Gratis online overzicht van ETL-tools opnieuw geactualiseerd

Update ETL-matrix

Norman Manley

Het gat tussen enerzijds wat ETL-leveranciers bedenken en anderzijds wat gebruikers willen, lijkt elk jaar groter te worden. Het begint al bij de naam. Gebruikers praten nog steeds over ETL (Extract, Transform en Load), terwijl leveranciers dit een gepasseerd station vinden. Zij praten over data-integratie, iets dat breder is dan de oorspronkelijke ETL, maar dezelfde problematiek beschrijft.

Ervan uitgaande dat Google-statistieken een indicatie geven van waar men in geïnteresseerd is, zien wij dat wereldwijd 240.000 mensen 'ETL' intikken als vraag per maand (2400 in Nederland). De zoekterm 'data integration' wordt 'slechts' door 74.000 mensen (720 in Nederland) ingetikt. Op de eerste Google-pagina van 'data integration' zien wij bijna uitsluitend informatie van leveranciers. De eerste 'ETL' pagina is bijna leveranciervrij, en bevat heel veel (neutrale) informatie over het onderwerp zelf.

Goede positionering van de producten blijft, zoals blijkt uit bovenstaand voorbeeld, een probleem. Door de overnames in de afgelopen jaren heeft een aantal van de grote leveranciers (IBM, Oracle) meerdere producten die min of meer hetzelfde lijken te doen. Bedrijven die meerdere producten in huis hebben en deze willen gaan rationaliseren, vragen zich af welk product in de toekomst blijft bestaan en nog belangrijker; welk product niet. Daarnaast zijn de Open Source producten sterk in opkomst. Twee jaar geleden waren deze nauwelijks meer dan een stuk hobbyisme, maar nu praten wij over producten die qua functionaliteit goed te vergelijken zijn met de marktleiders.

Selectiecriteria	IBM Information Server	Informatica PowerCenter	Pitney Bowes / Data Flow	Talend Open Studio & Integration Suite	Pervasive Data Integrator
Step by step running	yes	yes	no	yes	yes
Row-by-row running	yes	yes	no	yes	yes
Breakpoints	yes	yes	no	yes	yes
Software watchpoints	yes	yes	yes	yes	yes
Compiler/validate	yes	half	-	yes	yes
Corrections to syntax and/or field names		half		yes	

Informatie aanvragen / geselecteerde Producten

IBM Information Server Pervasive Data Integrator
Informatica PowerCenter Talend Open Studio & Integration Suite

De ETL-matrix is te vinden op www.dbm. nl, onderaan de pagina kunt u uw keuze invoeren.

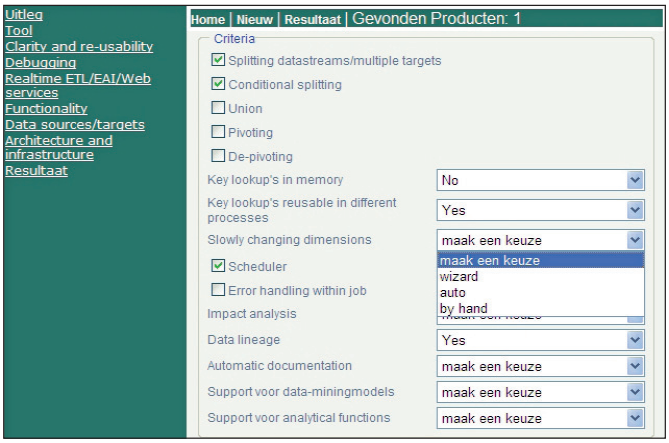
De rol van de producten is in de loop van de tijd ook veranderd. Vier of vijf jaar geleden waren er twee mogelijkheden voor een ETL-product, te weten:

