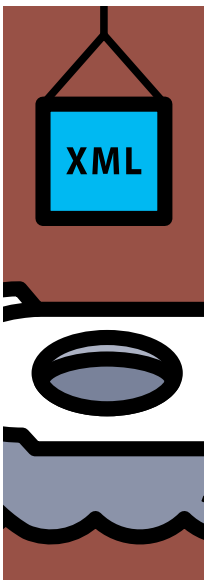


XML databases ExistDB en Galax onder de loep

Bruikbare alternatieven?

Jos van Dongen

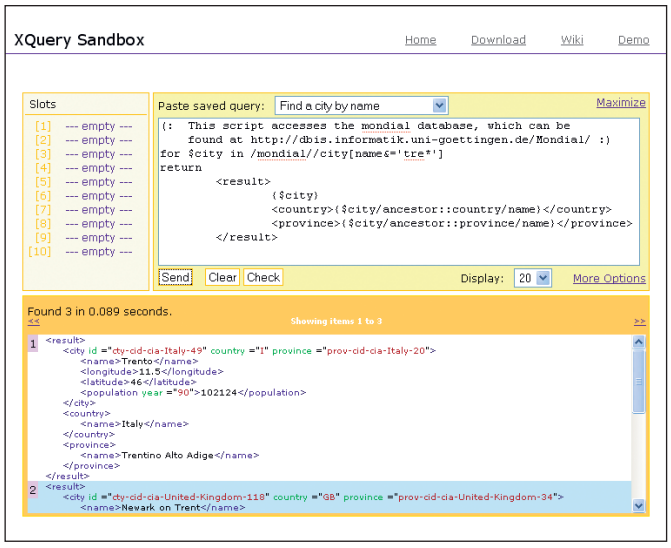
De markt voor pure play XML database vendors wordt er niet duidelijker op, nu vrijwel alle leidende RDBMS leveranciers hun producten uitrusten met native XML opslag-, index- en query-mogelijkheden. In de Open Source-wereld ligt het wat genuanceerder.

De bekendere spelers als Postgresql en MySQL beschikken (nog) niet over een native XML datatype en XQuery support, terwijl dit voor een embedded database als bijvoorbeeld BerkeleyDB wèl geldt. Ook het Nederlandse OSDB vlaggenschip MonetDB is in een XQuery variant beschikbaar en maakt zoals gebruikelijk weer gehakt van alle andere XML databases in de benchmarks. Ondanks de beschikbare alternatieven blijft, als we naar Open Source producten kijken, een tweetal pure play XML databases de populariteitslijst aanvoeren. Het gaat hier om de producten Exist en Galax, die hieronder verder besproken zullen worden.

Historie

Zowel Exist als Galax zijn zoals zoveel OS-projecten gestart als eenmansproject, beide zo'n zes jaar geleden. De man achter Galax is Jérôme Siméon, één van de editors van de W3C XQuery 1.0 standaard. Ook Mary Fernandez van AT&T Labs vinden we terug als editor van de XQuery werkgroep en als één van de drijvende krachten achter Galax. Het doel van het Galax team is dan ook om een volledige implementatie van de XQuery 'working drafts' te bieden. Galax positioneert zichzelf in eerste instantie als leer- en experimenteermiddel, wat niet verwonderlijk is gezien de research-achtergrond van het project.

