Onderzoeksrapport derde generatie ETL-tools interessant naslagwerk

ETL-tool kopen versus zelf bouwen?

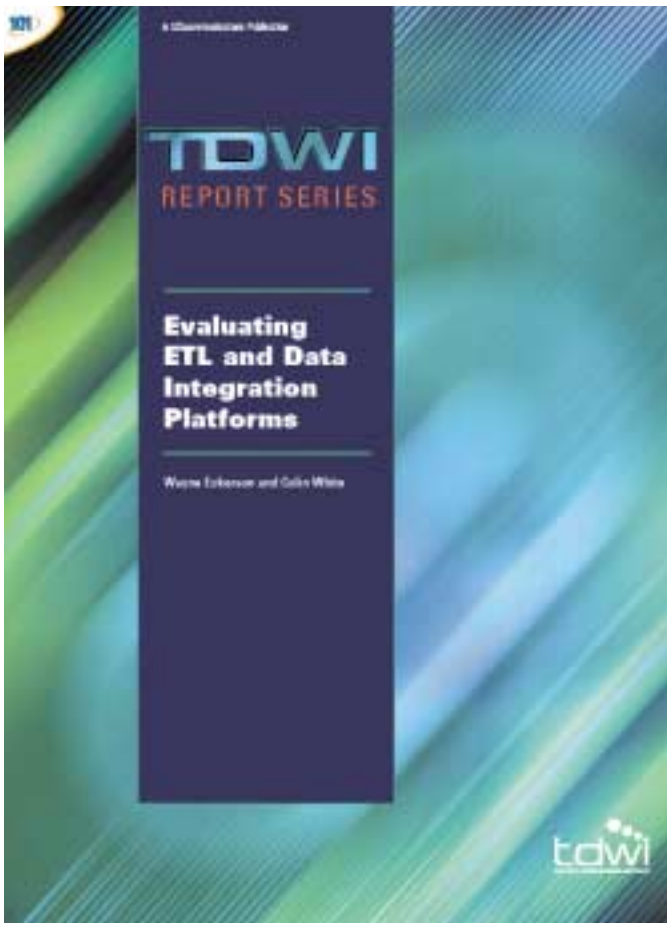
Paul van der Linden

The Data Warehousing Institute (TDWI) is een organisatie die zich al sinds 1995 bezig houdt met alle zaken betreffende Business Intelligence en datawarehousing. Sinds jaar en dag wordt gewaarschuwd voor valkuilen in ontwerp- en realisatietrajecten en worden best cases beschreven. Eén van de activiteiten van de TDWI bestaat uit het verrichten van onderzoeken op het werkterrein, hetgeen dan weer leidt tot een rapport. Zo hebben Wayne Eckerson en Colin White een rapport geschreven met de titel 'Evaluating ETL and Data Integration Platforms'. Het onderzoek waarop het rapport is gebaseerd betreft een survey die in november 2002 onder meer dan 1000 BI-professionals werd gehouden.

Het rapport behandelt de afweging tussen het kopen van een ETL-tool versus het zelf bouwen van deze functionaliteit. Daarbij wordt gesteld dat een nieuwe generatie van tools aan het ontstaan is, die door TDWI data integration-platforms worden genoemd. Tenslotte wordt een lijst met evaluatiecriteria behandeld aan de hand waarvan de selectie van een ETL-tool kan plaatsvinden.

Historie van ETL-tools

De eerste generatie van ETL-tools bestond uit codegeneratoren die op verschillende machines hun code achterlieten. Aangezien de programma's hierdoor dicht tegen de data aanzaten, leverde dit een goede performance op. Vanuit het oogpunt van beheer en onderhoud vormde deze decentrale constructie echter een minder geslaagde oplossing. Prism, Carlton en ETI vormden de belangrijkste producten uit deze eerste lichting.
De volgende generatie van ETL-tools betrof ETL-servers die centraal, vanaf een server, het hele ETL-proces afhandelden.
Het voordeel van deze *hub-and-spoke* gedachte bestaat vooral uit de vereenvoudiging van de aansturing van het hele proces. Ook het overzicht dat men heeft en houdt, is een niet-triviaal voordeel boven de eerste generatie producten. Het bekendste product van deze zogenaamde tweede generatie ETL-tools is PowerCenter van Informatica.
Inmiddels is er sprake van een derde generatie ETL-tools. The Data Warehousing Institute noemt dit het data-integratieplatform.



De eisen die door de business worden gesteld aan de ETL-tools houden onder andere in dat gegevens ook in real-time beschikbaar dienen te zijn. Dat heeft er in de markt al toe geleid dat ETL-tools en EAI-oplossingen naar elkaar toe zijn gegroeid.
ETL-tools zijn groot geworden in de context van datawarehouses. Door de combinatie met EAI, wat een applicatiegerichte context kent, ontstaat het data-integratieplatform. Dit betekent in wezen een verdere verzelfstandiging van ETL-tools.

Data-integratieplatform

Aan het data-integratieplatform worden diverse eisen gesteld die momenteel door geen enkel ETL-tool allemaal worden ingevuld. Het gaat dan om eisen gesteld aan het platform en aan de data-integratie.

