EAI en ETL groeien minder snel naar elkaar toe

Trends in ETL-land en de wereld van data-integratie (1)

Daan van Beek

De fenomenen EAI, data-integratie en ETL mogen zich momenteel verheugen in meer dan normale belangstelling. Organisaties willen een geïntegreerde bedrijfsprocesomgeving en gegevens één keer invoeren en meerdere keren gebruiken. Dit bespaart tijd en kosten en levert een veel grotere betrouwbaarheid op van data, informatie en kennis in de organisatie.

Thomas Siebel, oprichter van Siebel, spreekt van grote besparingen als organisaties "de waardevolle data die door de hele organisatie in gescheiden silo's liggen opgeslagen" kunnen combineren (Computable, 12 dec. 2003). ETL-tools zijn specifiek in het leven geroepen om gegevens uit verschillende informatiesystemen binnen een organisatie te integreren en zodoende één overzichtelijke silo te bouwen op basis waarvan betrouwbare management-informatie is te realiseren. Dit artikel gaat in op de meest in het oog springende technologische ontwikkelingen van deze ETL-tools en biedt een overzicht van de belangrijkste trends in de markt.

Infrastructuur

Database Magazine en passionned onderzochten de op de markt beschikbare ETL-tools. Leveranciers die niet in het Gartner-Magic ETL-Quadrant van oktober 2003 (zie afbeelding 1) voorkomen,

> Een grid met servers draait niet alleen maar ETL-processen

deden niet aan het onderzoek mee. In het verleden is heel veel aandacht uitgegaan naar BI tools voor de eindgebruiker, de rapportage- en OLAP-gereedschappen. Terwijl de ontwikkeling en het onderhoud van ETL-processen de meeste tijd en kosten van BI-projecten in beslag nemen. Frank Buytendijk, research vice president bij Gartner, stelt: "70 tot 80 procent van de kosten van het opzetten van een BI-project zit 'm in de infrastructuur, zoals ontwerp, ETL en gegevenskwaliteit. Ieder uur dat je besteedt aan het inrichten van die infrastructuur, wordt later meervoudig terugverdiend." Daarom is onderzocht in hoeverre ETL-tools inspelen op specifieke behoeften die het succes en het rendement van BI-projecten zouden kunnen verhogen en de doorlooptijden ervan kunnen verkorten.

Het onderzoek richtte zich op de volgende thema's: realtime ETL en EAI, herbruikbaarheid en gebruiksvriendelijkheid, geheugenbeheer, versiebeheer, foutopsporing, datakwaliteit en modelgedreven ontwikkeling van ETL-processen. Voor aanvang van het onderzoek leefden de volgende verwachtingen ten aanzien van ETL-tools:

- ze convergeren in rap tempo met EAI-tools;
- ze worden volwassen ontwikkelomgevingen voor data;
- foutopsporing en het What You See Is What You Get (WYSIWYG)-principe worden gemeengoed;
- herbruikbaarheid en decompositie breken door;
- laadvensters van ETL-processen kunnen korter door slimmer geheugenbeheer;
- de leveranciers gaan modelgedreven ontwikkeling omarmen;
- controle op datakwaliteit wordt 'standaard' onderdeel.

Afbeelding 1: Het Gartner-Magic ETL-Quadrant van oktober 2003.

Om alvast een tipje van de sluier op te lichten: de 'werelden' van ETL en EAI zullen ze nog wel enkele jaren naast elkaar blijven bestaan. Tevens blijkt uit het onderzoek dat ETL-tools steeds meer gaan lijken op software-ontwikkelomgevingen met faciliteiten voor foutopsporing, WYSIWYG en componentgebaseerd ontwikkelen. Er zijn daarnaast plannen gesignaleerd om ETL-tools modelgedreven te laten werken. Transformaties en de complete ETL-architectuur genereert de ETL-tool dan op basis van business -specificaties en verwijzingen naar desbetreffende tabellen en attributen in het bronsysteem. In een volgend artikel zullen de leveranciers ten opzichte van elkaar gepositioneerd worden op verschillende criteria en wordt tevens een matrix getoond die per ETL-tool inzicht verschaft in alle belangrijke kenmerken zoals gebruiksvriendelijkheid, foutopsporing, ETL-procesondersteuning voor bijvoorbeeld slowly changing dimensions, server-grid-technologie, herbruikbaarheid en decompositie.

