De ETL Matrix Unloaded en Reloaded

De ETL Matrix Unloaded

Rick Mutsaers

Met plezier heb ik het artikel de ETL Matrix Reloaded gelezen van Daan van Beek (DB/M 6 2005), waarin hij de ontwikkelingen van de ETL toolmarkt beschrijft.

Hij stelt daarin onder meer dat de ontwikkeling van ETL tools nog lang niet gestopt is en dat er de komende tijd nog een hoop te doen is. Daar ben ik het helemaal mee eens. In zijn matrix staat vervolgens van een 17-tal ETL tools uitgewerkt hoe deze tools op diverse fronten als functionaliteit, gebruikersvriendelijkheid en dergelijke scoren. De argeloze lezer zou dit kunnen lezen als een perfecte vergelijkingsmatrix op basis waarvan een keuze voor een bepaald tool gemaakt kan worden.
Echter, bij het uitvoeren van een gedegen toolkeuzetraject moet niet alleen gekeken worden naar de functionaliteit die het tool biedt, maar zeker ook naar zaken als marktpenetratie/aandeel, kennis in de markt, onderhoudbaarheid van de ETL-processen en bijbehorende documentatie; een voorwaarde in een markt waarbij veel van de ETL-processen projectmatig ontwikkeld worden door veelal externe medewerkers van IT-dienstverleners en het beheer en onderhoud juist door interne medewerkers worden uitgevoerd. Deze zaken ontbreken echter in de matrix.
Uiteraard is het mooi om te zien dat er tools zijn die 130 native connecties aan kunnen, maar in hoeveel gevallen heb je dat daadwerkelijk nodig? Natuurlijk is het prettig als het databaseformaat van je bron- en/of target-omgeving door het tool ondersteund wordt, maar de praktijk leert dat de meeste klanten (combinaties van) Oracle, Microsoft, DB/2 en Sybase als target-omgeving gebruiken. En in de meeste gevallen is het mogelijk om vanuit een exotisch bronsysteem, dat niet direct of door middel van een ODBC-koppeling door het ETL tool aangesproken kan worden, een export aan te maken in flatfile- of XML-formaat en daarmee kunnen alle ETL tools overweg.

Beauty contest

Het punt is eigenlijk dat de ETL Matrix een soort beauty contest is. Welk product heeft de meeste functionaliteiten, kan de meeste platforms aan en heeft het afgelopen jaar het grootste aantal nieuwe features erbij gekregen. Dat is op zich prima en biedt de lezer een goed beeld van wat een bepaald ETL tool allemaal kan. Het geeft echter mijns inziens een vertekend beeld van de werkelijke waarde van een ETL tool. Daar spelen meer aspecten een belangrijke rol en kunnen sommige aspecten beter anders weergegeven worden.
Neem een aspect als groeipotentie. Wanneer een nieuwkomer als BusinessObjects Data Integrator, na de eerste twee jaren goed naar zijn markt gekeken te hebben, het aantal features 'op niveau' van de concurrentie brengt, levert dat ineens een sprong op waardoor het lijkt alsof er veel groeipotentie is, terwijl in werkelijkheid een inhaalslag gemaakt is. 'Oudere' tools, zoals Informatica's PowerCenter en Ascential DataStage zijn geleidelijk met de markt meegegroeid en hebben daardoor een veel gestager groei meegemaakt. De curve in de groeipotentiegrafiek loopt daardoor minder steil, maar het aantal jaren markt- en ontwikkelervaring dat in zo'n tool is ingebouwd is vele malen groter dan bij een nieuwkomer. Dit spreekt echter niet uit de groeipotentiegrafiek. Groeipotentie zou daarom bijvoorbeeld beter weergegeven kunnen worden in termen van groei in marktaandeel.

