Informatica met PowerCenter 8 in Gartner's leiderskwadrant

Data-integratiedrieluik met SOA en ICC

Paul van der Linden

Vorig jaar werd door CEO Sohaib Abbasi tijdens Informatica World vooral de boodschap verkondigd dat Informatica niet meer gelijkstaat met ETL (extractie, transformatie en laden van data). Het nieuwe Informatica staat voor data-integratie.

Tijdens de Informatica World conferentie die vorig jaar juni in Washington werd gehouden beloofde ceo Sohaib Abbasi al een volgende versie van PowerCenter die een 10 keer betere performance en een 10 keer betere productiviteit levert dan de op dat moment actuele versie 7.1[1]. Die volgende versie, toen alleen nog bekend onder de codenaam project Zeus, is inmiddels beschikbaar als PowerCenter 8 en is bedoeld om zogenaamde 'mission critical enterprise deployments' aan te kunnen. Voor de herfst van 2006 staat weer een volgende versie gepland (project Hercules) die uniforme toegang tot alle data belooft (on demand data integration).

Van ETL tot data-integratie

Nog niet eens zo heel lang geleden was Informatica (samen met Ascential) de onbetwiste leider in ETL-producten. ETL staat voor het extraheren, transformeren en laden van data afkomstig uit verschillende bronnen (transactiesystemen) naar het datawarehouse. De ETL-productcategorie is dan ook nauw gelieerd aan datawarehousing. Het mag dan ook niet verbazen dat de ontwikkelingen in ETL voor een groot deel worden ingegeven door ontwikkelingen in Business Intelligence en datawarehousing.

> Het is nu mogelijk om een sessie dynamisch (at runtime) te partitioneren

De eerste golf van ETL-tools, aangevoerd door Carlton, ETI en Prism, bestond uit zogenaamde codegeneratoren. Dicht tegen de databronnen aan werden programma's uitgevoerd die uit alle hoeken van het applicatielandschap relevante data moesten extraheren om die vervolgens betekenisvol te combineren en in het datawarehouse klaar te zetten. Nadeel van deze zogenaamde eerste generatie ETL-tools was de verspreiding van al de ETL-softwareprogramma's ('software everywhere'). Als reactie hierop ontstond een nieuwe tweede golf van ETL-software die de ETL-processen vanuit een centrale plek (de transformation engine) uitvoerde. Informatica en voormalig ceo Gaurav Dhilon behoren tot de pioniers van deze tweede golf. De resulterende leidende positie in ETL heeft Informatica niet meer opgegeven. Het stelde hen in staat om een forse prijs te hanteren voor een ETL-product dat top-of-class is en dan ook door menige organisatie werd gekocht en ingezet.

Een van de belangrijkste actuele ontwikkelingen in BI en datawarehousing is de toegenomen snelheid waarmee data beschikbaar moeten zijn. Dat geldt ook voor data die door het datawarehouse worden geleverd. In tegenstelling tot vroeger wordt er geen clementie meer verleend, omdat we het hebben over grote hoeveelheden data, complexe bewerkingen en analyses en verschillende bronsystemen. Deze beweging naar near real-time en real-time data zal niet in alle organisaties in dezelfde mate spelen, maar de ontwikkeling is duidelijk. Gezien de ontwikkelingen in de markten waarin organisaties opereren is er een groeiende behoefte aan het sneller kunnen beschikken over adequate data. Deze ontwikkeling heeft de bestaande ETL-markt onder druk gezet. Immers, hier werden grote hoeveelheden data batchgewijs (periodiek) overgezet naar het datawarehouse. De regelmaat waarmee dit gebeurde kon best maandelijks of wekelijks zijn. Dagelijks of het real-time overzetten van data was in deze datawarehouses niet aan de orde. De tweede ontwikkeling die de bestaande ETL-markt onder druk zette, was de beschikbaarheid van goedkope ETL-tools die niet alle, maar wel de meeste functionaliteiten boden tegen een fractie van de prijs. Deze ETL-tools zijn geschikt voor de meeste organisaties. Denk hierbij bijvoorbeeld aan Data Transformation Services (DTS)[2], de ETL-tool die door Microsoft werd meegeleverd met hun SQL Server database. Maar ook Oracle's Warehouse Builder (OWB) wordt meegeleverd met de database en vormt in veel gevallen een voldoende goede ETL-oplossing.

