Oracle en het semantic web

RDF-ondersteuning in 10gR2 van groot belang

In situaties waarin het niet om klassieke transacties gaat, voldoen standaardtechnieken als SQL en Java niet altijd. RDF met de bijbehorende query-taal SPARQL maakt het mogelijk om vrijwel elke denkbare informatie te beschrijven en te bevragen. Voor toepassingen in de biowetenschappen is dat van groot belang, maar ook de dataintegratiemogelijkheden zijn indrukwekkend. Oracle ondersteunt – als eerste grote software-leverancier – RDF sinds 10gR2 en heeft zelfs een heel performante manier gevonden om deze data op te slaan en te bewerken. Optimize sprak met Dr. Susie Stephens, principal project manager life sciences.

Dr. Susie Stephens studeerde biologie. Eerst algemene biologie, daarna deed ze een fysiologie PhD, en ten slotte een postdoctorale studie microbiologie. Daarna werkte ze vier jaar bij Sun. Sinds 2001 is ze werkzaam bij Oracle.

Rekenkracht

Waarom bent u overgestapt van Sun naar Oracle?

Stephens: 'Ik werkte bij Sun Microsystems eerst drie jaar als pre sales systems engineer waarbij ik me vooral op biotechnologische bedrijven geconcentreerd heb en daarna werkte ik er twee jaar als global life sciences technical manager. Ik begon bij Sun omdat ik dacht dat er veel interessante IT-uitdagingen waren op het vlak van biowetenschappen, én dat de meeste daarvan gerelateerd waren aan rekenkracht. Na verloop van tijd werd ik me er echter steeds meer van bewust dat de werkelijke uitdaging bestond in het beheren van de informatie, en daarom stapte ik over naar Oracle, hét informatiemanagementbedrijf. Ik ben begonnen als productmanager voor de biowetenschappen. In die functie richtte ik me vooral op het verbeteren van Oracle core-technologieproducten (de database, de middleware en ook de collaboration tools) om ervoor te zorgen dat ze net zo goed werkten met wetenschappelijke data als met business data.
Toen ik nog maar pas bij Oracle werkte, kreeg ik het verzoek om support voor RDF in de database te brengen. Ik begon dus te werken met de semantic-webtechnologieën en begon RDF in de database te brengen. Die taak heb ik nu nog steeds, maar omdat er zoveel meer bedrijven geïnteresseerd zijn in de semantische mogelijkheden van de database slokt het een steeds grotere hoeveelheid tijd op. Ik heb bovendien mijn terrein een beetje buiten de biowetenschappen verlegd, omdat ik Oracle vertegenwoordig in discussies met W3C omtrent het semantic web en de bijbehorende standaarden en recentelijk ben ik uitgenodigd om de nieuwe chair-voorzitter te worden van de W3C over het *semantic web education and research group.* Dat heeft effect op verschillende industrieën, we doen meer dan alleen biowetenschappen.'

Triplet

RDF is heel anders dan de relationele manier om data te beschrijven en op te slaan. De structuur ervan lijkt op een taal waarin je alles kunt uitdrukken wat je wilt.

Stephens: 'RDF is gebaseerd op de notie van triplets en knopen, oftewel subject (onderwerp), predikaat (gezegde) en object (lijdend voorwerp), zodat je heel gemakkelijk uitspraken kunt doen over een feit, een associatie of zelfs uitspraken over uitspraken. Ik zal je mijn favoriete voorbeeld geven: ik kan een uitspraak doen over Susie. Het predikaat zou leeftijd kunnen zijn, en je zou een literal als object kunnen hebben, en dan zeggen: 21. Erik (*Erik Zwanenburg, red.*) zou die informatie grinnikend kunnen aanhoren en het niet geloven, dus je zou Erik als het subject van een ander triplet kunnen nemen, als predikaat zou je kunnen kiezen "gelooft niet" en dan kun je verwijzen naar de eerste triplet. Je hebt dus heel veel flexibiliteit met de triplet-representatie, maar je hoeft niet de volledige flexibiliteit te gebruiken als dat niet nodig is. Een databasetabel is in feite gebaseerd op de notie van een triplet: je hebt een rij, een kolom en een waarde.'

