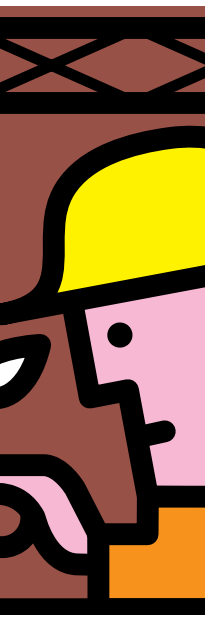

XML uitermate geschikt voor e-mail archivering

Zoeken in niet-gestructureerde data

Teus Molenaar

Als je nagaat dat relationele data minder dan tien procent omvatten van de totale informatie binnen een organisatie, dan is een XML database een goed alternatief. “E-mail archivering bijvoorbeeld wordt heel belangrijk”, zegt Jeroen van Rotterdam, directeur van X-Hive. “En dan heb je het vaak over documenten die in XML modellen zijn gegoten. Dergelijke documenten indelen in een relationele database is erg moeilijk en zeker niet efficiënt.”

We spreken Jeroen van Rotterdam een paar dagen nadat publiek is gemaakt dat EMC het Rotterdamse bedrijf X-Hive heeft ingelijfd. Deze Amerikaanse leverancier van opslagsystemen heeft de laatste jaren zijn werkveld verruimd naar alles dat te maken heeft met informatieverwerking onder de bijpassende slagzin ‘Where information lives’. Het bedrijf heeft tal van overnames gedaan om producten verder te ontwikkelen voor doelmatige en kostenefficiënte opslag en vindbaarheid van documenten. De belangrijkste aankoop op dit vlak was Documentum, leverancier van Enterprise Content Management software, eind 2003. Met de groei van wat heet ‘ongestructureerde’ data was EMC op zoek naar een database die goed overweg kan met dit type informatie.

> Een technische omschrijving heeft een andere structuur dan een lijst met vragen en antwoorden

“Ze hebben alle bekende databases onderzocht en wij zijn uiteindelijk als de beste uit de bus gekomen”, aldus een trotse Van Rotterdam. “Toen hebben ze nog overwogen of EMC een OEM-overeenkomst zou sluiten met X-Hive, maar dat zou de onderneming té afhankelijk maken van onze strategieën. Dus is besloten ons te kopen.”
Het voordeel voor X-Hive is dat het complete verkoop- en marketingapparaat van EMC de producten van de Rotterdamse fabrikant gaat slijten. Ook zullen de services- en support-afdelingen van EMC zich gaan inzetten voor X-Hive. “Dat is al heel wat”, zegt Van Rotterdam. “Wij zijn namelijk altijd alleen een productenbedrijf geweest. Natuurlijk leveren we wel ondersteuning, maar geen consultancy en zo. Wij maakten en maken software, niets meer, niets minder. De helft van onze omzet is altijd in research gegaan.”

Grootste klanten

Waarom heeft EMC zijn oog laten vallen op het softwarehuis uit onze havenstad? Wat is er zo interessant aan een organisatie die een XML database weet te bouwen? De overstelpende vloed aan documenten die de wereld overspoelt, zo luidt het antwoord. Een relationele database moet een vertaalslag maken om XML documenten netjes op te slaan en vervolgens weer terug te vertalen als ze worden opgevraagd. Dat kost tijd en dat was in sommige gevallen (zoals bij vliegtuigmaatschappijen) toen al een probleem, maar de verwachting is dat dit bij meer organisaties een uitdaging gaat worden. De strakkere regelgeving voor bedrijven en overheden eist dat veel meer informatie moet worden bewaard en moet worden overlegd als een toezichthouder daar om vraagt. Vaak gaat het dan om Office-documenten (zoals tekstbestanden en presentaties) en natuurlijk e-mail.
Het product X-Hive/DB is voor het eerst in december 2000 uitgebracht. Inmiddels is de zevende versie (7.6) op de markt. “Er is een enorme hoeveelheid productontwikkeling in gaat zitten. En we zijn alleen maar beter en beter geworden”, zegt Van Rotterdam.
Hij vertelt dat het bedrijf zich altijd op de grootste klanten heeft gericht; die met veel data en complexe documenten. Zo zijn Boeing, Fokker Services, Harley Davidson en Northwest Airlines gretige afnemers van de XML database. Daarbij gaat het voornamelijk over de handleidingen en beschrijvingen van de producten die deze bedrijven maken. “Het verhaal gaat dat als je alle documentatie van een vliegtuig op papier publiceert, datzelfde vliegtuig te klein is om die papierstroom te herbergen”,

