Een gemeenschappelijke querytaal voor XML en relationele structuren

Loskoppeling van opslag-structuur en betekenis

Maurice Gittens

Waarom is de vraag of data door middel van relationele, XML of andere structuren is vastgelegd relevant, als men juist geïnteresseerd is in de *informatie* die in de betreffende structuren schuil gaat? Is het niet mogelijk om met één en dezelfde taal de raadpleging en manipulatie van de *informatie-inhoud* van een databron te faciliteren, terwijl de gebruiker van de interne structuur van de databron geen weet heeft?

En als dat al mogelijk is, in welke mate? En welke onderscheidende abstracties zouden kenmerkend voor zo'n taal kunnen zijn? Welke voordelen zouden voorhanden zijn bij de inzet van zulk een taal? Teruggrijpend op basisbeginselen uit de elementaire wiskunde, het relationele model, de referentiële betekenistheorie en de informatiemodellering gaat dit artikel introducerend in op deze en verwante vragen.

Informatie en haar toegankelijkheid

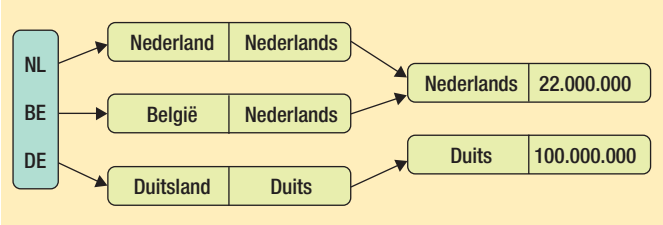
Veelal worden informatiesystemen gebouwd aan de hand van, soms zeer uitgebreide, informatiemodellen. Deze informatiemodellen geven in de regel een conceptueel beeld van de structuur van het gemodelleerde informatiedomein. Vereenvoudigd geldt dat bij de realisatie van een concreet informatiesysteem op basis van een informatiemodel keuzes gemaakt worden over *hoe* het gemodelleerde vorm zal krijgen binnen het te realiseren informatiesysteem. Concepten uit het informatiemodel worden geprojecteerd op abstracties uit oplossingsdomeinen. Abstracties uit de objectgeoriënteerde, documentcentrische en relationele benaderingen worden in dit verband, vaak in combinatie, ingezet als er invulling gegeven wordt aan de vraag *hoe* een informatiesysteem zal worden ingericht. De vraag *waarmee* een informatiesysteem ingericht zal worden, wordt beantwoord als een keuze gemaakt is voor wat tegenwoordig door sommigen de *technology stack* genoemd wordt.

Let wel, de beschouwing van de wat-, hoe- en waarmee-vragen met betrekking tot een informatiemodel geeft antwoorden die respectievelijk met verschillende doelgroepen te associëren zijn. De doelgroep die zinvol antwoord kan geven op wat-vragen is in het algemeen anders dan de doelgroep die dat kan op de hoe- en waarmee-vragen. Deze doelgroepen worden respectievelijk tot *wat-professionals, hoe-professionals* en *waarmee-professionals* gedoopt. Het is belangrijk om te beseffen dat deze drie groepen professionals, in het algemeen, ieder een eigen terminologie hanteren en vaak ook eigen soorten vragen stellen aan een informatiesysteem. Vanuit het perspectief van de business zijn de vragen van wat-professionals in het algemeen leidend. Vragen van wat-professional worden doorgaans via hoe-professionals met behulp van een waarmee-professional beantwoord.

> Informatiesystemen voor business professionals zijn vaak onnodig transparant

Door de verschillende vertaalslagen die in het geschetste traject plaatsvinden, worden vragen van wat-professionals, en zeker ad hoc vragen van deze professionals, in het algemeen niet op een kosteneffectieve wijze beantwoord. Vaak gestelde wat-vragen worden via hoe- en waarmee-hulpmiddelen doorgaans op vrij inflexibele wijze in functionaliteit van informatiesystemen omgezet. Hierdoor gaan aanpassingen vaak met hoge kosten en doorlooptijden gepaard. Een belangrijke reden waarom dit gebeurt, is maar al te vaak dat informatiesystemen voor business professionals, qua implementatie, onnodig transparant zijn. Dit wil bijvoorbeeld zeggen dat het te vaak gebeurt dat mensen die inzicht in de business hebben desondanks niet in staat zijn om dat inzicht in voldoende mate te laten renderen, omdat niet pertinente implementatiedetails zich op de voorgrond dringen. Om dit verschijnsel te verhelpen wordt in dit artikel voorgesteld om de *informatie* en de *significantie* ervan centraal te stellen. Wanneer informatie en haar significantie centraal staat, geldt dat de professional bij de interpretatie van informatie nimmer gehinderd mag worden door hoe of waarmee de informatie is opgeslagen.

