Strijdplan ontvouwd tijdens Informatica World 2005

Op weg naar dominantie in Data-Integratie

Paul van der Linden

Van 5 t/m 7 juni vond in Washington de bijeenkomst Informatica World 2005 plaats. Meer dan 1.000 bezoekers waren aanwezig op de inmiddels zevende International User Conference.

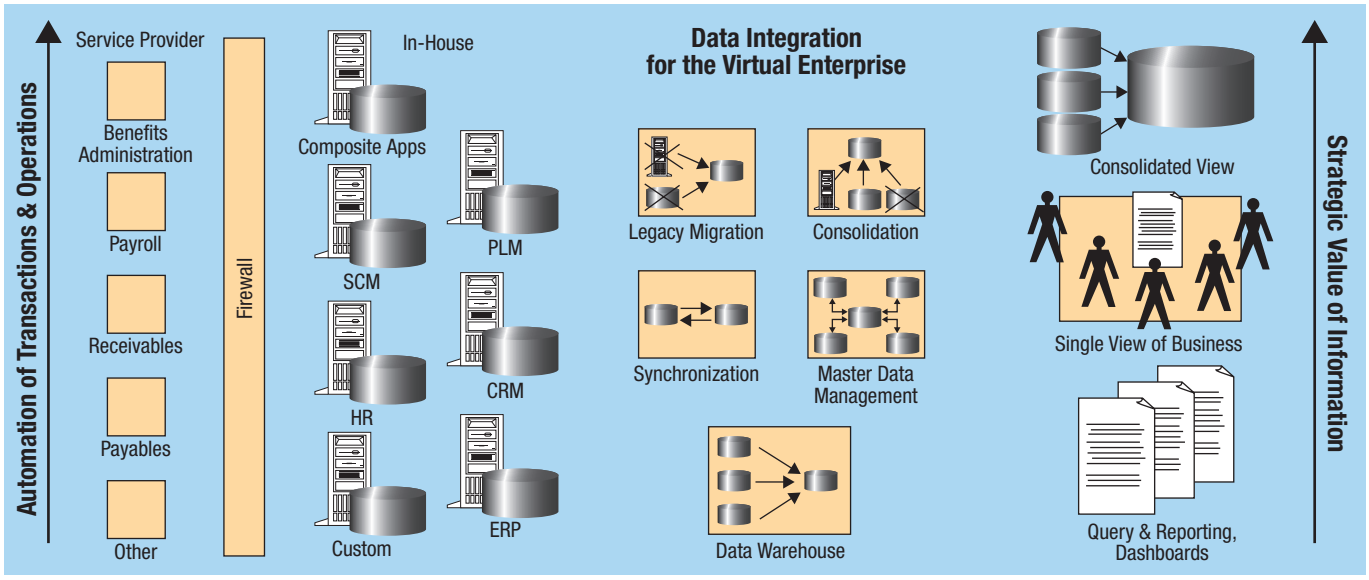
Keynote speeches van onder andere Gartner (Frank Buytendijk en Ted Friedman) en Geoffrey Moore (TCG Advisors) moesten het belang van datakwaliteit en data-integratie onderstrepen. Door president en CEO Sohaib Abbasi werd Informatica's dominantie in data-integratie beleden. Onder het thema 'Data Directions' werd drie dagen lang besproken waar het met IT naar toe gaat en welke rol Informatica daarin kan en wil spelen.