zegt Van Rotterdam. Harley Davidson gaat er prat op dat elke motorfiets een uniek exemplaar is, ook al wijkt een bepaalde motor alleen maar af van zijn broertje door een andere kleur benzinetank. Het geeft te raden wat een hoeveelheid gegevens zo'n administratie oplevert. Elke verkochte motorfiets heeft zijn eigen beschrijving, gekoppeld aan de (huidige) eigenaar.

> XQuery heeft veel meer mogelijkheden dan SQL voor een relationele database

Hetzelfde geldt voor vliegtuigen. Elke moertje en nippeltje is gedocumenteerd. En voordat een vliegtuig weer de lucht in gaat, moeten heel veel onderdelen eerst nog eens worden gecontroleerd. Daarom is een database waarin deze informatie heel snel is terug te vinden niet alleen voor vliegtuigfabrikanten van belang, maar evenzeer voor de vliegtuigmaatschappijen, zoals Air France-KLM. De piloot mag immers pas het vliegtuig in de lucht brengen als de protocollen zijn doorlopen. Voor de controleurs en monteurs is het dan zaak heel snel te kunnen zoeken in die enorme berg informatie.

Blob's

De XML database kan als een eigen server runnen, vertelt Van Rotterdam. "Wij ondersteunen replicatie, maar de volgende stap is dat we naar een gedistribueerde database gaan. Daar werken we nu aan in het ontwikkelteam. Dan heb je de mogelijkheid om te clusteren. Dat moet ook wel, want als je kijkt naar e-mail archivering bij grote organisaties dan heb je het toch wel over honderden miljoenen documenten per jaar die ergens een plekje moeten krijgen en vervolgens toch snel zijn op te zoeken. In de praktijk blijken bedrijven die e-mail archieven toch vrij lang doorzoekbaar te houden."

"Wij kunnen nu een heel eind opschalen", reageert Van Rotterdam op de vraag of de huidige database niet is opgewassen tegen dergelijke hoeveelheden. "Wij hebben klanten in een Terabyte-omgeving met alleen XML documenten. En dan zitten we nog steeds op subsecond vindsnelheden; dat is heel netjes. Wij testen zelf met een database van een halve Terabyte en dan komen we op een vindtijd van dertien milliseconden."

In de XML database zitten dus XML documenten, hoe zit het met beeldmateriaal? "Nou, dat is wel interessant", begint Van Rotterdam. "Wij kunnen in onze database ook Blob's (Binary Large Object) opslaan. Al onze klanten doen dat ook. Het voordeel is dat je een transactiemechanisme, versioning en branching, en dergelijke hebt. Je hangt daar metadata aan en daar kun je full text searches op doen. Je kunt zelfs Xquery's over metadata van Blob's runnen. Het interessante is om te zien dat bijna alle binary formaten langzaam opschuiven naar XML. Microsoft Word is daar een mooi voorbeeld van. De versie 2007 is eigenlijk een XML formaat. In images zie je nu in het image-formaat XML metadata ontstaan. Als je dan een JPEG-image in de database opslaat, dan is het een koud kunstje om de XML metadata eruit te halen en die als XML document in X-Hive/DB op te slaan. Dan kun je het volledig doorzoeken over de metada die in het image zelf zitten. Adobe is erg actief op dat gebied."