Afbeelding 1: Voorbeeldfunctie.



Afbeelding 2: Samengestelde functie.

Wat hier gesteld wordt is niet nieuw. Het is een toepassing van het *orthogonaliteitsbeginsel*. De wat-vraag is orthogonaal ten opzichte van de hoe- en waarmee-vragen. De implementatie mag niet transparant zijn voor wat-professionals en de significantie des te meer. De professional die accepteert dat de opslagstructuur van informatie voor de betekenis van deze informatie niet wezenlijk is, heeft zich geëmancipeerd van een groot struikelblok.

XML en relationele opslag

Dat de opslagstructuur van informatie niet wezenlijk is voor de significantie van de informatie, is bijvoorbeeld in te zien wanneer men beschouwt dat het mogelijk is om voor een XML taal, zoals bijvoorbeeld XHTML, een informatiemodel te maken. Dit XHTML informatiemodel zou natuurlijk prima naar bijvoorbeeld een relationeel datamodel *geforward-engineerd* kunnen worden. In dit artikel zal door gebrek aan ruimte hierop niet verder worden ingegaan. Feit blijft wel dat XML opslagstructuren en ook relationele opslagstructuren in bestemde gevallen prima geschikt zijn voor de opslag van informatie. Het zou dan ook het mooist zijn als het niet transparant zou zijn dat informatie in een bepaald geval dan wel met XML of relationele structuren is opgeslagen. Wie de opslagstructuur van informatie wil loskoppelen van de betekenis van die informatie, zal een abstractie voor de betekenis van informatie dienen aan te wenden die geschikt is om de betekenis van informatie te belichamen, zonder dat daarbij de opslagstructuren van de informatie ter zake doen. Een wezenlijke vraag is daarbij welke abstractie voor dit doel het beste dienst zou kunnen doen.

Van de referentiële betekenis-theorie van Frege, via Tarski's modellen van de predikatenlogica en Chomsky's frasestructuur-grammatica's tot Codd's relationele structuren voor de opslag van gegevens en verschillende informatiemodelleringmethoden: telkens geldt dat er één en dezelfde fundamentele abstractie te identificeren valt. Deze abstractie, die dienst doet bij de belichaming van de significantie van informatie, is de notie van een functie. Een *wiskundige* functie om precies te zijn. Deze functies zijn onder verschillende namen bekend. Te denken valt aan termen als: associaties, relaties, mappings, objecten, klassen, koppelingen, transformaties, grammatica's, enzovoort. Wiskundige functies zijn in wezen vrij eenvoudige abstracties die we intuïtief en vaak ook formeel kennen. Deze functies zullen worden ingezet om talen te ontwerpen die geschikt zijn om databronnen te ontsluiten, terwijl de interne opslagstructuur van de betreffende databronnen niet relevant is.

Afbeelding 1 toont een voorbeeldfunctie, die een landcode koppelt aan een landnaam en de taal die primair in dat land gesproken wordt. Als *domein* van deze functie geldt de verzameling land-codes, terwijl als *bereik* van de functie het product van de landnaam, landcode en de taal van een land geldt. Dit voorbeeld is uit te breiden door op te merken dat de verzameling talen die in verschillende landen gesproken worden ook een interessante verzameling vormt. Afbeelding 2 toont een samengestelde functie die het mogelijk maakt om een landcode via een tussenfunctie te koppelen aan een verzameling talen die in verschillende landen gesproken zouden kunnen worden.

Relationele concepten als functies

Het is verhelderend om relationele concepten als relaties, *candidate keys* en *foreign keys* in het licht van wiskundige functies te zien. In het algemeen geldt bijvoorbeeld dat iedere relatie *tegelijkertijd* ook minimaal één wiskundige functie belichaamd. Dit is eenvoudig in te zien wanneer we beseffen dat iedere relatie *per definitie* minimaal één candidate key onderscheidt. De bovenstaande functie als relatie wordt getoond in afbeelding 3.

