

Klaas Brant over IBM's DB2 Universal Database

Over airbags en query re-write

Is er nog iets nieuws uit te vinden in databaseland? Alles wat we nodig hebben, is reeds uitgevonden. Als dat waar was, konden auto-ontwerpers ook wel stoppen met hun research. Op die manier zou de airbag er nooit gekomen zijn.

De meeste programmeurs gebruiken een SQL waarvan je soms de tranen in de broek springen. Onder het motto 'onbekend maakt onbemind' worden constructies als WHERE EXISTS vermeden, om nog maar te zwijgen over OUTERJOIN's. Natuurlijk, uitzonderingen bevestigen die regel, en ook de BI-leveranciers benutten alle mogelijke SQL-constructies. Echter, hun tools hebben een SQL-generator die weliswaar complexe SQL-constructies kan genereren, maar vaak tegenstrijdigheden en redundante constructies oplevert. Schoolvoorbeeld is dat deze generatoren vaak een WHERE NOT NULL-predikaat genereren voor een kolom die per definitie geen NULL's kan bevatten. Dom? Niet echt. Probeer maar eens met alle facetten van database- en objectdefinitie rekening te houden. En dan moet je ook nog eens meerdere databases ondersteunen. Nee, we zullen ermee moeten leren leven dat er veel ronduit slechte SQL in de wereld is. Voor de databaseleveranciers ligt er dan ook de uitdaging hiervan het beste te maken.

In de beginjaren van de relationele databases nam de database SQL aan, en die werd één-op-één vertaald in een access path om de data op te halen en het resultaat te creëren. Overigens doen sommige databases dit nog steeds en is de constructie van SQL maatgevend voor de snelheid waarmee het resultaat terugkomt. Andere databases gaan wel een stapje verder. Zo neemt DB2 de SQL aan om die vervolgens af te breken in een parsing-tree. Daarna gaat de optimizer aan de slag. Met behulp van complexe algoritmen kijkt hij of de SQL anders geformuleerd kan worden.

Kan dat zo maar? Natuurlijk. In deze stap, *query re-write*, wordt gebruik gemaakt van de verzamelingenleer uit de wiskunde. Het is niet voor niets een relationele database. Bij deze leer hoort ook de booleaanse algebra, waarin absorptie- en transitie regels bestaan. Dit terugvertalend naar SQL: door de regels toe te passen kunnen we subquery's omschrijven naar joins en onzin uit SQL bijvoorbij wegstrepen of juist de SQL uitbreiden. Want op die manier

kan de database meer access paths onderzoeken en kiezen voor het beste, goedkoopste en snelste access path.

De aangeboden SQL hoeft niet het beste te zijn om toch het beste resultaat te krijgen. IBM doet hiernaar veel research, en ik mag wel zeggen dat de query re-write van DB2 waarschijnlijk de beste ter wereld is. Waarschijnlijk, omdat onafhankelijke partijen de verschillen tussen databaseleveranciers op dit punt tot nog toe

weinig hebben onderzocht. Vreemd, en ook jammer, want query re-write is in mijn ogen vele malen belangrijker dan een TPC-benchmark.

Research naar query re-write is belangrijk, en van de voordelen kunnen alle applicaties onmiddellijk gebruik maken. Maar soms wordt ook veel research gedaan naar nieuwe concepten. Zo spendeert IBM de nodige tijd aan research van indexen en data-opslag. In DB2 Versie 8 (Unix, Linux en Windows) komt men met een nieuw concept dat Multi Dimension Clustering (MDC) heet. Met MDC worden de data zo opgeslagen, dat ze altijd 100% geclusterd zullen blijven. Maar dat kan ook volgens meerdere indexen. Dit dankzij een bijzonder knappe manier van data-opslag met blocks, cells en slices. Daarbovenop komt een nieuw type index (*block index*), die deze MDC-opslag begrijpt en op een slimme manier kan gebruiken. 'Multi dimension' doet sterk denken aan datawarehousing en dat maakt de voordelen natuurlijk snel duidelijk. Maar hebt u in een *data-ontwerp* nooit voor de moeilijke keuze gestaan welke index de clusteringindex zal worden? Deze keuze is dus binnenkort niet meer nodig.

In de auto-industrie en elders worden nog dagelijks veel dingen uitgevonden die onmiddellijk bruikbaar zijn. Soms zichtbaar als een airbag, soms onzichtbaar als een ABS. Zo ook in de database-industrie: soms zichtbaar, zoals MDC, soms onzichtbaar, zoals query re-write. Sommige automerken hebben relatief veel nieuwe technologie uitgevonden. Denk aan Saab. In de databasewereld is het niet anders: IBM heeft met DB2 veel hightech features uitgevonden, die later opduiken in andere databases. ●

Klaas Brant (kbrant@kbce.nl) is DB2-specialist en directeur van KBCE. Meer informatie over DB2 is te vinden op www.kbce.nl