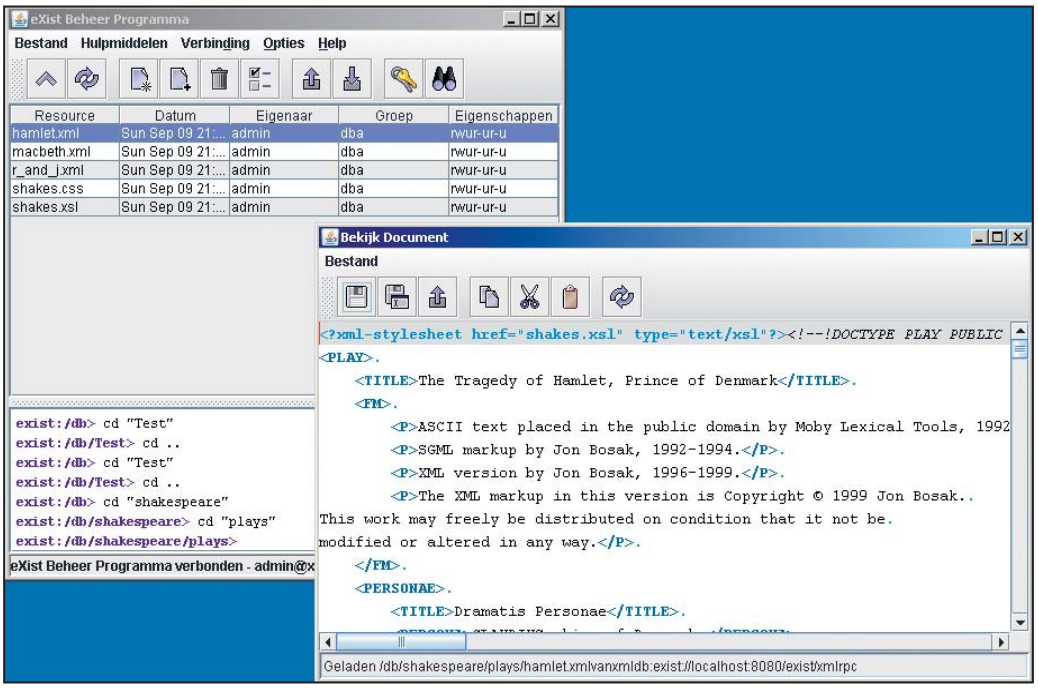
ExistDB heeft een wat andere achtergrond en is door Wolfgang Meier oorspronkelijk ontwikkeld als XML middleware voor MySQL. Het idee was om XML documenten relationeel op te slaan en door middel van XPath opvraagbaar te maken. In de huidige vorm biedt ExistDB een complete XML opslag- en opvraagoplossing zonder de noodzaak voor aanvullende software. De plannen voor de toekomst gaan een stuk verder dan een volledige XQuery ondersteuning, hoewel daar in de roadmap (zie ook http://exist.sourceforge.net/roadmap.html) wel de hoogste prioriteit aan wordt toegekend. Exist is namelijk niet alleen een XML database maar tevens een omgeving om webapplicaties te ontwikkelen met behulp van XQuery en gerelateerde standaarden als XSLT, XHTML en CSS, eventueel aangevuld met JavaScript voor het creëren van de tegenwoordig zo populaire AJAX-effecten.



Afbeelding 1: ExistDB XQuery Sandbox.

De praktijk

Om met beide producten te experimenteren hoeft er voor de verandering helemaal niets gedownload en geïnstalleerd te worden. ExistDB biedt een 'XQuery Sandbox' waarmee voorgedefinieerde query's kunnen worden uitgevoerd, maar waarmee ook zelf naar hartelust geknutseld kan worden (zie afbeelding 1). Ook met Galax kan online worden gewerkt, met als leuke bijkomstigheid dat alle onderdelen van het verwerken van de XQuery kunnen worden getoond. Maar goed, online spelen is leuk ter lering ende vermaak, voor het echte werk is het toch



Afbeelding 2: ExistDB beheer.

nog altijd het beste om de spullen gewoon te installeren op de eigen PC. Dit laatste is overigens alleen gebeurd met Exist; wie zich wil wagen aan de installatie van Galax verwijs ik graag naar www.galaXQuery.org/doc/install.html.
ExistDB is volledig in Java geprogrammeerd en heeft alleen een JVM nodig om te kunnen draaien. Dat betekent dat na het downloaden en uitvoeren van het java -jar eXist-[Version].jar commando de software is geïnstalleerd. Na het starten van de database engine is Exist klaar voor gebruik. Exist komt met een eigen voorgeïnstalleerde webserver (Jetty) en door in de browser het lokale adres http://localhost:8080/exist/index.xml in te tikken verschijnt een bekend scherm. De Exist website en de lokale versie die zojuist is geïnstalleerd zijn dan ook identiek aan elkaar.

> ExistDB is volledig in Java geprogrammeerd en heeft alleen een JVM nodig

Er zijn twee manieren om met de database te communiceren: de Java front-end en de web-interface. Welke er gebruikt wordt hangt van het doel af: om documenten te bekijken die in de database zijn opgeslagen werkt de Java client wat vlotter (zie afbeelding 2), terwijl de web-interface weer meer informatie geeft over bijvoorbeeld de omvang van de documenten. Hier valt ook meteen te zien dat de indexering van de documenten zo'n 30 procent extra opslagruimte kost. XML documenten worden opgeslagen in zogenaamde 'collections' die dienen als container, zoals ook folders dienen als container voor files. Ook zijn er voorzieningen om gebruikers en gebruikersgroepen te beheren en vervolgens aan een collectie of een individueel document bepaalde groepen wel of geen lees-, schrijf- en update-rechten toe te kennen.