Prijs/kwaliteit

Ook de grafiek waarin de verhouding prijs/functionaliteiten wordt weergegeven vertekent sterk. Uiteraard ligt de prijs per functionaliteit het laagst bij leveranciers waar het ETL tool meegeleverd wordt met de database, zoals bij Microsoft en Oracle het geval is. Maar wat heb je aan een gratis ETL tool als het een enorme inspanning in termen van ontwikkeltijd kost om er een ETL-proces mee te bouwen. Een betere maatstaf zou zijn om een aantal standaard ETL-processen te bedenken en te meten hoeveel moeite in termen van ontwikkel-, implementatie- en onderhoudstijd en hoeveel ETL-objecten er nodig zijn om deze ETL-processen uit te werken in een bepaald tool. Op basis daarvan kan dan een indexcijfer bepaald worden, dat verrekend kan worden met de aanschafprijs. De vraag is dan of Microsoft en Oracle nog steeds bovenaan staan.

Architectuur

Een punt dat ook niet heel expliciet aangehaald wordt is bijvoorbeeld dat een product als Oracle WarehouseBuilder min of meer een Oracle database als target vereist en dat Microsofts Integration Services een Microsoft SQL Server database vereisen. En dat beide producten hun ETL-processen 'in' de database uitvoeren. Dit beperkt de keuze van het target databaseplatform. Bovendien zijn zaken als schaalbaarheid en fail-over minder makkelijk in te regelen wanneer de ETL-component en database-

component aan elkaar vastzitten zoals bij Oracle en Microsoft duidelijk het geval is. Een Oracle PL/SQL-procedure (het resultaat van de deployment van een ETL-proces) kan maar op één node tegelijk werken, terwijl bij een tool als PowerCenter de mogelijkheid bestaat één sessie op te splitsen in meerdere gepartitioneerde streams die elk op een afzonderlijke ETL server kunnen draaien. Ook tools als Cognos DecisionStream en BusinessObjects Data Integrator werken het best in combinatie met de 'eigen' reporting-omgevingen, waardoor de effectieve inzetbaarheid bij grote organisaties – waar vaak meer smaken reporting tools gebruikt worden – kleiner is. Uiteraard geldt wel dat wanneer gebruik gemaakt wordt van het bijbehorende reporting tool, de synergievoordelen groot zijn. Bij Cognos Decisionstream kan bijvoorbeeld rechtstreeks het model voor een PowerPlay-kubus gegenereerd worden.

Werkelijke waarde

Uiteraard geeft de ETL Matrix op een groot aantal objectieve criteria inzicht in de kwaliteiten van een ETL-product. Dat staat buiten kijf. De werkelijke toegevoegde waarde van een dergelijke ETL Matrix zou echter sterk vergroot kunnen worden door ook te kijken naar de werkelijke waarde van een ETL-product.
Deze bestaat, naast de objectieve zaken zoals in de ETL Matrix vermeld, met name uit zaken als:

- Financiële positie van de leverancier. Is de leverancier in staat voldoende investeringen in het product te (blijven) doen om het verder te ontwikkelen;
- Marktaandeel van het ETL-product. De leiders qua marktaandeel hebben meestal de features in huis waar de meeste klanten om vragen. De match tussen featurevraag uit de markt en feature-aanbod in het product is bij de leiders meestal het grootst en daarmee de kans dat een product breed inzetbaar is binnen een bedrijf;
- Kennis in de markt. Hoe gemakkelijk is het om (extern) personeel met kennis van en ervaring met het product in de markt te vinden;
- Werkelijke ontwikkelkosten. Hoe lang duurt het (in termen van doorlooptijd van ontwikkeling en implementatie) voordat een ETL-proces van tekentafel- naar productieversie is om te vormen. Dit hangt samen met de gebruikersvriendelijkheid, taakcompatibiliteit, maar ook met zaken als het effectief in kunnen zetten van kant-en-klare brokjes ETL-functionaliteit (zoals Slowly Changing Dimension of automatisch genereren van metadata ten behoeve van reporting tools);
- Onafhankelijkheid leverancier. Wanneer het ETL-product de core business van een leverancier is, is het waarschijnlijk dat hij dit product koestert en verder ontwikkelt. Stilstand is immers achteruitgang. Wanneer het ETL-product een zogenaamd neven-product is, is het nog maar de vraag hoeveel ontwikkel-effort de leverancier in zijn ETL-product blijft steken. Met andere woorden, hoe zeker is de toekomst van een bepaald ETL tool.