Dus het werkt ook de andere kant op.

Stephens: 'Je kunt dus RDF-data als een tabel weergeven, maar ook in een boomstructuur om het meer als XML eruit te laten

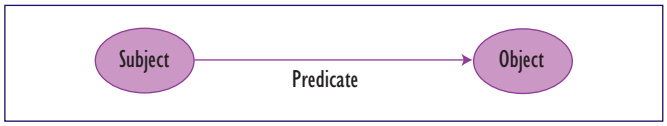
zien, of je zou er een graaf van kunnen maken, wat een meer natuurlijke representatie is *(red.: een graaf bestaat uit een verzameling punten, knopen genoemd, waarvan sommige verbonden zijn door lijnen, de zijden of kanten)*. De triplet heeft dus een zekere correlatie met het relationele model.
Je hebt de onderliggende triplet-representatie en de *unique indentifiers* voor iedere component van de triplet. Je maakt met de *unique indentifiers* heel precies duidelijk waarover je het hebt. Dus in plaats van een *loose term* (afzonderlijke term) te gebruiken, zeg je zeer expliciet waarover je uitspraken wilt doen voor iedere component van een triplet. Om een voorbeeld uit de biowetenschappen te gebruiken: als je een unieke identifier geeft aan een potentieel medicijn dat door het medicijnontwikkelproces gaat, kun je beginnen alle informatie over dat kandidaat-medicijn die unieke identifier mee te geven. Anders is het heel moeilijk om dat verband te leggen omdat die informatie verdeeld is over heel veel silo's door het gehele bedrijf. Het bedrijf kan dan nooit een compleet beeld krijgen. Als je dus data in RDF hebt, kun je linken naar welke andere data dan ook in RDF. Daardoor wordt het een zeer flexibel integratieplatform. Je kunt het vergelijken met webpagina's die naar iedere andere webpagina kunnen linken, met dit verschil dat je praat over data in plaats van over een type document, een fijnere granulariteit dus.'

Ontologie

Daarmee wordt het dus ook geschikt om integratie tot stand te brengen, misschien zelfs voor een nieuw soort ESB-ontwerp?
Stephens: 'Ja, maar de data in de back-end zouden opgeslagen kunnen zijn in XML, een spreadsheet of een Word-document en via het web beschikbaar gesteld kunnen worden. Om terug te gaan naar de basisidee: RDF is de kern die onder de representatie van het semantic web ligt. De webontologielaag die daarboven zit, is het niveau waarop je informatie over een zeker domein kunt representeren. Dus wanneer je je domein gemodelleerd hebt, kan RDF dat onmiddellijk verbinden met de ontologie. De ontologie heeft hiërarchieën en kan data representeren op verschillende abstractieniveaus. Dus het onderliggende basismodel van het semantic web is RDF en OWL (Web Ontology Language) standaarden. Wat het representeren daarvan betreft kun je data onmiddellijk in RDF creëren of data erin converteren en ze dan opslaan in Oracle's RDF-model. Een andere benadering die minder rijp is: het *on the fly* mappen van een bepaalde databron naar RDF, mensen werken aan *on the fly*-mapping van relationele databronnen naar RDF en van bijvoorbeeld XML naar RDF. Met die mappings proberen ze doorgaans data uit veel verschillende bronnen te integreren: relationeel, XML, spreadsheets, noem maar op.'

Hoe gaat dat vertalen in zijn werk?
Stephens: 'Je hebt standaarden voor RDF en OWL, en mensen werken nog aan de tools en standaardtechnologieën om mappings te doen van de databronnen naar RDF. Het W3C heeft een werkgroep voor vertalingen van XML naar RDF, maar op het moment is er nog niets voor de vertaling van relationeel naar RDF. Maar het doel is wel om ook dat mogelijk te maken.'