Het installeren van dezelfde voorbeeldcollecties en -files die ook online beschikbaar zijn is vanuit de Web admin module een fluitje van een cent, waarna ook meteen alle voorbeeld query's beschikbaar zijn in de 'Sandbox'. De voorbeelden zijn beperkt van omvang, dus het zal geen verbazing wekken dat de performance erg goed is. Om toch een indruk te krijgen van de prestaties met wat grotere documenten is de XMark generator van het CWI gebruikt om een 11 MB (factor 0.1) document te maken en vervolgens in de database op te slaan. Zelfs dit relatief kleine document geeft Exist al aardig wat te doen, blijkens de responstijd van de XMark query's. Met vervolgens alleen al het opslaan van een 110 MB (factor 1) document heeft Exist behoorlijk wat moeite, laat staan het behalen van een acceptabele responstijd. Exist lijkt dan ook eerder een lichtgewicht database die kan dienen als basis voor een interactieve, op XQuery gebaseerde website, dan een full-blown XML database met de bijbehorende schaalbaarheid en performance. Toch zijn er verschillende organisaties die Exist voor dit doel inzetten. Eén van de meest in het oog springende is ASML, dat de database gebruikt voor de opslag van ontwerpdocumenten.

Uitdagingen

Voor zowel Galax als ExistDB geldt dat schaalbaarheid en performance de grootste uitdagingen vormen. Schaalbaarheid

heeft in dit opzicht meerdere betekenissen: het gaat daarbij niet puur om het totale opslagvolume, maar ook om de omvang van een enkel XML document. Als gekeken wordt naar berichten in user groups of recente presentaties op congressen, blijken databases groter dan 100 GB of individuele documenten groter dan 100 MB nog voor aardig wat problemen te zorgen, wat ook overeenkomt met mijn eigen waarnemingen. Om de diverse XML database-producten onderling vergelijkbaar te maken voor wat betreft prestaties is er bij het Centrum voor Wiskunde en Informatica (CWI) een XML benchmark ontwikkeld, de al eerder genoemde XMark. Zoals ook bij andere database benchmarks gebruikelijk is, wordt er een aantal (20 in dit geval) verschillende query's gedefinieerd waarbij de te bevragen database telkens een stuk groter wordt gemaakt. Op http://monetdb.cwi.nl/projects/monetdb/XQuery/Benchmark/XMark/ kunt u zelf zien dat ExistDB bij de 110 MB XMark niet meer in het lijstje voorkomt (wat inhoudt dat geen enkele query uitgevoerd kon worden), en dat hetzelfde geldt voor Galax bij de 1,1 GB XMark-test. De enige producten, naast uiteraard MonetDB XML, die de 11 GB test nog aankunnen zijn het commerciële product X-Hive, en BerkeleyDB XML van Oracle. Naast de prestaties geven ook de 'draft' status van de diverse XQuery onderdelen de producenten de nodige kopzorgen, maar dit zou volgens de W3C planning over enige maanden voorbij moeten zijn.

Conclusie

De meeste Open Source XML databases ademen nog sterk een R&D sfeer. Het is niet verwonderlijk dat Galax zichzelf vooral positioneert als leermiddel om kennis te maken met XQuery, gezien de achtergrond van de drijvende krachten achter dit product. ExistDB is wat ambitieuzer en wil naast een XML database een complete op XQuery gebaseerde web-ontwikkel-omgeving bieden. Wat ook opvalt is dat het weliswaar XML databases zijn, maar dat het DBMS-deel onvergelijkbaar is met producten als MySQL en PostgreSQL. Voorzieningen voor concurrency, transacties, beveiliging, backup/restore, etcetera zijn niet of nauwelijks voorhanden. Kortom, wilt u kennismaken met XML databases en Xquery, dan voldoet (vooral door de eenvoudige installatie) Exist prima.

Bent u echter op zoek naar een XML database voor bedrijfskritische toepassingen, zoek dan vooral nog even verder. Met het complete overzicht op www.rpbourret.com/xml/ProdsNative.htm kunt u vast een begin maken.

Jos van Dongen
Jos van Dongen (jvdongen@tholis.com) is Senior Consultant bij Tholis Consulting.