De trends

Bij aanvang van het onderzoek zijn zeventien leveranciers benaderd die in het Gartner-Magic ETL-Quadrant oktober 2003 staan vermeld. Eén leverancier is los van het Magic Quadrant meegenomen in het onderzoek. In totaal zijn dus achtien leveranciers benaderd waarvan er zestien aan het onderzoek hebben meegewerkt. Zij lieten zien hoe hun ETL-tools organisaties helpen om het ETL-proces te vereenvoudigen. Aan de hand van productpresentaties, interviews en de helpbestanden worden de volgende ontwikkelingen duidelijk.

Huwelijk tussen ETL en EAI is nog ver weg; real-time ETL breekt door.

Hoewel de werelden van ETL en Enterprise Application Integration (EAI) aan elkaar verwant zijn (het gaat immers bij beiden om data-integratie) en naar elkaar toegroeien, is er nog lang

De essentie van het ETL-proces

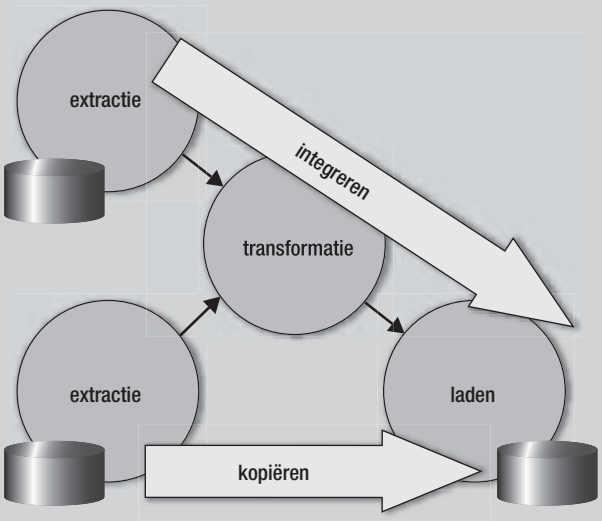
De afkorting ETL staat voor Extractie, Transformatie en Laden. Het proces van ETL begint bij de extractie van gegevens en eindigt bij het laden van de getransformeerde gegevens. Deze drie woorden geven weliswaar de essentie van ETL goed weer maar doen toch het ETL-proces tekort. Er gebeurt meer dan dat. Inherent aan het ETL-proces is het kopiëren van de gegevens van de ene machine naar de andere of van de ene database naar de andere. Daarnaast vindt er heel vaak door het ETL-proces integratie plaats van gegevens uit meerdere bronsystemen. Afbeelding 2 geeft het ETL-proces in essentie weer.

Tevens vindt er echter achter de schermen nog meer plaats, vooral waar het doel van de transformatie een datawarehouse-omgeving is. Er vindt dan vaak ook synchronisatie (van dimensies) en opschoning plaats om respectievelijk de gegevens met elkaar in overeenstemming te brengen en de kwaliteit van de gegevens te verbeteren. Bij elkaar komen we dan op de volgende formule:

E T (s o i) L : K

s = synchronisatie, o = opschoning, i = integratie, K = kopiëren.

Over Extractie zegt het woordenboek: het maken van een extract, een uittreksel. Dat is precies wat er tijdens een extractieproces ook gebeurt. Er wordt een uittreksel gemaakt van de gegevens op het bronsysteem, er wordt bijna nooit compleet gespiegeld. Sommige tabellen of kolommen zijn niet relevant (een horizontaal uittreksel op basis van de structuur of kolommen) en vaak worden bij aanvang bepaalde historische gegevens niet meegenomen (een verticaal uittreksel op basis van rijen).
Het woordenboek spreekt over transformatie als een gedaanteverandering, een verandering van vorm. In natuurkundige zin spreekt men van een omzetting van elektriciteit in licht. Deze heldere toelichting raakt precies het doel waarvoor organisaties ETL-transformaties in het leven roepen: het licht laten schijnen op opmerkelijke trends 'verborgen' in de bedrijfsgegevens.
De letter L uit ETL spreekt voor zich: het laden van de getransformeerde, geëxtraheerde gegevens in een database.
Het is niet voor niets dat dit artikel stilstaat bij de essentie van het ETL-proces. Dit proces slorpt maar liefst 70 tot 80 procent van de tijd, complexiteit en het budget van BI-projecten op en ETL-leveranciers weten zich daarin gesteund om hun gereedschappen aan de man te brengen. De complexiteit van het ETL-proces zit vooral in de letter T die het omzettingsproces van data tot informatie moet faciliteren.