Het vertalen van XML naar RDF lijkt me moeilijk, omdat XML-documenten zo verschillend van structuur kunnen zijn, beter gezegd: niet zo gestructureerd zijn als je zou wensen.
Stephens: 'W3C werkt aan een technologie genaamd griddle die dat voor zijn rekening neemt. Ik ben geen expert op het gebied van de details van de griddle-specificaties omdat ik meer geïnteresseerd ben in relationeel naar RDF-mapping, maar ze hebben een hoop vooruitgang geboekt op dat gebied. Alhoewel RDF in zijn meest natuurlijk representatie een graaf is, kan het



RDF is gebaseerd op de notie van triplets en knopen.

op verschillende manieren geserialiseerd worden en kan er dus een XML-serialisatie naar RDF plaatsvinden. Wanneer mensen eerder een conversie willen doen dan een *on the fly mapping*, zullen ze kijken naar technologieën als XSLT of XQuery. Het gebruiken daarvan voor de mapping van XML naar RDF is vrij volwassen, dus ze zullen gebruik maken van die kennis voor *on the fly mapping*.'

De structuur van RDF lijkt heel erg geschikt om taal te analyseren, met gebruik van de Chomkiaanse transformationele generatieve grammatica.

Stephens: 'Ik heb me meer gericht op het gebruik van RDF voor data-integratie dan op meer complexe zaken. Het semantic web kan ook gebruikt worden voor zoekmogelijkheden. Oracle gebruikt het RDF-datamodel samen met klassieke secure enterprise search software, en software van Hyperion voor onze OTN-website. Ik kom ook mensen tegen die geïnteresseerd zijn in de inferentiecapaciteiten van RDF. Het zou bijvoorbeeld artsen kunnen helpen de beste behandeling te kiezen, op basis van alle condities die ze kennen (het gehele ziektebeeld tot dan toe). Wanneer er iets nieuws gediagnosticeerd wordt, helpt ze dat om de patiënt te behandelen op basis van de gehele geschiedenis van mogelijke behandelingen van ziekten en misschien de andere geneesmiddelen die ze gebruiken op dat moment.'

Inferentie

Stephens: 'Veel triplets kunnen ook ouder-kind relaties zijn waarbij ook de capaciteiten van Oracle om de hoek komen kijken. In 10gR2 hebben we RDF- en RDFS-ondersteuning in de database, en we hebben een objectrelationele implementatie en we gebruiken de subject en de objecten van de triplet. Dat wijkt af van wat andere mensen doen waardoor we een veel schaalbaarder implementatie hebben. We zijn de enige enterprise vendor die RDF ondersteunt maar er zijn een aantal kleinere partijen die triplet- of RDF-stores hebben en sommige hebben een implementatie die gericht is op subject, predikaat en object boven drie kolommen. Ze zetten alles in een triplet. Je kunt je wel voorstellen dat dat niet goed schaalt. Je krijgt dan een heel lange dunne tabel en bovendien is de knoop klein; het zou alleen een URI kunnen zijn, maar het kan ook een complete publicatie zijn, en het opnieuw opslaan van een hele publica-

SQL / RDBMS
- Concise, efficient transactions
- Transaction metadata is embedded or implicit in the application or database schema

XQuery/ XML
- Transaction across organizational boundaries
- XML wraps the metadata about the transaction around the data

SPARQL / RDF
- Information sharing with ultimate flexibility
- Enables semantics as well as syntax to be embedded in document

Voorbeeld van RDF

```
INSERT INTO family_rdf_data VALUES (29,
SDO_RDF_TRIPLE_S('family', 'http://www.example.
org/family/motherOf',
'http://www.w3.org/2000/01/rdf-schema#domain',
'http://www.example.org/family/Female'));
```

Zie verder: http://www.oracle.com/technology/tech/semantic_technologies/pdf/semantic_tech_rdf_wp.pdf

tie. Iedere keer als je er iets over wil zeggen, is dat dus heel slecht voor de schaalbaarheid. Wij slaan het gewoon één keer op en geven ieder item een unieke identifier en gebruiken die identifiers om het een veel schaalbaarder oplossing van te maken. Bovendien is de manier waarop we het opslaan een objectrelationele implementatie dus het heeft een relationele basis. We hebben een tabel die linktabel heet die de predikaat informatie restored. We zeggen waar de link begint en eindigt in de linktabel, dus de link representeert de complete triplet. Als we extra informatie over de triplet willen opslaan kunnen we dat op een soort relationele manier doen. We wijken min of meer af van de W3C-manier, maar ik houd erg van de manier waarop we het gedaan hebben.'