> Het is verhelderend om relationele concepten in het licht van wiskundige functies te zien

Omdat de landcode als unieke sleutel geldt, is het mogelijk om gegeven een landcode, de hieraan gekoppelde gegevens te herleiden. Dit is mogelijk juist omdat iedere relatie tegelijkertijd ook een functie vertegenwoordigt. De tekstuele weergave van een tuple in de 'landrelatie' is bijvoorbeeld:

```
land["nl"]
```

code	naam	taal
nl	Nederland	Nederlands
be	België	Nederlands
de	Duitsland	Duits

Afbeelding 3: Relationele functie.

naam	populatie
Nederlands	22.000.000
Duits	100.000.000

Afbeelding 4: Taalrelatie.

Interessant hierbij is om op te merken dat iedere tuple op zijn beurt ook als een wiskundige functie gezien kan worden. Het zal dan duidelijk zijn dat iedere tuple van een relatie een functie is die een attribuutnaam aan een attribuutwaarde koppelt. Dat deze kennis bij u op zijn minst latent aanwezig is, blijk uit het feit dat u zeer waarschijnlijk weet dat de expressie

```
land["nl"].taal
```

correspondeert met de waarde van de attribuut 'naam' van de tuple dat correspondeert met de landcode 'nl'. Ook een foreign key-relatie kan gezien worden als een wiskundige functie, waarvan de child-relatie het domein vormt terwijl de parent-relatie dienst doet als bereik van de functie.
Als een foreign-key 'taalref' tussen ons voorbeeld relatie 'land' en de in afbeelding 4 getoonde relatie 'taal' wordt gedefinieerd, geldt dat 'taalref' gezien mag worden als een functie die landcodes via de 'land'-relatie koppelt aan de 'taal'-relatie. Syntactisch blijkt deze koppeling bijvoorbeeld uit de expressie

```
land["nl"].taalref.populatie
```

die de waarde van de populatie van mensen die Nederlands spreken aanwijst.

XML structuren als wiskundige functies

XML structuren kunnen op hun beurt ook als wiskundige functies worden gezien. Dit is aannemelijk als men beseft dat de bovenstaande expressie afgezien van syntactische details in essentie een XPath expressie zou kunnen zijn. De details blijken in het geval van XML structuren, waarschijnlijk door de SGML historie, iets complexer in vergelijking met relationele structuren. Om deze reden wordt in dit artikel hier niet op ingegaan. Dat neemt niet weg dat ik het in dit artikel gepresenteerde betoog vorm heb gegeven in een hulpmiddel dat de ontsluiting van SQL-, CSV- en XML-databronnen met één en dezelfde syntax en semantiek faciliteert. De voordelen van deze aanpak lijken wel evident.

Een leuke eigenschap van wiskundige functies is dat deze tegelijkertijd ook volwaardige wiskundige relaties zijn. Dus de relationele algebra is zonder enige aanpassing toepasbaar op informatie die in termen van haar functionele significantie is uitgedrukt. Die complexe joins die in de relationele wereld zo gewaardeerd worden, kunnen bij de in dit artikel gepresenteerde benadering ook op XML-, CSV- en SQL-databronnen worden

toegepast. Het aardige is hierbij, dat het mogelijk is om gegevens uit drie verschillende soorten databronnen (XML, CSV, SQL), simultaan, in een join-expressie te incorporeren.

De loskoppeling van de betekenis van informatie en de opslagstructuur van informatie maakt de weg vrij voor nieuwe en verfrissende benaderingen om tot de ontsluiting van informatie te komen. Een aantal ontsluitingstechnieken waar ik onderzoek naar doe is bijvoorbeeld:

- 2D- en 3D-visuele representaties van betekenisstructuren voor informatie;
- Multi-linguïstische ontsluiting van databronnen;
- Auditieve ontsluiting van databronnen.

De loskoppeling van de opslagstructuur en betekenis is dan ook een eerste stap bij verbetering van het kennisrendement van wat-professionals.

Conclusies

De primaire stelling die in dit artikel is geponeerd, is dat het de aanbeveling verdient om de nadruk te leggen op de significantie van informatie, losgekoppeld van specifieke opslagstructuren voor informatie. De doelgroep die zijn kennis en expertise kan laten renderen is bij deze aanpak groter dan wanneer er onnodige segregatie plaatsvindt op basis van de hoe- en waarmee-kennis.