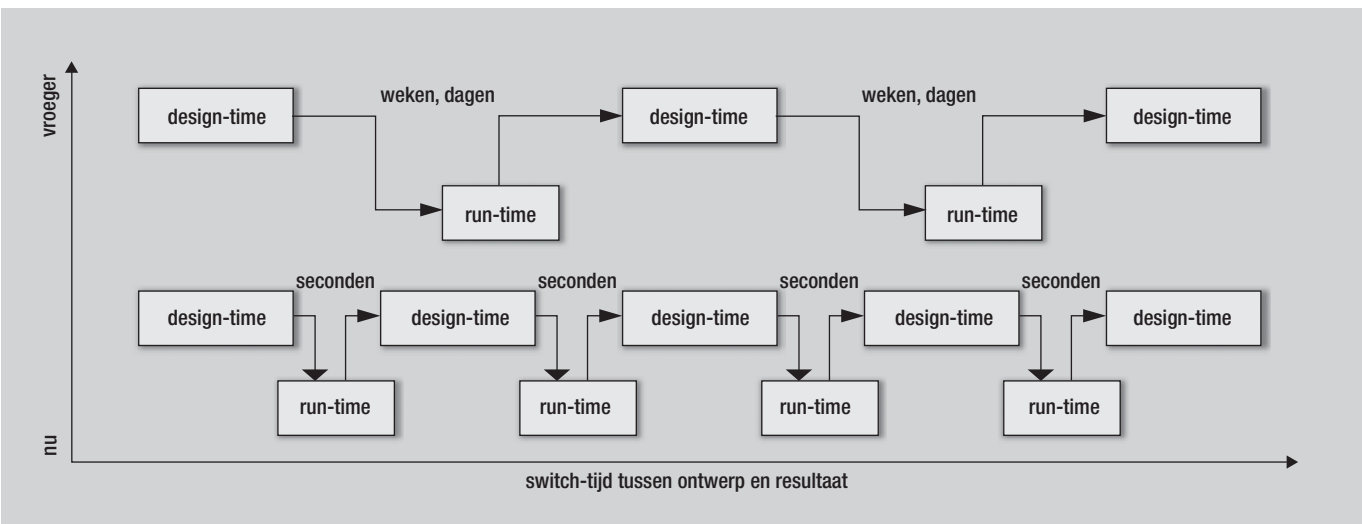
Afbeelding 2: Het ETL-proces in essentie is meer dan alleen extractie, transformatie en laden.

ETL Winter Survey 2004 (1)

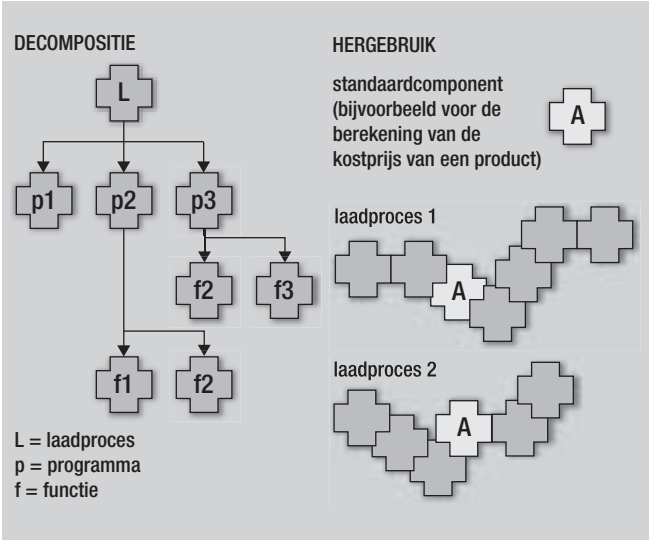
Daan van Beek van passioned deed voor DB/M een uitgebreid en indringend onderzoek naar ETL-tools. In de vele persoonlijke interviews met de vendors focuste hij op real-time ETL en EAI, herbruikbaarheid en gebruiksvriendelijkheid, WYSIWYG, geheugenbeheer, versiebeheer, foutopsporing, datakwaliteit, slowly changing dimensions en modelgedreven ontwikkeling van ETL-processen.
In dit nummer van DB/M worden de ontwikkelingen rond ETL tools uit de doeken gedaan; in het mei-nummer wordt ingezoomd op de kenmerken, overeenkomsten en verschillen van de individuele producten en presenteert DB/M u De ETL Matrix.
Op 8 juni 2004 organiseert Array een Expert Meeting over dit onderwerp, waarin aan de hand van het ETL Winter Survey 2004 de uitkomsten worden bediscussieerd door auteur Daan van Beek, Harm van der Lek en enkele leveranciers. Voor meer informatie: www.expertmeetings.nl