Conclusie

De ETL Matrix is een uitstekende basis voor het doen van een selectietraject voor een ETL tool. Het mist echter een aantal essentiële aandachtspunten die tijdens een dergelijk traject zeker bekeken moeten worden. Wanneer de matrix met deze punten uitgebreid wordt, heeft de lezer een instrument in handen waarmee hij de perfecte ETL tool voor zijn toepassing kan selecteren. En wanneer deze extra aandachtspunten verwerkt zijn in de matrix zouden de winnaars van de vorige ronde wel eens hoge ogen kunnen gooien.

Rick Mutsaers (rick.mutsaers@ordina.nl) is senior BI Consultant bij Ordina VisionWorks.

Reloading the Matrix

Daan van Beek

Ik ben blij met de positieve waardering van Rick Mutsaers over de ETL Matrix Reloaded. Het overgrote deel van de zaken die hij noemt is echter onterecht en hij slaat de plank meerdere malen flink mis.

Marktpenetratie. Dit zegt natuurlijk niet veel over de kwaliteit van de ETL tool en in hoeverre de tool past binnen de IT-infrastructuur en bij het niveau van de ETL-ontwikkelaars.
Het huidige marktaandeel zegt iets over het verleden, en niets over de toekomst. Ooit had de tekstverwerker WordPerfect ook een enorm marktaandeel. De groeipotentie zoals deze binnen het ETL-onderzoek is gedefinieerd (zie DB/M 6 van 2005) geeft aan hoe vaak een leverancier een nieuwe *waardevolle* functie uitbrengt, volgens mij een belangrijke graadmeter voor innovatie en toekomstige prestaties. Dat nieuwkomers het kunstje van de 'oudere' tools afkijken is ook niet waar. Business Objects heeft bijvoorbeeld een aantal interessante innovaties toegepast – waaronder data auditing – die niet zitten in de 'oudere' tools.
Het is daarnaast erg jammer dat PowerCenter en DataStage geen out-of-the-box ondersteuning bieden voor veel voorkomende taken zoals het opbouwen van historie en slowly changing dimensions.

Zelfs 'nieuwkomer' Microsoft Integration Services biedt standaard die mogelijkheid! Overigens is het zo dat Data Integrator (voorheen Acta) van Business Objects slechts twee jaar later op de markt is gekomen dan PowerCenter. Hoe opmerkelijker is het dat het evenveel features bevat en stukken gebruiksvriendelijker is.

Bovendien is er inmiddels in de markt voldoende kennis van Data Integrator, Microsoft Integration Services, Oracle Warehouse Builder – tools met een hoge gebruiksvriendelijkheid – en ook minder bekende tools als Sunopsis ETL. Misschien (nog) niet bij de grotere integrators, maar wel bij de kleinere partijen in de markt. Tegenwoordig kunnen veel ETL-ontwikkelaars meerdere tools aan, en is het niet zo heel moeilijk om vanuit die kennis en ervaring een andere tool aan te leren. Kennis in de markt van ETL tools met een kleiner marktaandeel is dus ofwel aanwezig, of gemakkelijk aan te leren.
Tot slot: de complexiteit van de 'oudere' tools vraagt om externe specialisten van IT-dienstverleners om ETL-processen te bouwen.

Gratis ETL

Inderdaad heb je niets aan een gratis ETL tool wanneer je daarmee tegen allerlei moeilijkheden aanloopt en het ontwikkelproces een enorme inspanning vergt. Goedkoop is in dat geval duurkoop. Het feit wil echter dat juist de zogenaamde (bijna) gratis ETL tools uitstekend scoren op gebruiksvriendelijkheid en functionaliteit. Laten we het omdraaien: hoe zou het zijn om in een dure Mercedes met pech langs de weg te moeten staan?
Overigens zijn de tools niet helemaal gratis, je moet altijd de database erbij kopen, iets wat we al eerder hebben opgemerkt in ons artikel.
En stel nu dat men een goedkope ETL tool aanschaft: in plaats van € 100.000,- aan licenties te betalen kan men een externe ETL-ontwikkelaar ruim 100 dagen lang laten ontwikkelen! Daar kan men toch een aardig datawarehouse mee bouwen.

