

Informatica neemt Itemfield over

Power Exchange for Complex Data

Hans Lamboo

Onlangs nam data-integrator Informatica het bedrijf Itemfield over. Hiermee verkrijgt Informatica datatransformatie-technologie voor uitwisseling van en toegang tot ongestructureerde en semi-gestructureerde data.

Sinds Informatica een aantal jaren geleden stopte met het aanbieden van BI aan het front-end, zoals tools voor rapportage en analyse, is de focus geheel gericht op de integratie van data in de breedste zin van het woord. Dus niet alleen data uit relationele databases, maar ook flat files, XML-berichten, messages van Tibco, WebSphere MQ, SAP NetWeaver, EDI, enzovoort. De blik

richt zich sinds enige tijd ook op ongestructureerde, semi-gestructureerde en geografische data. Onder semi-gestructureerde data verstaat Informatica de standaards, vaak in XML vastgelegd, zoals die bijvoorbeeld gebruikt worden bij de banken (SWIFT), in de gezondheidszorg (HL7) en de verzekeraars (Acord); geografische data komen bijvoorbeeld uit applicaties als ESRI, Bentley en MicroStation.

Er is vanuit organisaties steeds vaker vraag naar multi-format data-integratie. Een van de problemen daarbij is dat sommige 'standaards' helemaal niet zo standaard blijken te zijn. Zo'n 'standaard' wordt vaak handmatig geprogrammeerd en er bestaan dus tientallen versies van. Komt er een nieuwe versie uit, dan moeten meestal alle programma's worden aangepast. Ook daarbij wil Informatica een rol spelen.

Verizon

Het telecombedrijf Verizon gebruikt de Itemfield-technologie om uit de financiële administratie gegevens op het web te kunnen publiceren. In hun mainframe-systeem vindt alle billing plaats van het gebruik van mobiele netwerken, en daar komen papieren nota's uit. Als extra service aan de klanten, wilde Verizon de klanten hun nota's, maar ook extra informatie zoals bijvoorbeeld het gebruik van de laatste maanden, op het web aanbieden. Verizon kon daarvoor een applicatie bouwen die de data uit het source-systeem gaat aggregeren in een datawarehouse, welke data naar de website worden gesluisd. Maar men realiseerde zich dat op de papieren nota eigenlijk alle informatie staat die daarvoor nodig is. Bovendien zijn die nota's allemaal gecontroleerd, komen uit een secure back-office systeem; de business rules en dergelijke zijn allemaal al toegepast in het printprogramma, en de nota's bevatten dus de juiste informatie. Met behulp van Itemfield-technologie zijn de verschillende velden gedefinieerd: het print-formaat ziet er zo uit, daar staan de gebelde minuten, daar staat dit gegeven en daar dat. Dat document is 'geleerd' aan Itemfield. Bij het mainframe staat een printstraat, waar alle nota's die uit het mainframe komen eerst door Itemfield heengaan, de data worden omgezet in een ander XML-formaat dat gepresenteerd kan worden aan de end-users, en daarna worden de nota's per post verstuurd aan de klanten. Met deze oplossing kostte het dus veel minder tijd en geld om de beoogde extra service te realiseren, dan dat daarvoor een nieuwe applicatie had moeten worden gebouwd.

XML transformatie-engine

Itemfield is een Israëliisch bedrijf; de afdeling Research & Development bevindt zich in Tel Aviv. Zo'n twee jaar geleden sloot Informatica een OEM-overeenkomst met Itemfield, en bood de software binnen PowerCenter aan als optie, de Unstructured Data Option, voor de handling van ongestructureerde en semi-gestructureerde data. Omdat Informatica de data-integratiemarkt wil blijven domineren was het onvermijdelijk om het in de eigen 'stack' te hebben; bovendien bestond het gevaar dat een andere partij Itemfield zou overnemen, waardoor zelfs de OEM onder druk zou komen. En dus werd Itemfield ingelijfd.

De software van Itemfield omvat een grote hoeveelheid voor-gedefinieerde library's van standaards zoals Acord, SWIFT en HL7 en zo nog wel honderd verschillende branche-standaards. Bij een aanpassing van de standaard is binnen 48 uur de nieuwe library-versie beschikbaar voor de klant, die daarmee snel zelf de transformatie kan regelen van oude bestanden naar de nieuwe standaard.

Tot de Intellectual Property van Itemfield behoort ook software die kan 'leren' nieuwe structuren te maken. De klanten van Itemfield gebruiken het voor transformaties van objecten die niet in een relationele database zitten, bijvoorbeeld afkomstig uit medische systemen. De software kan geleerd worden wat de structuur van de data is, wat er mee gedaan moet worden en ze naar een database sturen. Het hart van de Itemfield-technologie is een hele slimme XML-transformatie engine. Want wat er eigenlijk gebeurt is: eerst het definiëren, dan de transformatie naar een XML-

schema, dan het lezen en vergaren van data, dan het omzetten naar andere XML-toepassingen of het sturen van de data naar een database of datawarehouse.