Geen wonder dus dat de ETL-leveranciers *pur sang* zich genoodzaakt voelden om hun positie te heroverwegen. Aangezien prijsconcurrentie met partijen als Microsoft geen (plezierige) optie is was het te volgen pad duidelijk: toevoegen van extra functionaliteit. In dit geval het toevoegen van real-time data-integratie

aan de reeds aanwezige batchfunctionaliteit. Hiermee ontstond het 'nieuwe' veld van data-integratie. Dit is ook het speelveld waarnaar Informatie is verhuist. Uiteraard biedt het nog steeds een prima ETL-oplossing, maar nu is het slechts een van de dingen die ze te bieden heeft. De totale oplossing die met PowerCenter 8 wordt geboden heet: data-integratie. PowerCenter is beschikbaar als Standard Edition en als Advanced Edition. Het verschil zit hem in de beschikbaarheid van PowerAnalyzer (BI), SuperGlue (metadata management) en team-based development en server grid.

Componenten van data-integratie

Wat zijn nu de componenten die tot data-integratie kunnen worden gerekend?

- Eenduidige kijk op de organisatie ('one version of the truth'): centraal staat een eenduidige kijk op de organisatie en haar resultaten. Om zo'n 'single view' of 'single version of the truth' te verkrijgen wordt meestal een datawarehouseproject gestart;
- Financiële consolidatie: bij elkaar brengen en combineren van financiële gegevens, bijvoorbeeld van werkmaatschappijen naar een centrale holding;
- Masterdata management: erop gericht om consistentie te verkrijgen in deze beschrijvende data. Denk bij masterdata bijvoorbeeld aan een geografische indeling, kalender of productcategorieën;
- Legacy-migratie: het eenmalig overzetten van data die zich bevinden op een legacy-platform naar een ander platform;
- Datasynchronisatie: doorzetten van (wijzigingen van) data tussen verschillende systemen. Een voorbeeld hiervan is Enterprise Application Integration (EAI).

The Data Warehousing Institute (TDWI) heeft eind 2002 de eisen benoemd waaraan volgens hen een zogenaamd data-integratie-platform dient te voldoen. Deze eisen vallen uiteen in eisen die gelden voor het platform en eisen die men stelt aan de data-integratie.

Tot de platformaspecten rekent TDWI de volgende punten: goede performance en schaalbaarheid; ingebouwde data cleansing en profiling; complexe, herbruikbare transformaties beschikbaar; betrouwbare operatie en robuuste administratie. Tot de data-integratie-aspecten: diverse bronnen en doelsystemen; update- en capture-faciliteiten; near real-time processing; global metadata management.

De meeste van deze aspecten zullen geen uitleg nodig hebben. Enkele vragen echter om nadere uitleg. Zo betekent data profiling dat alle mogelijke waarden en formaten van velden en kolommen, alsmede de afhankelijkheden tussen bestanden in kaart kunnen worden gebracht. Deze worden vervolgens gebruikt als een soort steen van Rosetta. Data profiling verschaft hiermee een goed beeld van de data die feitelijk in de bronsystemen zitten. Onder het kunnen omgaan met verschillende bronnen en doelen wordt ook begrepen web services en XML (inmiddels beide gangbare formaten). Bij globaal metadata management wordt het Common Warehouse Metamodel (CWM) van de OMG genoemd. Ook hier geldt dat CWM de afgelopen jaren aan populariteit heeft gewonnen.