Begin in ETL

Informatica staat alom bekend als marktleider in ETL. ETL staat voor extractie, transformatie en laden van data en vormt binnen datawarehousing een belangrijke component. Het is het traject waarbij data die door verschillende transactionele systemen zijn aangeleverd worden gevalideerd, bewerkt en aan elkaar gesmeed. Het is dan ook niet verwonderlijk dat het merendeel van de inspanning van een datawarehouseproject hier plaatsvindt.
De eerste generatie van ETL-software bestond uit codegeneratoren, waarvan ETI, Carlton en Prism de bekendste zijn. Deze eerste generatie werd opgevolgd door transformation engines die alle ETL-bewerkingen centraal op een server verrichten. Informatica is met haar tweede-generatie-tool bijzonder succesvol geweest. Ondanks de toch heftige aanschafprijs is het onbetwist marktleider. In Nederland gebruikt maar liefst 31 procent Informatica als ETL-tool (cijfers Nationaal Datawarehouse Onderzoek 2005). Volgens Abbasi maakt 78 procent van de Fortune 100 bedrijven gebruik van Informatica.
Inmiddels is ETL als onderwerp minder sexy geworden en hebben de ETL-leveranciers het liever over data-integratie. Abbasi sprak tijdens Informatica World dan ook van het doortrekken van de dominantie in datawarehousing (niet ETL!) naar een dominante positie in Data Integration.

De vele gezichten van Data-integratie

De van Oracle afkomstige opvolger van mede-oprichter Gaurav Dhilon heeft dan ook een bijzonder interessant doel voor ogen. Volgens marktanalistenbureau IDC hebben organisaties in de komende twee jaar te maken met (gemiddeld) maar liefst 26 data-integratieprojecten[1]. De waarde van al deze DI-projecten komt neer op zo'n 13 miljard dollar. Het gaat hierbij om projecten die als doel hebben te zorgen voor een single view op de organisatie, financiële consolidatie, masterdata management, legacy-migratie of data-



Afbeelding 1: Cross-Enterprise Data-integratie in de visie van Informatica.

CEO Sohaib Abbasi: 78 procent van de Fortune 100 bedrijven maakt gebruik van Informatica.

synchronisatie. Outsourcing is in de ogen van Abbasi een beweging die (ook) leidt tot meer datafragmentatie – en dus meer behoefte om al die data weer te kunnen integreren. Het zal duidelijk zijn dat de IT-wereld in de ogen van Abbasi er veel belovend uitziet.

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

- Single view op de organisatie: hierbij gaat het om het verkrijgen van 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).

ETL is binnen de data-integratieverzameling niet zozeer een aparte component (zoals hierboven benoemd) als wel een aspect van verschillende van deze componenten. Vandaar ook dat Informatica denkt met name geschikt te zijn om de data-integratiemarkt te veroveren.
Maar ETL is niet het enige aspect wat ingevuld dient te worden. Dat datavolumes nog steeds toenemen (elke anderhalf jaar een verdubbeling van data volgens Gartner) is genoegzaam bekend. Waar minder bij wordt stil gestaan is dat het hier met name om ongestructureerde data gaat. Dit betreft data die niet zijn vastgelegd in de rijen en kolommen waarover databases beschikken. Denk bijvoorbeeld aan de immense hoeveelheden e-mails of aan informatie die in PowerPoint-presentaties of Word-documenten worden vastgelegd. Volgens Abassi gaat het zelfs om 90 procent ongestructureerde data en slechts om 10 procent gestructureerde data. De uitdaging is dus ook om deze ongestructureerde data zodanig vast te leggen dat ze makkelijk te ontsluiten, gebruiken en te combineren zijn. Tijdens Informatica World werd getoond hoe zoiets gedaan kan worden.
Informatica ondersteunt inmiddels ook Enterprise Information Integration (EII). Hierbij is sprake van het integreren van data zonder dat deze fysiek worden verplaatst. Het voordeel van EII ten opzichte van bijvoorbeeld een datawarehouse-aanpak is evident (sneller), de beperkingen ervan zijn eveneens voor de hand liggend. In voorkomende gevallen kan EII een passende oplossing bieden. Het komt daarmee tegemoet aan de eisen van steeds snellere dataleveranties.