Toepassingen

Een voorbeeld: een bedrijf ontvangt elke maand via de e-mail facturen in PDF-formaat. Itemfield 'leest' de PDF door te zoeken naar tevoren door de gebruiker gedefinieerde velden: achter het woord 'klantnummer' is het eerste getal het klantnummer, onder het woord 'order' staan de orderregels, etcetera. De gebruiker 'leert' de software de PDF's dus te lezen aan de hand van een voorbeeld. De zo verkregen data worden omgezet in een XML-schema, dat vervolgens gestuurd kan worden naar bijvoorbeeld het datawarehouse, operationeel systeem, of op een MQ-bus gezet. Kortom: via de e-mail als attachments binnenkomende facturen worden automatisch verwerkt en de data worden geïntegreerd in andere systemen.

Wat Itemfield met een PDF kan, geldt ook voor bijvoorbeeld spreadsheets. Bedrijven die leven bij de gratie van spreadsheets zien zich geconfronteerd met wet- en regelgeving op het gebied van compliance, die meer grip vereisen en dus betere mogelijkheden voor auditing en tracing. Die spreadsheets zouden daartoe moeten worden gelezen, geïmporteerd en vastgelegd in een database-structuur. De software van Itemfield maakt dat mogelijk. Een andere toepassingsmogelijkheid is het consolideren in een datawarehouse van diverse spreadsheets, bijvoorbeeld vanuit verschillende divisies, in één corporate spreadsheet. Je kunt het ook gebruiken als elke divisie binnen een bedrijf een spreadsheet aanlevert, om dat gemakkelijk te consolideren in een datawarehouse of in een corporate spreadsheet. Nog een aardig toepassingsvoorbeeld: een bedrijf wil de websites van de concurrentie in de gaten houden. De gebruiker bezoekt de websites van de betreffende concurrenten en definieert zekere velden van de sites in de Itemfield-technologie, die vervolgens wijzigingen in deze velden registreert en ter analyse kan opslaan in een database.

Op dit moment kan de software nog geen écht ongestructureerde data, zoals e-mails en documenten, verwerken. Er wordt wel gewerkt aan verbreding van de toepassingsmogelijkheden. Maar Informatica is zeker niet van plan zelf alle nieuwe functionaliteit te ontwikkelen, maar zoekt wel technologieën die geïntegreerd kunnen worden of bedrijven waarmee partnerships kunnen worden aangegaan – bijvoorbeeld integratie met search-technologie, met business rules engines en met data mining producten. Een voorbeeld daarvan is een telecombedrijf dat Informatica's PowerCenter gebruikt voor alle data-integratie. De telco gebruikt voor de search op gestructureerde en ongestructureerde data de technologie van de Fast-engine. Die kan naast de 'zittende' data in databases ook de data zien die door PowerCenter heengaan en ze vervolgens indexeren. Daarmee is het niet meer nodig te weten hoe de kubus in elkaar zit, de eindgebruiker geeft een paar kernwoorden en krijgt dan een zoekresultaat. Sommige hits



Bert Oosterhof: "De combinatie Informatica plus Itemfield biedt uitkomst".

komen misschien uit een Oracle database, andere komen uit documenten. Als de gebruiker op hits uit de database klikt, wordt vanzelfsprekend geen document getoond, maar wordt er een query afgevuurd die de data ophaalt. Met dit soort technologieën wil Informatica graag nauw samenwerken. Men is zeker niet van plan zelf een nieuwe search-engine te gaan ontwikkelen. "Onze architectuur is zodanig open, dat elk data-format door de PowerCenter-engine verwerkt kan worden", zegt Bert Oosterhof, Technisch Directeur bij Informatica Europa. "In de logica definieer je de afkomst en het doel van de data; daartussen heb je de transformatieregels. Die kunnen ook door derden gemaakt worden. Een van onze telecom-klienten heeft op deze manier PowerCenter met search-technologie van Fast Search geïntegreerd, waardoor de zoekmachine niet alleen documenten en webpagina's kan terugvinden maar ook data die door PowerCenter zijn verwerkt."

Oosterhof vervolgt: "Voor Informatica is de overname van Itemfield een grote stap voorwaarts. Om onze sterke marktpositie te behouden moet het aanbod almaar breder worden, moeten we steeds meer kwaliteit bieden, moet de software sneller worden, grotere datavolumes en meer datatypes aankunnen en daar dan ook weer méér functionaliteiten op los kunnen laten. Daar zit onze groei. En belangrijk is: onze onafhankelijkheid in de markt behouden."