Nieuwe functionaliteit in PC8

Tot de nieuwe functionaliteiten (vergeleken met versie 7.1) behoren de volgende opties.

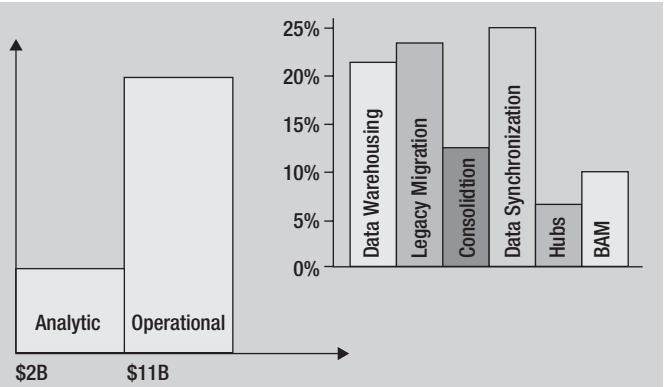
Data federation optie. Dit is de mogelijkheid om een virtuele datafederatie of Enterprise Information Integration (EII) toe te passen. Dit houdt in dat een virtuele dataview kan worden gegenereerd op basis van verschillende operationele bronnen zonder dat daarbij de data fysiek wordt verplaatst. Hierbij wordt gebruik gemaakt van data caching. Het grote voordeel van deze aanpak is dat op een snelle manier, zonder veel extra kosten en zonder extra complexiteit te creëren de vereiste informatie kan worden geleverd.

Unstructured data optie. Een add-on die het mogelijk maakt om ongestructureerde data (documenten, e-mails, binaire files, Excel etcetera) en bepaalde semi-gestructureerde data (bijvoorbeeld HIPAA, ACORD, FIX, SWIFT) te benaderen en beschikbaar te maken. Hierbij hoeft niet geprogrammeerd te worden. Zoals bekend zal zijn bestaat 80 procent van alle data in ongestructureerde vorm.

High availability optie. Deze draagt zorg voor een hoge beschikbaarheid (uptime) van de omgeving. Dit wordt bereikt door onder andere naadloze fail-over en herstel van gestopt/interrupted werk; eenvoudige setup en management door gebruik van een configureerbaar service framework; service monitoring voor onder andere het opsporen van fouten.

Enterprise Grid optie. Dit is de mogelijkheid om een sessie te verdelen over meerdere nodes (servers) in een grid ('Session on Grid'). Hierdoor ontstaat de mogelijkheid om bedrijfsbrede, mission-critical data-integratie in te zetten. Onderdeel van de Enterprise Grid optie is de eerder beschreven High availability optie en de Session on Grid-mogelijkheden.

Pushdown Optimization optie. Dit betreft het verbeteren van de sessieperformance door het verschuiven van de transformatie-logica naar de database. Het uitvoeren van de SQL kan gebeuren



Afbeelding 1: Data-integratie projecten.

door de brondatabase, doeldatabase of in de integratieservices. PowerCenter kan op basis van de gedefinieerde mapping en sessieconfiguratie zelf hier een optimale keuze maken. Ook het afhandelen van flat files is verder verbeterd.

Verbeteringen in PC8 betreffen de volgende punten.
Data cleanse and match optie (voorheen de Data Cleansing optie). Hiervan maken deel uit: data parsing (ook van ongestructureerde data), datastandaardisatie, adres/naam-verbeteringen met behulp van geografische en zipcode-informatie, matching met behulp van fuzzy logic of business rules, record merging (recordconsolidatie) en business data parsing.
Data profiling optie. In vorige versies ging het met name om cijfers over data: statistieken over kolommen, redundantie, uniekheid en businessregels. Hier zijn nu bijgekomen: bepalen van de functionele afhankelijkheden tussen de verschillende kolommen in een bronsysteem; multi-column sleutelanalyse en interferentie tussen verschillende bronnen; column look-up domain en multiple join columns.