Van Rotterdam zegt dat we Blob's niet moeten onderschatten. Hij meent dat de inhoud van documentgeoriënteerde systemen voor tachtig procent uit Blob's bestaat. Dan gaat het niet alleen om stilstaand beeld, maar ook om bewegend beeld. Dat groeit heel erg hard. Videovergaderingen willen we immers ook bewaren en doorzoekbaar maken. XML is daar de oplossing voor. Als je metadata aan de beelden kunt toekennen, dan kun je ze ook doorzoeken. "Voor X-Hive is dit een gunstige ontwikkeling."

Andere architectuur

Toen X-Hive zijn XML database aan het bouwen was, waarde Tamino van Software AG al rond. Volgens Jeroen van Rotterdam, directeur van X-Hive, heeft zijn database een heel andere

Van eigen bodem

X-Hive/DB, dat inmiddels de zevende versie kent, is van eigen bodem. Fabrikant X-Hive heeft zijn hoofdzetel in Rotterdam. Er is al langer een nevenvestiging in de VS, maar het ontwikkelwerk blijft, ook na de overname door EMC, in de havenstad. "Het heeft geen enkele zin om naar de VS te verhuizen met ons bedrijf. Dankzij internet maakt het niet uit waar je bent gevestigd. En 'emigreren' geeft alleen maar veel overlast en nauwelijks enig voordeel. Misschien is er wel meer talent in Silicon Valley – en dat kunnen we nu wel gebruiken, omdat we flink gaan uitbreiden – maar talent heeft daar de neiging om binnen het jaar al weer te vertrekken naar de volgende baan. Ons product is erg ingewikkeld. Het kost wel een half jaar om daar goed in te geraken. Dan blijf ik liever in Rotterdam. Onze medewerkers zijn erg loyaal aan het bedrijf; wij hebben nauwelijks verloop", vertelt X-Hive-directeur Jeroen van Rotterdam.

Het bedrijfje is overigens niet meteen begonnen met de bouw van een XML database. "We zijn in januari 1996 begonnen met dit bedrijf. Aanvankelijk als research club om software te ontwikkelen. In tussentijd zouden ze projecten doen met het idee dat er wel een goed idee zou komen. Maar dat pakte niet zo goed uit. Als één van ons een idee had, dan zeiden de andere twee: 'Nwah; dat wordt niks.' Zo hebben we – achteraf gezien – briljante ideeën laten lopen."

Van Rotterdam vertelt dat allengs bleek dat er een behoefte ontstond aan medianeutraal uitgeven. "Hoe sla je informatie efficiënt en doelmatig doorzoekbaar op? Op zo'n manier dat diezelfde informatie op verschillende manieren is te gebruiken. Zo is onze XML database ontstaan, we waren er vroeg bij. Vergeet niet dat de XML standaard uit 1998 stamt."

Combinatie VMware en X-Hive

"Het is toch wel erg leuk om te constateren dat een Nederlands bedrijfje zich zo in de kijker heeft weten te spelen", is het eerste wat Hans Timmerman kwijt wil. Hij is als Technology Officer mede verantwoordelijk voor strategische samenwerkingen bij EMC in Nederland. Hij weet niet wat de echte redenen zijn geweest van de raad van bestuur om X-Hive te kopen, maar heeft er wel een persoonlijke mening over.

Hij vertelt dat EMC interessante bedrijven opkoopt om deze verder door te ontwikkelen naar een eigen product. Dat is gebeurd met het Vlaamse Filepool dat heeft geleid tot de archiveringsoplossing Centera van EMC. "We kennen relationele databases. Objectgeoriënteerde databases zijn nooit echt doorgebroken; de technologie bleek te complex. Maar met XML kan dat wel; zeker voor ongestructureerde data. Op dit moment zijn er drie belangrijke database-leveranciers: Oracle, IBM en Microsoft. Met X-Hive heeft EMC de mogelijkheid uit te groeien tot de vierde leverancier in rangorde, maar dan wel XML gebaseerd."