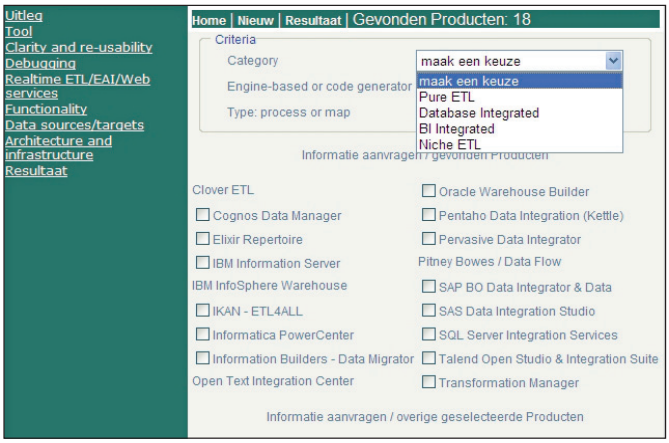
1. Het werd gebruikt om één keer in de zoveel tijd gegevens over te brengen van één of meer bronsystemen naar een Datawarehouse. Onderweg waren de gegevens enigszins schoongemaakt, ontdubbelt en voor zover mogelijk gevalideerd;
2. Het werd ingezet bij een datamigratie. Eenmalig werden de brongegevens overgebracht van het oude systeem (vaak zelfgebouwd, dat vervangen werd door een nieuw aangekocht ERP-systeem) en geladen, met veel pijn en moeite, in het nieuwe systeem. Een proces dat zo moeilijk kon zijn dat veel bedrijven hiermee stopten, en helemaal geen historie hebben meegenomen.

Dit waren hele lange procedures, die vaak 's nachts draaiden, of als het echt te veel was in het weekend. Veelal ging het fout, en moest men de volgende dag (of het volgende weekend) het opnieuw gaan proberen. Beide bovengenoemde problemen bestaan nog. Het datawarehouse wordt tegenwoordig wat vaker up-to-date gehouden, en soms wordt dat op een wat slimmere manier gedaan, maar dit is wat ETL-tools altijd gedaan hebben en waarschijnlijk ook altijd zullen blijven doen.

Om van dit nogal saaie imago af te komen hebben de leveranciers 'data-integratie' uitgevonden, maar gezien de cijfers van Google nog niet goed in de markt gezet. Data-integratie is geen tool maar een architectuur, het fundament dat je nodig hebt voor het bereiken van het ultieme; "Een bron van de waarheid". Data-integratie zorgt ervoor dat er bedrijfsbreed voor ieder data-element maar één definitie is. Geen discussies meer over wat bijvoorbeeld een klant is, ergens in je geïntegreerde omgeving staat het éénduidig gedefinieerd. Data-integratie is een droom, en dromen mag, maar dromen kunnen ook nachtmerries worden. Veelal weten we niet goed wat voor data we in huis hebben en pas op het moment dat we gaan proberen om gegevens aan elkaar te koppelen komen we erachter dat het anders ligt dan we altijd gedacht hadden. Op zich een probleem dat wel degelijk aangepakt moet worden, maar het maakt nonsens van de planning die daarvoor lag.

Ten opzichte van de traditionele ETL-producten zijn er enkele belangrijke verbeteringen in de data-integratieproducten. De belangrijkste gebieden waar de leveranciers aandacht aan hebben besteed zijn:





1. Real-time toegang tot databronnen;
2. Datakwaliteit, afkomst van de data(lineage) en dataprofielen;
3. Cloud computing en SaaS;
4. Masterdata management.