> Het ICC moet dan ook gezien worden als een hands-on club

Partitioning optie. Het is nu mogelijk om een sessie dynamisch (at runtime) te partitioneren op basis van het aantal partities, het aantal nodes in een grid en de partitionering van de bron. Voor Oracle en DB2-bronnen geldt dat PowerCenter zijn partitioneringsschema kan afstemmen op de partitionering van de bron teneinde performancewinst te bepalen.
Team-based development. Dit is alleen beschikbaar in PC Advanced Edition. Betreft de mogelijkheid om een expliciete check-out te doen en de mogelijkheid om oudere versies van objecten te zien.

Markt

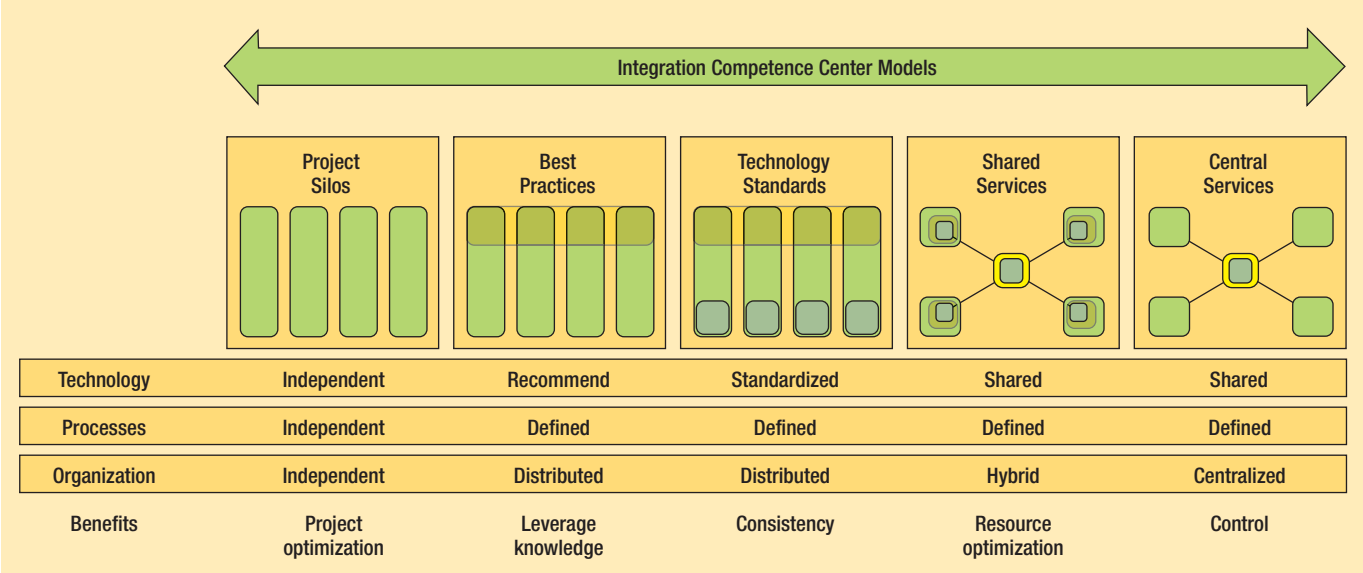
Uit het Nationaal Data Warehouse Onderzoek (DWO 2005)[3] dat vorig jaar door Array Publications en Atos Origin werd uitgevoerd, bleek dat deelnemende bedrijven in 12 procent van de gevallen niet over een ETL-beschikken. Een percentage van 68 beschikt wel over een ETL-tool en 20 procent van de organisaties zelfs over twee ETL-tools of meer. Deze cijfers komen grotendeels overeen met de bevindingen van The Data Warehousing Institute (TDWI) die voor de VS uitkomen op 18 procent van de organisaties die handmatig ETL-processen schrijven en 82 procent van de organisaties die hiervoor een ETL-tool inzetten[4].
Informatica is – niet verrassend – de meest gebruikte ETL-tool in Nederland. Van de deelnemers aan het DWO 2005 geeft 31 procent aan dat Informatica wordt ingezet. Oracle Warehouse Builder (OWB) volgt met 22 procent, met daarachter een peloton van Microsoft, SAS, Business Objects, Cognos en Ascential (nu IBM WebSphere). Deze indeling komt redelijk overeen met hetgeen Gartner laat zien in het ETL Magic Quadrant[5]. Alleen de positie van Ascential, inmiddels overgenomen door IBM, wijkt af van de Nederlandse plaatsing. In het Magic Quadrant is Ascential namelijk de absolute leider. Op ability to execute zit Informatica op één lijn; wat visionaire inzicht moet men Ascential voor laten gaan.
Dat data-integratie steeds belangrijker is geworden staat buiten kijf. Een belangrijke drijfveer hierbij is dataconsistentie. Wet- en regelgeving als Sarbanes-Oxley, Basel II en IFRS hebben het belang van consistente data alleen maar vergroot. Maar ook de noodzaak om steeds sneller (de juiste) data op meerdere plaatsen beschikbaar te hebben, operationele systemen te synchroniseren en real-time te monitoren zijn even zovele redenen waarom data-integratie hoog op de agenda staat.