Dat X-Hive een rol gaat spelen binnen het Documentum-portfolio acht Timmerman vanzelfsprekend. Een goede XML database helpt de op XML en Java gebaseerde content management suite vooruit. "Maar misschien is er meer. EMC werkt met VMware aan Virtual Desktop Infrastructures en de toepassing van virtual appliances. Je beschikt dan over een lege PC of notebook en op een host of een USB stick heb je een stukje besturingssysteem met de nodige appliances staan, zodat je altijd jouw eigen omgeving bij je hebt. Je doet je werk op de PC, haalt de USB stick er weer uit of je *offloadt* naar de host. Bij diefstal van de machine ben je dan in elk geval geen gegevens kwijt. Je kunt de appliance ook helemaal in Java schrijven, dan heb je zelfs geen besturingssysteem meer nodig. Een heel interessante ontwikkeling, bijvoorbeeld voor kritieke en geclassificeerde omgevingen. Je hebt natuurlijk een database nodig; gebaseerd op XML in dit geval. Met VMware en X-Hive heb je dan een krachtige combinatie te pakken."

architectuur dan Tamino. "Onze kernel is een persistente DOM-implementatie. En ons storage-model – met name de indexstructuren – is volledig geoptimaliseerd voor XML data, dus niet afhankelijk van schema's of DTD's (Document Type Definition). Daardoor kunnen we heel snel zoeken. Wij zoeken onafhankelijk van de schema's." DOM staat overigens voor Document Object Model, een specificatie van W3C voor het aanpassen en bekijken van webpagina's, onafhankelijk van de gebruikte taal of het platform.

> Het interessante is om te zien dat bijna alle binary formaten langzaam opschuiven naar XML

Die onafhankelijkheid is ook wel nodig, omdat geen enkele klant maar één schema of één DTD hanteert. In een handleiding bijvoorbeeld zitten verschillende documentstructuren. Een technische omschrijving heeft een andere structuur dan een lijst met veelgestelde vragen en hun antwoorden. Een zoekopdracht moet in beide soorten documenten even snel zijn weg kunnen vinden. "Je kunt ook heel gericht zoeken. Dat is wat XQuery doet: heel duidelijk zoeken in de context. Dat is de kracht ervan. Je kunt bijvoorbeeld zeggen: ik wil zoeken waar een bepaald woord voorkomt in de titel van het document, met meneer Pieterse als auteur en de aanwezigheid van een samenvatting in het document. Dat soort combinaties kun je als zoekopdracht meegeven."

Terug naar het concurrentieveld. Van Tamino zegt Van Rotterdam nauwelijks 'last' te hebben. "Die komen we bijna niet tegen in de praktijk. In eerste instantie moeten wij altijd bewijzen dat we sneller zijn dan Oracle, of DB2. Want dat is toch de corporate standaard. Wij concurreren met traditionele databases. Maar er is ook een andere klasse XML databases zoals Viper van IBM, de DB2-database die van nature XML bestanden kan opslaan. "Wij opereren echter alleen bij de heel grote organisaties, en daar komen we niet veel concurrenten tegen."

XQuery

Als je documentgeoriënteerde data hebt, dan heb je een prachtige, wat Van Rotterdam noemt: *processing pipeline*.
"Je hebt de database, je runt een XQuery, daar komt een XML document uit (daar kun je constructors voor gebruiken in XQuery) en daar kun je vervolgens een transformatie op loslaten om bijvoorbeeld pdf of html te krijgen. De applicatie verschuift dan heel erg naar het schrijven van Xquery's in plaats van dat je programmeert. Database, XQuery, XML document als resultaat, transformatie, eindgebruiker. Dat is het proces", stelt Van Rotterdam.
XQuery, ontstaan uit Quilt, is als query-taal ontwikkeld door de standaardencommissie W3C, waarvan X-Hive sinds 2001 lid is. Van Rotterdam is er helemaal weg van. "XQuery heeft veel meer mogelijkheden dan SQL voor een relationele database. Het is bijna een programmeertaal. Het is extreem flexibel."
Met X-Hive kan iemand heel makkelijk uit de voeten. Onder voorwaarde dat de standaarden als DOM, XQuery, XPath, XLink en XPointer gesneden koek zijn. "Als je daarmee bekend bent, dan leer je in een halve dag met X-Hive/DB om te gaan."
De database is overigens in puur Java geschreven, zodat hij platformonafhankelijk is in te zetten.