Tot de platform-aspecten 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 zijn inmiddels genoegzaam bekend. Een aantal vraagt 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 worden ook web services en XML begrepen.

> De eisen houden in dat gegevens ook in real-time beschikbaar dienen te zijn

Bij globaal metadata management wordt de Common Warehouse Metamodel (CWM) van de OMG genoemd. De TDWI vindt het te vroeg om aan te geven of deze metadata-standaard doorzet. Toch is de toekomst met vertrouwen tegemoet te zien. In eerste instantie zullen softwareleveranciers CWM gebruiken om zich te onderscheiden van hun concurrenten. Vervolgens gaan organisaties deze vraag ook stellen aan leveranciers, waardoor meer leveranciers zich (serieus) gaan bezighouden met CWM. Zie hier het ontstaan van een de facto standaard.

Buy or build?

Uit het onderzoek van de TDWI blijkt dat 45 procent van de ondervraagde organisaties gebruik maakt van ETL-tools en 18 procent kiest voor zelf bouwen. Daarnaast bestaat er echter een groep van 37 procent, die gebruik maakt van een ETL-tool, maar daarnaast ook zelf ETL-programma's in elkaar sleutelt.
De belangrijkste reden die wordt genoemd om een ETL-tool te kopen is de tijdwinst die geboekt wordt bij het ontwikkelen en onderhouden van code. Echter, het duurt zo'n drie maanden om een ETL-tool onder de knie te krijgen en vervolgens nog minstens een half jaar om er in te kunnen lezen en schrijven. De twee belangrijkste redenen om zelf aan de slag te gaan zijn dan ook dat dat goedkoper is en dat men daarmee sneller resultaat boekt. Hierbij speelt mee dat verschillende bedrijven de mening zijn toegedaan dat de bestaande ETL-tools onvoldoende aansluiten op hun business en dat ze met name de eenvoudige taken automatiseren, en van mening zijn dat de meer complexe taken nog steeds handmatig moeten worden opgepakt. Vandaar deze groep van 37 procent, die in principe heeft gekozen voor een ETL-tool, maar de strategie hanteert dat waar dat beter past handmatig gegenereerde code wordt ingezet.
Gezien het doel van ETL-tools en de reden dat toch naar handwerk wordt gegrepen, kan dit niet anders worden gezien dan als een uitdaging aan ETL-leveranciers om hun producten verder te ontwikkelen. Het doel is om het aantal regels code dat wordt geschreven buiten de tool te minimaliseren. Behalve het onderhoud en beheerargument geldt hierbij ook dat ontwikkelaars die gebruik maken van een ETL-tool vijf tot zes keer zo productief zijn vergeleken bij hen die programmeren vanaf de prompt volgens Pieter Mimno, algemeen beschouwd als een van de belangrijkste analisten op het gebied van ETL- en BI-tools. Bespottelijk blijven de hoge prijzen die betaald moeten worden voor ETL-tools. In dit verband wordt dan ook gesproken van 'sticker shock'. De betreffende tools kosten zo rond de 200.000 dollar en dat is voor een heleboel bedrijven een drempel die ze niet (of niet in een keer) kunnen nemen. Gelukkig dat het gebundeld aanbieden van ETL-functionaliteit, als onderdeel van de database of een BI-suite, enige druk op deze markt uitoefent.

Selectiecriteria

Het derde deel van het rapport, waarin de criteria worden besproken om een ETL-tool (of data-integratieplatform) te kunnen evalueren, is wellicht het minst belangrijk. Immers, het zal duidelijk zijn dat de geboden en gevraagde functionaliteiten zich steeds verder ontwikkelen. Het is geen goede zaak om te kiezen voor een specifieke tool alleen op basis van de grotere functionaliteit die het op een bepaald tijdstip biedt. Haasje-over is immers de realiteit in de IT-markt. Belangrijker is de blijvende steun en het commitment die men van de leverancier mag verwachten, iets wat in het rapport ook wordt gezegd. Hiermee loopt vervolgens de lucht uit de genoemde selectiecriteria.

Conclusie

Het beste deel van 'Evaluating ETL and Data Integration Platforms' bestaat uit de resultaten van het onderzoek. Hierin staan harde gegevens hoe ETL in de praktijk wordt gebruikt.
De beschouwingen over de ontwikkelingen van de ETL-tools naar het data-integratieplatform en de genoemde selectiecriteria zijn valide maar bieden, vrees ik, toch een schijnzekerheid. Wat je als gebruiker eigenlijk wenst is een concrete vertaling vanuit een business requirement naar een passende IT-oplossing, in plaats van een opsomming van alle criteria die van toepassing zouden kunnen zijn. Dat dit ontbreekt is een gemiste kans. Hierdoor is het rapport wel een interessant naslagwerk over ETL, maar te weinig een praktische handleiding om concrete beslissingen te nemen.

Paul van der Linden (Paul.PFH.vanderLinden@AtosOrigin.com) is senior consultant Data Warehousing/BI bij AtosOrigin.