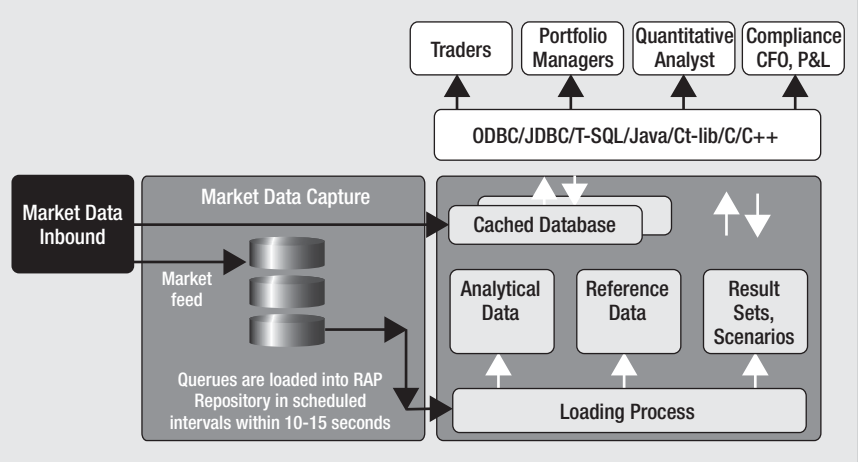
Maurice Gittens is zelfstandig IT-consultant.

Update

MiFID, het nieuwe fenomeen in de financiële wereld

Vanaf 1 november 2007 treedt een nieuwe Europese richtlijn in werking, Markets in Financial Instruments Directive, kortweg MiFID. De richtlijn heeft vooral betrekking en invloed op de manier waarop in aandelen en effecten gehandeld wordt, zowel aan de koop- als aan de verkoopkant, zowel bij banken als op de beurzen. Meer transparantie en meer concurrentie zijn hierbij kernwoorden.
Wat is de impact? Gavin Quinn, European Marketing Manager Financial Services Industry bij Sybase, zegt daarover: "Financiële instellingen moeten veel meer informatie langer bewaren; bovendien moet die informatie snel kunnen worden teruggevonden en gepresenteerd. Belangrijk: het gaat hierbij niet alleen om transactiegegevens, maar ook om achtergrondinformatie over de omstandigheden waaronder de transacties plaatsvonden. Marktgegevens dus, pre- en post-trade analyses, aan wie wat wanneer gerapporteerd is en wie wanneer welke beslissing heeft genomen op basis van welke gegevens. Zet dat in het licht van groeiende datavolumes, meer typen en soorten effecten, meer valuta's, meer derivaten en meer databronnen, tegen een achtergrond van snellere markten, meer geautomatiseerde handel, Algo/Quant, en je hebt een problematiek die moeilijk te bevatten is. MiFID verandert de dynamiek van de branche volledig."

Volgens Quinn verandert de hele zakelijke huishouding door MiFID, door minder grote spreiding en lagere commissies. Meer handel zal elektronisch geschieden zonder tussenkomst van de verkoperszijde. "Risk Management, daar gaat het om. Compliance wordt daar de aanjager van. Basel II dwingt bedrijven hun traditionele risicobepaling te herdefiniëren, en een operationeel Risk Framework te ontwikkelen." Om verder te gaan: "Sybase zag dat al aankomen en ontwikkelde een, laten we zeggen, 'time-series database', het Risk Analytics Platform (RAP). Het probleem ligt immers niet zozeer bij de opslag van deze enorme hoeveelheden data; de rapportage-eisen van MiFID en effectieve risico-analyse eisen dat de data snel inzichtelijk kunnen worden gemaakt. RAP is bij uitstek geschikt om de problemen die ontstaan door MiFID het hoofd te bieden. RAP reduceert het informatie- en analyseproces door snellere processing van de in-bound time-series data en consolidatie van actuele en enorme hoeveelheden historische gegevens in een continu updatende datastore."
RAP is eigenlijk een combinatie van Sybase's ASE en IQ. In het ASE gedeelte vindt de I/O en OLTP plaats, in IQ geschiedt de datacompressie en OLAP. Opslag en analyse in één en dezelfde omgeving dus, tot wel 2 Petabyte. "De Security Exchange Commission en Barclay's Global Investors hebben RAP al draaien", besluit Gavin Quinn. "En er zullen er nog vele instellingen volgen."



Afbeelding 1: Sybase Risk Analytics Platform.