Tooling

Informatica's bekendste product is zonder twijfel PowerCenter. Op dit moment wordt het product in de Standard Edition (versie 7.1.2 GA) en als Advanced Edition (versie 7.1.1 FCS) aangeboden. De Standard Edition (SE) bestaat uit het standaard integratieplatform. Bij de Advanced Edition (AE) is PowerCenter gecombineerd met SuperGlue en PowerAnalyzer. Hierbij geldt dat er sprake is van een 'single install' en is de belofte dat de drie producten zich gedragen als een enkelvoudig product. SuperGlue vormt hierbij de metadatacomponent. PowerAnalyzer is bedoeld voor analyse en rapportage en beschikt over dashboarding.
Daarnaast zijn er bij beide varianten nog verschillende add-ons beschikbaar (PowerCenter Options) onder andere voor real-time, data profiling, partitioning, data cleansing en X Connect voor het connecteren naar metadatabronnen. Het beeld wordt gecompleteerd met PowerExchange, bedoeld voor het mainframe. PowerCenter Advanced Edition is Informatica's product om brede data-integratie in te vullen.

Voor het najaar is de volgende versie van PowerCenter gepland. Deze product-release (codenaam: project Zeus) biedt volgens opgave van Abbasi een 10 keer betere performance en een 10 keer betere productiviteit dan het huidige product. Daarbij wordt beloofd dat het tien keer zoveel data aankan. Niet dat de tevredenheid van de huidige klanten te wensen overlaat. Maar liefst 94 procent van de gebruikers vernieuwt het maintenance-contract, tegen een industriegemiddelde van 86 procent. De versie die project Zeus moet opleveren is bedoeld voor 'mission critical enterprise deployments'. In de herfst van 2006 volgt weer een volgende versie die een uniforme toegang tot alle data belooft (on demand data integration).

Integration Competency Centers

De uitdagingen waarmee ondernemingen de komende jaren te maken krijgen zijn volgens Abbasi enorm. Enerzijds is er sprake van steeds meer data, anderzijds moeten die data ook steeds sneller en op meer plaatsen beschikbaar zijn. Indien Business

Intelligence inderdaad afdaalt naar de werkvloer (bekend als operational BI) betekent dit dat ook steeds meer personen toegang moeten hebben tot die data. Daarbij leidt de huidige outsourcingstrend (BPO, IT outsourcing, ASP) tot een verdere fragmentatie en versnippering van data. Organisaties zullen derhalve moeten overgaan tot de inrichting van een Integratie Competency Center (ICC). Alleen op deze wijze is overzicht en sturing mogelijk in een steeds complexer worden datalandschap.

Een ICC is een clustering van beschikbare kennis en ervaring op het gebied van data-integratie. Het gaat hierbij zowel om mensen als om processen en technologie. Doel is om zoveel mogelijk gebruik te maken van de kennis, ervaring en vaardigheden die een organisatie reeds heeft om niet het wiel uit te moeten vinden. Vanuit het ICC worden dan ook IT-diensten aangeboden aan de verschillende projecten die betrekking hebben op data-integratie. Gezien de verwachtingen van analistenbureaus als IDC is het zeer zeker de moeite waard om zo'n investering in een ICC te doen. Het model dat Informatica hanteert maakt een onderscheiding in vijf fasen.

De eerste fase is de situatie waarin elk project voor zichzelf naar oplossingen zoekt en haar eigen ideeën implementeert. Vanuit organisatieoogpunt is dit een minder prettige situatie die leidt tot een zeer complex totaalplaatje. In deze situatie is dan ook nog geen sprake van een ICC. Fase 2 tot en met 5 kennen wel een ICC, waarbij steeds meer aspecten van de drie-eenheid mensen-processen-technologie worden ingevuld. De eerste vorm van een ICC wordt gekenmerkt door best practices. De processen zijn hierbij bepaald en vanuit de best practices worden aanbevelingen gedaan welke technologieën het beste kunnen worden ingezet. De personen met ICC-kennis zitten overal verspreid in de organisatie. Het grote voordeel van deze fase ten opzichte van de beginsituatie is dat de beschikbare kennis en ervaring nu worden gedeeld. In fase 3 is sprake van standaardisatie op technologie waardoor meer consistentie wordt bereikt. Dit betekent ook dat in termen van beheer en onderhoud een aanzienlijke vereenvoudiging wordt bereikt.