Integration Competence Center

Tijdens de BI Summit die Gartner begin februari in London hield was de belangstelling voor competence centers (CC's) overweldigend. De meeste organisaties beschikken inmiddels over een competence center. Een competence center is het antwoord op de al jaren geleden geconstateerde kloof tussen de vraag naar gekwalificeerde IT-krachten en het aanbod ervan (de 'competence gap'). Ook als het gaat om integratie wordt in een Integration Competence Center (ICC) de oplossing gezien. Waar de eerste ICC's voornamelijk bestaan uit personen met een achtergrond in applicatie-integratie (Enterprise Application Integration) kan geconstateerd worden dat meer recente ICC's ook data-integratie-specialisten, database-experts en deskundigen op het terrein van datakwaliteit herbergen. Een ICC is de verzameling van al deze experts binnen de organisatie. Door ze als competence center te organiseren wordt maximaal resultaat verkregen doordat kennis en ervaring kan worden uitgewisseld en best practices ontstaan. Het ICC moet dan ook gezien worden als een hands-on club die betrokken is bij de verschillende projecten binnen de organisatie waarin integratie een rol speelt. Ook Informatica heeft een duidelijke mening over ICC's en hanteert een model bestaande uit vijf fasen.

Data-integratie en SOA

Een van de interessante kruispunten is die van data-integratie en SOA. Het begrip Service Oriented Architecture is al een aantal jaren oud. Een SOA gaat uit van software services die zich binnen, maar ook buiten de organisatie kunnen bevinden en op het moment dat daar behoefte aan is kunnen worden aangeroepen. Hiertoe is het nodig dat duidelijk is welke services beschikbaar zijn en dat via een broker een contract kan worden afgesloten onder welke voorwaarden gebruik kan worden gemaakt van beschikbare services. Protocollen zoals XML (Extensible Markup Language), SOAP (Simple Object Access Protocol) en WSDL (Web Services Desciption Language) hebben de implementatie van SOA's mogelijk gemaakt. Volgens een onderzoek dat AMR Research in 20056 uitvoerde beschikt momenteel 21 procent



Afbeelding 2: Integratie van het Competence Center Model.