Doordrongen

Volgens Van Rotterdam is het bedrijfsleven inmiddels wel doordrongen van het nut van metadata aan bestanden toevoegen. Hij ziet dit belang onder andere terug bij e-mail archivering. En daar is de noodzaak vooral ingegeven door de scherpere wet- en regelgeving.
"E-mail kun je heel makkelijk structureren in XML en de volumes zijn zo groot, en de combinatie van gestructureerd en full text search is zo belangrijk, dat dit alles een goede aanzet is om een XML database in te zetten om contextgevoelig te kunnen zoeken. Deze ontwikkeling is uiterst interessant."

Teus Molenaar is freelance journalist.

Update

Cognos gaat Applix overnemen

Cognos en Applix hebben samen bekendgemaakt een definitieve overeenkomst voor de overname van Applix door Cognos verder uit te werken.
Met deze overname verwacht Cognos zijn positie als toonaangevende onafhankelijke leverancier van financiële Performance Management te versterken.

De geplande overname betreft de verwerving van aandelen in contanten voor een bedrag van netto circa 306 miljoen USD. Applix zal Cognos 8 Planning, Cognos 8 Controller en Cognos 8 Business Intelligence gaan completeren, in het bijzonder wat betreft financiële Performance Management, bijvoorbeeld verbeterde analyse en optimalisatie van grote hoeveelheden complexe financiële data, sterke functionaliteit voor zelfservice zoals business rules management; nieuwe oplossingsgebieden zoals analyse van de winstgevendheid; en innovatieve technologie zoals Applix TM1, een gepatenteerde, 64-bit, in-memory en multidimensionale OLAP-server.

Bull in zee met DATAllegro en QlikTech

DATAllegro, leverancier van datawarehouse-applicaties, en Bull zijn een strategische samenwerking aangegaan.
Als onderdeel van de overeenkomst introduceren DATAllegro en Bull de NovaScale Datawarehouse-serie.
Deze bestaat uit een combinatie van de DATAllegro-software en Bull NovaScale-servers. De oplossing wordt geproduceerd en verkocht door Bull. Ondersteuning en onderhoud worden verzorgd door beide partijen.
Door de samenwerking tussen DATAllegro en Bull wordt een complete oplossing geboden: van productie tot een Europees netwerk voor onderhoud en support. Daarnaast profiteren klanten van de kennis en ervaring van Bull op het gebied van hoogwaardige mainframe-omgevingen en implementaties van grote, bedrijfsbrede datawarehouses.
Speciaal voor de 'midmarket' ontwikkelt Bull in samenwerking met QlikTech, pionier op het gebied van 'in-memory' BI, onder de naam QlikView een innovatieve benadering van BI. Partijen willen zo middelgrote bedrijven en organisaties uit de publieke sector geïntegreerde BI met een snelle ROI aanbieden.

Microsoft breidt BI verder uit

Onlangs is de Microsoft Office PerformancePoint Server 2007 gelanceerd. Dit product is een belangrijke uitbreiding op het gebied van Business Intelligence. Office PerformancePoint Server 2007 heeft de bekende interface van Microsoft Office. Microsoft meldt dat de applicatie is gebouwd om te voldoen aan de functionele en prestatie-eisen van grote ondernemingen, maar zal zodanig concurrerend geprijsd worden dat bedrijven van elke omvang Office PerformancePoint Server 2007 breed kunnen inzetten.
De geïntegreerde Microsoft Business Intelligence-productlijn, die verder bestaat uit SQL Server en het Microsoft Office system, geeft meer mogelijkheden om BI-oplossingen te bouwen die zijn afgestemd op de specifieke behoeften van een bepaalde klant of branche. Microsoft heeft inmiddels wereldwijd bijna 1000 partners opgeleid voor Office PerformancePoint Server 2007.