Oracle wijkt daarvan af, maar volgt wel de standaarden?
Stephens: 'Ja, ik begrijp dat dat een beetje vreemd klinkt. Volgens de W3C-specificatie zou je moeten zeggen: de eerste node waarover ik iets wil zeggen is deze node, de link waarover ik wil praten is deze link, de tweede node waarover ik wil praten is deze, en ten slotte is dit wat ik over alle drie wil zeggen. Om een ding over de triplet te zeggen moet je er vier aan toevoegen. We kunnen dat doen als je dat wilt, maar wat we met onze implementatie kunnen doen is dat we simpelweg kunnen dat iets van toepassing is op een zekere link, want de link representeert de complete triplet. We hebben dus een alternatieve methode. We moeten ook de RDF-data kunnen query'en binnen het RDF-datamodel, en we hebben SQL uitgebreid om dat mogelijk te maken. De query-mogelijkheid is ingevoerd samen met een aantal regels, die inferentie mogelijk maken. Ik vind zelf altijd een aardig voorbeeld, dat je een hoop ouder-kind relaties binnen het datamodel kunt opslaan. Je kunt bijvoorbeeld een regel schrijven die zegt dat een ouder van een ouder een grootouder is. Als die regel er eenmaal is, kun je die opslaan in de rulebase, daarna je rule toepassen op de originele dataset, en je krijgt alle grootouders. De informatie over kleinkind-relaties wordt dan opgeslagen in de regelindex, en vervolgens kun je die regelindex in je query opnemen, wanneer je iets van die informatie die je uitgekozen heb in de dataset wilt gebruiken.'

Het klinkt als een heel eenvoudige oplossing voor die situaties waarin 'gewone' SQL te gecompliceerd wordt, zeker voor de doorsnee ontwikkelaar die geen SQL-crack is.

Sterke groei

Er moeten toch ook veel mogelijkheden zijn om RDF verder te gebruiken, anders zou Oracle er niet in geïnteresseerd zijn?
Stephens: 'Ja, we geloven heel sterk dat het een technologie is die voor veel industrieën een grote belofte inhoudt. Kijk bijvoorbeeld naar de gezondheidszorg. Vaak worden patiënten doorverwezen van het ene ziekenhuis naar het andere ziekenhuis, maar ieder ziekenhuis heeft zijn eigen relationele model. Daardoor raakt het patiëntendossier verdeeld over verschillende ziekenhuizen en is het niet mogelijk om een compleet beeld te krijgen van het medisch dossier van die persoon. Wat ze nu gaan doen, is het bouwen van een ontologie, een soort schema dat boven de database zit. Vervolgens kun je de relevante informatie van de ziekenhuizen mappen in de ontologie die boven de database zit en dan kun je naar de totale data kijken. Dat werkt veel beter, omdat het zo moeilijk is om relationele modellen samen te voegen. RDF gaat helemaal over de mogelijkheid onafhankelijke databronnen samen te voegen die nooit opgezet zijn vanuit de bedoeling ze samen te voegen. Door de unieke identifiers kun je dat met RDF wel doen. Je geeft unieke indentifiers aan alles wat in de database zit. Daardoor wordt de database in plaats van alleen een lokale data-repository meer globaal. RFID is bijvoorbeeld een technologie die heel sterk groeit en die helemaal gebaseerd is op het geven van unieke identifiers. Dat is heel gemakkelijk te koppelen aan het RDF-model. Mensen beginnen het in andere situaties ook te gebruiken. Je zou het overal kunnen gebruiken waar je data moet integreren.'

Tekst en fotografie: **Dré de Man.**