van de organisaties over SOA. Gartner verwacht dat de meeste organisaties binnen vijf jaar een SOA hebben.
In de eerste SOA-projecten werd met name gebruik gemaakt van EAI software, messaging middleware, enterprise service bussen en op J2EE en .NET gebaseerde middleware. De nadruk lag hierbij met name op applicaties en het losweken van de business logica zodat het als aparte services kon worden aangeboden. Door de manier waarop data-integratie in de meeste organisaties is geïmplementeerd (door zogenaamde point-to-point solutions) is het alleen maar een kwestie van tijd dat SOA ook hier wordt toegepast. Wat een goede data-integratie-oplossing biedt (data-kwaliteit, dataconsistentie, data governance, semantiek/metadata, toegang tot alle soorten data, messaging en bulkdataverwerking) zal in de vorm van services worden aangeboden om een maximale herbruikbaarheid en helderheid met minimaal TCO op te leveren. Kortom, de applicatiecentrische SOA-wereld en de ontluikende data-integratiewereld zitten op een koers die zal leiden tot wederzijdse versterking. SOA zal een data-integratiebasis krijgen onder de applicatieservices terwijl dat data-integratie via services zal worden aangeboden. PowerCenter 8 wordt door Informatica dan ook gepositioneerd als het data-integratieplatform dat dataservices levert aan een service oriented architecture. De derde bouwsteen om te komen tot best practice data-integratie is het eerder beschreven Integration Competence Center. Zie hier het drieluik dat data-integratie heet.

Conclusie

Vorig jaar werd door CEO Sohaib Abbasi tijdens Informatica World vooral de boodschap verkondigd dat Informatica niet meer gelijkstaat met ETL (extractie, transformatie en laden van data). Het nieuwe Informatica staat voor data-integratie. ETL is daar slechts een van de componenten van. Financiële consolidatie, masterdata management, datasynchronisatie en legacy-migratie zijn enkele van de andere componenten. In een wereld die gekenmerkt wordt door outsourcing en verdere datafragmentatie lijkt het voor Informatica altijd feest.
Niet dat Informatica op zijn lauweren rust. Met PowerCenter 8 is onder andere nieuwe functionaliteit op het gebied van ongestructureerde data, datafederatie en hogere beschikbaarheid van data toegevoegd. Ook belangrijke verbeteringen in data cleansing en matching, data profiling en partitionering zijn doorgevoerd.
Minstens even belangrijk is de visie op data-integratie die Informatica met PowerCenter 8 neerlegt. PowerCenter 8 sluit aan bij de opkomende Service Oriented Architectures en zorgt voor een completere invulling daarvan. Door inzet van een ICC, een best practice benadering, zijn de voorwaarden voor een succesvolle data-integratie-oplossing aanwezig. Hiermee biedt Informatica zowel een visie op data-integratie als met PowerCenter 8 een praktische invulling ervan. Terecht dat Gartner Informatica in het leiderskwadrant plaatst.

Noten

1. *PowerCenter Standard Edition versie 7.1.2 GA en PowerCenter Advanced Edition versie 7.1.1 FCS.*
2. *In de huidige versie van MS SQL Server is DTS omgedoopt/opgevolgd door Integration Services.*
3. *'Tussen 36 en 120 minuten BI per dag. Het Nationaal Data Warehouse Onderzoek 2005: uitkomsten, resultaten en aanbevelingen', P.F.H. van der Linden, uitgeverij Array Publications 2005 ISBN 90-74562-11-6.*
4. *'Evaluating ETL and Data Integration Platforms', Wayne Eckerson en Colin White, 2003.*
5. *'Magic Quadrant for Extraction, Transformation and Loading, 1H05', Gartner 11 mei 2005.*
6. *'Service-Oriented Architecture: Survey Findings on Deployment and Plans for the Future', Eric Austvold en Karen Carter, AMR Research Market Analytix Report: Market Trends Series 2005.*

Paul van der Linden (Paul.PFH.vanderLinden@AtosOrigin.com) is senior consultant Data Warehousing/BI bij Atos Origin en geeft leiding aan Data Warehousing Cost & Lifecycle Management (CLM).