Migratie-optie uitgebreid

Voor wie Itemfield nog niet kent: veel ziekenhuizen in Nederland gebruiken de software als onderdeel van G2SPEECH-oplossin-

gen. Ook Philips Medical Systems gebruikt de technologie. Daarnaast zijn er klanten, die SAP's Conversion Agent gebruiken, zoals Ziekenhuis Amstelland. Deze Conversion Agent is een OEM-versie van Itemfield. Voor de verwerking van ongestructureerde data worden nog veel meer partnerships gesloten. Informatica blijft daar als onafhankelijke dataservices-leverancier tussen zitten.

Bert Oosterhof: "We zullen zowel de klanten van Itemfield PowerCenter gaan aanbieden als andersom. Ongeveer de helft van onze klanten gebruikt onze data-integratie- en datakwaliteit-oplossingen voor het laden en bijwerken van het datawarehouse, de andere helft gebruikt Informatica voor datamigratie en andere projecten waarbij data verplaatst moeten worden. In datamigratieprojecten zit ook altijd een stukje ongestructureerde data. Er staat misschien een oud mainframe-systeem, en omdat niet alle gegevens daarin zitten, gebruikt het bedrijf nog wat spreadsheets en Access. Nu is Siebel of SAP aangekocht en moet er gemigreerd gaan worden – dan praat je dus over een verzameling ongestructureerde data en mainframe-data, die allemaal getransformeerd moeten worden. De combinatie Informatica plus Itemfield biedt daarbij uitkomst. Er zullen ook Itemfield-klanten zijn die misschien in het verleden te klein waren voor Informatica-producten of die puur naar een oplossing van een probleem met ongestructureerde data gezocht hebben. Deze bedrijven hebben ongetwijfeld meer data-integratie

issues, die ze met Informatica in dezelfde laag met Itemfield kunnen stoppen – en dus geen verschillende producten hebben voor hetzelfde probleem."

Tenslotte

Inmiddels zijn de R&D-afdelingen van Itemfield en Informatica organisatorisch al geïntegreerd – fysiek niet, maar ze rapporteren aan dezelfde manager. R&D van Itemfield kan zich nu helemaal toeleggen op dataformaten en transformaties op die dataformaten. Alle access en connectiviteit die Informatica al heeft, kan nu ook door hen worden gebruikt. Dat heeft weer voordelen voor de toepasbaarheid van Itemfield-technologie. De combinatie met Informatica maakt het bijvoorbeeld mogelijk van een ongestructureerd object te controleren of de gegevens al voorkomen in een mainframe-database. Het ongestructureerde object wordt 'gelezen' en tegelijkertijd wordt het mainframe geraadpleegd, de uitkomsten worden gecombineerd, of gemapped, en zie daar. "Dat kon Itemfield vroeger ook wel met een ODBC-connectie naar een database, maar nu hebben ze connectie naar alles waarop Informatica kan connecteren", legt Bert Oosterhof uit. Om te besluiten: "Onze oude OEM-optie, de Unstructured Data Option, gaat nu Power Exchange for Complex Data heten. Dat kan stand-alone, maar is al geïntegreerd in PowerCenter als optie. En dat blijft zo."

Hans Lambou is hoofdredacteur van Database Magazine.

Ambitieuze BI & CMS specialisten zoeken nieuwe collega's met passie voor hun vak



Ambitieuw

Onze medewerkers zijn ons grootste onderscheidend vermogen: ambitieuze (tool)specialisten met een afspraak = afspraak mentaliteit en passie voor hun vak. We hebben de ambitie om ons team op korte termijn met een aantal specialisten uit te breiden. Ben jij de collega die we zoeken?

Passie

VLC is een ICT dienstverlener met passie voor Business Intelligence en Content Management. Twee dynamische vakgebieden die steeds meer naar elkaar toegroeien. We volgen de ontwikkelingen op de voet en stoppen veel tijd en energie in het volgen van cursussen, het bezoeken van seminars, het lezen van vakliteratuur etc. Op deze manier zorgen we ervoor dat onze specialisten specialist blijven.

VLC zoekt

Business Intelligence & Content Management specialisten, met een afgeronde HBO/academische opleiding, goede communicatieve vaardigheden en minimaal 2 jaar ervaring met een of meer van de volgende BI tools: WebFOCUS, Business Objects, Cognos, PowerCenter, Oracle Warehouse Builder, SAS of CMS tools: Tridion, GX WebManager.

VLC biedt

Ambitieuze collega's, passie voor het vakgebied, uitdagende opdrachten, persoonlijke ontwikkeling & uitstekende arbeidsvoorwaarden.

Interesse?

Stuur dan een reactie naar mathijs.kreugel@vlc.nl of kijk op www.vlc.nl.