Eén essentieel verschil tussen traditionele ETL en data-integratie is de mogelijkheid om de real-time transacties af te kunnen tappen en deze gegevens meteen te laden in een bestand (data-warehouse), niet met een separaat product, met eigen metadata en een eigen gebruikersinterface, maar met één geïntegreerd product. In het verleden zijn vele datawarehouse initiatieven mislukt omdat de informatie die eruit kwam niet actueel genoeg was, en dus ook niet interessant. Het is moeilijk om de dagplanning van een fabriek te maken als je niet weet hoeveel van het personeel zich ziek c.q. beter heeft gemeld. Als de informatie pas morgen in het datawarehouse komt kan je nu geen beslissingen nemen omtrent de productiecapaciteit. Het is duidelijk dat niet alle informatie binnen vijf minuten beschikbaar hoeft te zijn, maar bepaalde beslissingen kunnen alleen genomen worden op basis van gegevens die wel volledig up-to-date zijn, en dat is een groot voordeel van data-integratie ten opzichte van het oude ETL. Het ETL-proces is al jaren onderschat qua ingewikkeldheid en dus ook qua kosten. Het lezen van 'legacy' bestanden waarvan zowel de inrichting als de inhoud onduidelijk is behoort tot de categorie 'uitdaging' als je van Amerikaanse komaf bent. Een Nederlander die van klare taal houdt noemt het meestal wat het is – een probleem. Er zijn enkele problemen; de belangrijkste betreft waarschijnlijk metadata, dat wil zeggen wat is de inhoud van een data-element en waar komt het vandaan. Het komt voor dat men in een bestand 'omzet' ziet en in een ander bestand 'revenue', die worden netjes bij elkaar opgeteld om te komen tot een nieuw data-element dat we 'totaal omzet' noemen. Het probleem is echter dat de omzet vermeld was in euro's en de revenue in dollars; 'totaal omzet' is daardoor een numeriek veld geworden waarvan de inhoud je geen informatie geeft om beslissingen te ondersteunen. Na heel veel slechte ervaringen op dit gebied zijn de meeste ETL-tools inmiddels voorzien van uitgebreide mogelijkheden om de kwaliteit van de data te controleren, de herkomst te registreren en door middel van profiling vreemde data te signaleren om eventuele verbeteringen aan te brengen. Het enige wat nu overblijft is deze faciliteiten gebruiken!
Cloud computing en SaaS (Software as a Service) zijn twee begrippen die veelal samen gaan. In de eerste plaats wordt deze combinatie bekeken vanuit een pricing perspectief. In de ETL-matrix hebben wij alle leveranciers gevraagd om een offerte uit te brengen voor twee hardware-configuraties. Eén vrij klein gebaseerd op een Windows server, de tweede middelgroot, met als basis een UNIX server. De prijsverschillen tussen leveranciers, voor zover men bereid was om prijzen te publiceren, waren tonnen, en dat is zonder hardware, installatie en configuratie. SaaS wordt over het algemeen gezien als een veel eerlijker prijsmodel, je betaalt voor wat je gebruikt en als het blijkt dat je meer capaciteit nodig hebt kan je gemakkelijk beschikken over meer capaciteit. De Cloud geeft ook veel meer flexibiliteit dan een eigen rekencentrum, zonder boetes voor outsourcing van de software, zonder aanschaf van de hardware, backups enzovoort. Wel moet men rekening houden met security, in veel gevallen gaat dit om bedrijfsgevoelige informatie en het is vaak moeilijk om uit te leggen dat je eigenlijk geen idee hebt van waar de data zich bevinden. Het is soms ook niet gemakkelijk om software van verschillende leveranciers te koppelen in de Cloud, wat de keus aanzienlijk kan beperken.
Masterdata management is ook een relatief nieuw fenomeen binnen de ETL/data-integratie infrastructuur. Een van de grootste datakwaliteitsproblemen komt door het feit dat bepaalde belangrijke gegevens op meer dan één plaats onderhouden worden. Als voorbeeld worden vaak NAW-gegevens gebruikt. Hier praten wij niet alleen over het feit of het adres correct is; dat wil zeggen dat het adres bestaat, het een juiste postcode heeft en de straatnaam correct is gespeld, maar het moet ook actueel zijn – de klant is niet verhuisd. Masterdata management kan zorgen dat sleutelgegevens in verschillende bestanden gesynchroniseerd worden en er één bron van de waarheid is.
Kijk op www.dbm.nl voor de geactualiseerde ETL-matrix.

Norman Manley is Managing Partner bij Passionned. Hij doet regelmatig surveys op de ETL-markt, zie www.etltool.com. De inhoud van DB/M's ETL-matrix is gebaseerd op deze onderzoeken.