In fase 4 is sprake van gedeelde technologie (shared services). Een deel van de personen met ICC-kennis behoort nu tot een centraal ICC. Overigens gaat het hierbij om een virtueel samentrekken; het hoeft niet zo te zijn dat mensen ook fysiek bij elkaar worden gezet. Het meest uitgewerkte ICC-model wordt gekenmerkt door centrale diensten. Er is sprake van een (virtueel) centraal ICC met gedeelde technologieën en gedefinieerde processen. In deze situatie maakt de organisatie maximaal gebruik van de kennis, ervaringen en vaardigheden die ze heeft op het gebied van data-integratie. Informatica stelt zich met haar visie op een Integratie Compententie Center en haar vlaggenschip PowerCenter duidelijk kandidaat voor de functie van data-integratiestandaard.

Kansen

Welke kans heeft Informatica om inderdaad de nummer 1 in data-integratie te worden? Duidelijk is dat haar dominantie op ETL-gebied een goede uitgangspositie verschaft. Maar ETL is natuurlijk nog geen data-integratie. In het in 2003 verschenen rapport 'Evaluating ETL and Data Integration Platforms' van TDWI beschrijven Wayne Eckerson en Colin White aan welke eisen een DI-oplossing idealiter moet voldoen. Uiteraard moeten grote hoeveelheden data verwerkt kunnen worden en verschillende databronnen kunnen worden ontsloten. Daarnaast moeten oplossingen worden geboden om de krimpende batchwindows te kunnen ondervangen en moet tegemoet worden gekomen aan het steeds operationeler worden van BI. Denk dan bijvoorbeeld aan 24*7 beschikbaarheid van data. Datakwaliteit en metadatamanagement zijn verdere vereisten en tenslotte vragen organisaties om packaged solutions. Voorwaar geen geringe lijst van eisen. Hoe ver komt Informatica hiermee? Voor wat betreft de eerste vier eisen heeft Informatica met PowerCenter een sterk product. Uiteraard zijn er uitwerkingen te definiëren die momenteel nog niet aanwezig zijn en zal men gezien de exploderende datavolumes zeker niet op haar lauweren kunnen rusten, maar op deze punten wordt goed gepresteerd. Waar het gaat om metadata-management wordt het op het Common Warehouse Metamodel (CWM)-gebaseerde SuperGlue naar voren geschoven. Datakwaliteit is momenteel een optie en beperkt tot naam- en adresgegevens. Blijft over packaged solutions.

De grote concurrent Ascential is inmiddels overgenomen door IBM. Abbasi ziet dit niet als een bedreiging voor zijn bedrijf. 'Een overname betekent dat de betreffende organisatie minimaal twee tot drie jaar bezig is om de componenten in elkaar te sleutelen. Dat is een interne focus. Wij gebruiken die tijd om naar buiten, naar de markt te kijken en verder te komen'.

Conclusie

Informatica wil onder de nieuwe aanvoerder Sohaib Abbasi haar dominante rol in ETL doortrekken naar leiderschap in data-integratie. Gezien de brede scope van data-integratie is dat een enorme uitdaging voor elke leverancier. Informatica beschikt met PowerCenter Advanced Edition en haar visie op Integration Competency Centers over misschien wel de beste kaarten. Van de door The Data Warehousing Institute genoemde zeven eisen die idealiter gesteld kunnen worden aan een data-integratieoplossing worden er al (bijna) zes ingevuld. Niet of, maar wanneer de resterende aspecten gerealiseerd gaan worden is de vraag. Met de overname van de voornaamste concurrent Ascential door IBM staat Informatica weinig in de weg in haar queeste naar data-integratiedominantie.

Noot

1. Bron: '2004 Custom IDC Data Integration Market Sizing Study'. Gebaseerd op 150 interviews met grote organisaties.

Paul van der Linden

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).