend BI-applicaties. Let wel, al die applicaties moeten niet overlappende gegevensbehoeftes hebben, want anders zijn er nog veel meer nodig. De realiteit is uiteraard dat BI-applicaties juist wel dezelfde gegevens gebruiken. Hoeveel applicaties hebben we dan wel niet nodig om een datawarehouse van vijftig Terabytes te vullen?

Het probleem bij deze vergelijking is dat als er gesproken wordt over de grootte van een datawarehouse men over het algemeen de bruto grootte bedoelt, terwijl Nigel Pendse het over de netto grootte heeft. De bruto grootte van een datawarehouse is die hoeveelheid bytes die door de gehele datawarehouse-omgeving wordt ingenomen, dus inclusief het datawarehouse zelf, de datamarts, de kubussen, het ODS, enzovoort. Het omvat als het ware het totaal aan alle opgeslagen gegevens. Daarentegen is de netto grootte de hoeveelheid bytes die de gegevens zouden innemen als we het in een sequentieel bestand zouden opslaan zonder enige vorm van redundantie.

Het verschil tussen deze twee grootheden kan aanzienlijk zijn. De reden is de forse hoeveelheid redundante gegevens die opgeslagen wordt. Bijvoorbeeld, in veel gevallen zijn alle gegevens in de datamarts volledig redundant ten opzichte van wat er in het centrale datawarehouse ligt opgeslagen (mits deze aanwezig is); dimensies worden soms in verschillende datamarts herhaald; vele kubussen zijn ook vaak volledig redundant; tussen een eventueel ODS en een datawarehouse bestaat veel overlap; en hetzelfde geldt voor datamarts en een eventuele persistent staging area. Daarnaast bevat een datawarehouse zelf ook veel redundantie. De gegevens in de indexen, die in de materialized views, en de kolommen met geaggregeerde gegevens: ze zijn allemaal redundant.

Ik zou wel eens willen weten wat nu gemiddeld het verschil is tussen de netto grootte en de bruto grootte van datawarehouses. Het zou mij niet verbazen als blijkt dat bij veel organisaties de netto grootte slechts 20 procent is van de bruto grootte. En dit is een hele conservatieve voorspelling, omdat we weten dat als we gegevens in een klassieke SQL database server laden, we al met een explosiefactor rekening moeten houden van tenminste drie. Is dat belangrijk? Ik denk het wel. Het opslaan van redundante gegevens zal ongetwijfeld nuttig zijn om de performance van lastige query's te versnellen, maar hebben we enig idee wat de kosten van al die redundante gegevens zijn? Hoe duur is eigenlijk een datamart? Wat kost het om die redundante gegevens bij te werken? In hoeverre vertragen ze het laadproces? Verminderen ze de flexibiliteit van een omgeving?

Zijn onze datawarehouses dus nu echt zo groot? Uiteraard zijn veel datawarehouses niet klein en uiteraard groeien ze hard, en uiteraard bezitten eBay en WalMart fenomenaal grote databases, maar het leeuwendeel van al die gegevens is gewoonweg redundant. Dus de cijfers zijn inderdaad wat misleidend. Onze gehele datawarehouse-omgeving, inclusief alle datamarts en kubussen, is inderdaad zo groot, maar de netto hoeveelheid gegevens is een stuk minder groot. Het wordt tijd dat we gaan bestuderen of al die redundantie nog wel nodig is en of onze architecturen misschien gesimplificeerd kunnen worden.

Opmerking: een werkelijke definitie van wat precies netto grootte is moet wel wat preciezer zijn dan wat ik hier aangeef. Hierin moeten aspecten zoals onder andere compressie en het normaliseren van gegevens meegenomen worden.

**Rick van der Lans** is zelfstandig IT-consultant.