Over de *native connecties* kan ik kort zijn. Hoewel dit voor slechts enkele procentpunten meetelt in de eindscore van een ETL tool, kan het toch een belangrijk selectiecriterium zijn. Bijvoorbeeld het werken met flatfiles, wat Rick Mutsaers als oplossing aandraagt voor slechte connectiviteit, veroorzaakt verlies van essentiële metadata tijdens het eerste deel van het ETL-proces. Dit draagt niet bij aan eenduidige overdracht van informatie en verhoogt het risico op fouten in het ontwikkelproces. Ook zijn sommige applicaties zó 'gesloten' dat die überhaupt geen mogelijkheid kennen om flatfiles te genereren. En dan is het toch handig als er een kant-en-klare adapter voorhanden is. Zelfs XML, dat veel ballast met zich meebrengt, is niet altijd wenselijk en vormt zeker geen ideale oplossing bij grote hoeveelheden data die in batch verwerkt moeten worden. Daarnaast is changed data capture (alleen de wijzigingen oppakken) alleen effectief wanneer er een native connectie gelegd kan worden naar een database of server. Flatfiles en een niet onbelangrijke functie als changed data capture kunnen dus niet goed samengaan.

Architectuur

Oracle Warehouse Builder vereist inderdaad een Oracle database als target, daarvoor heeft men ook minder punten gekregen in het onderzoek. Hetzelfde geldt voor Microsoft dat alleen draait op Windows. Het is wel zo dat veel datawarehouses draaien op Unix in combinatie met Oracle, en NT met SQL Server. Als men Oracle al als database gebruikt, dan zou het op zijn minst onverstandig zijn om alleen af te gaan op het marktleiderschap van Informatica en/of de financiële positie van de leverancier. Overigens biedt ook Oracle partitionering aan en kunnen verschillende clusters of kan zelfs een grid worden gedefinieerd, echter dit gaat alleen in combinatie met de database (zie Matrix). Ook hier zijn in het ETL-onderzoek minder punten voor toegekend. Dat Cognos en Business Objects het beste werken met de eigen producten is vrij logisch. Dat zogenaamde onafhankelijke ETL tools beter zijn in het uitwisselen van metadata met andere BI-producten is daarentegen zeer de vraag. Veel (grotere) organisaties die meerdere smaken BI-producten in huis hebben, zijn op dit moment bovendien bezig met het consolideren van hun BI-omgeving en willen toe naar één BI-leverancier. Ook al is de kwaliteit van één onderdeel – bijvoorbeeld ETL – wat minder in vergelijking met anderen. De opbrengsten van standaardisatie zijn soms velen malen groter dan de voordelen van een best-of-breed oplossing.

Beauty contest

De werkelijke waarde van een ETL tool hangt natuurlijk niet alleen af van de aangeboden functionaliteiten, het gebruiksgemak, de connectiviteit, de platformondersteuning en de mate van innovatie. Maar het geeft wel een goede uitvalsbasis om tools te vergelijken en bijvoorbeeld een shortlist op te stellen. Uiteindelijk zal een proof-of-concept uitsluitsel moeten geven of de functionaliteiten ook echt werken, of er voldoende performance behaald wordt en of de snelheid van ontwikkelen verhoogd kan worden, ook bij complexe ETL-processen. Natuurlijk moet men ook kijken naar de financiële positie van de leverancier en of er voldoende kennis van de ETL tool in de markt op dat moment beschikbaar is.

De opmerkingen en aandachtspunten van Rick Mutsaers (naast senior adviseur bij Ordina ook mede voorzitter van de Nederlandstalige Informatica-gebruikersgroep) snijden naar mijn mening eigenlijk geen hout. Dat is jammer want er zijn best wel zinvolle aanvullingen te bedenken voor de Matrix. Bijvoorbeeld de kwaliteit van de respons van een supportafdeling van de ETL-leverancier, data auditing, ondersteuning van model driven architectures etcetera. Ook een ETL rad race zou een welkome aanvulling betekenen op de Matrix. Misschien wordt het tijd om die te gaan organiseren!

Daan van Beek (daanvanbeek@passionned.nl) is Managing Consultant bij Passionned, auteur van het boek 'De intelligente organisatie: prestatieverbetering en organisatie-ontwikkeling met Business Intelligence' en organisator van de Business Intelligence Awards (www.biaward.nl).