geen sprake van een daadwerkelijke versmelting van deze twee concepten. Babylonische spraakverwarring is hiervan de oorzaak. EAI en realtime ETL zijn vaak op één hoop gegooid en worden als één wereld beschouwd. Diverse ETL-gereedschappen bieden mogelijkheden om realtime berichten van wachtrijen af te lezen en te verwerken, maar is er dan sprake van *applicatie-* en *proces-*integratie zoals bij EAI?
Beide concepten zullen voorlopig naast elkaar blijven bestaan, ondanks dat 42 procent van de ETL-tools al mogelijkheden biedt voor realtime data-integratie. Zo brengt Microsoft naast haar EAI-tool BizTalk 2004, aan het einde van dit jaar een verbeterde versie van Data Transformation Services (codenaam Yukon) op de markt. Ook andere leveranciers zoals Ascential en Information Builders bieden nog gescheiden applicaties aan op dit vlak. Ten tijde van het onderzoek was er geen enkele ETL-leverancier die ETL én EAI aanbiedt binnen één omgeving. Dit geeft aan dat pure applicatie-integratie en ETL op dit moment nog niet echt samensmelten. Het blijkt dat EAI en ETL te snel op één hoop gegooid zijn. Hoever zijn we hier dan nog van verwijderd, klanten willen toch immers één complete omgeving voor data- en procesintegratie? Het groeit zeker naar elkaar toe, doordat enerzijds ETL-gereedschappen real-time mogelijkheden krijgen (mede voortgekomen uit het succes van XML en message queing) en anderzijds doordat EAI-gereedschappen faciliteiten krijgen voor analyse en Business Activity Monitoring (BAM). En als het huwelijk tussen ETL en EAI echt zo hard loopt hoe moet men het dan gaan noemen? Wie het weet mag het zeggen!

ETL-tools worden volwassen ontwikkelomgevingen voor data-integratie.
Programmeertalen en ontwikkelomgevingen bieden al veel langer mogelijkheden voor foutopsporing, versiebeheer, herbruikbaarheid, decompositie en What You See Is What You Get (WYSIWYG). ETL-tools worden ontwikkelomgevingen voor data met nagenoeg dezelfde faciliteiten als in de traditionele programmeertalen, echter het tempo waarin leveranciers deze faciliteiten aanbieden verloopt nog langzaam.
Foutopsporing en WYSIWYG hebben veel met elkaar te maken. Het doel is immers om in zo kort mogelijke tijd een foutloos programma op te leveren en dan liggen design- en run-time liefst niet te ver uit elkaar. Dit principe is weergegeven in afbeelding 3. Heel veel jaren geleden ging de programmeur in de weer met ponskaarten en zat weken kaartjes te 'ponsen' en eindelijk was het dan zo ver en moest hij in vijf minuten de ponskaarten er doorheen halen en afwachten wat het resultaat zou zijn.
Een klein foutje en hij moest weer weken wachten voordat hij opnieuw kon testen. We zien dat design- en run-time steeds dichter bij elkaar komen tot het moment dat we vrijwel direct het resultaat kunnen zien van wat we modelleren en programmeren. Het belang hiervan begint langzaam door te dringen tot de ETL-



Afbeelding 3: Door de tijd tussen design- en run-time te verkleinen kan sneller en met minder fouten een ETL-proces worden opgeleverd.



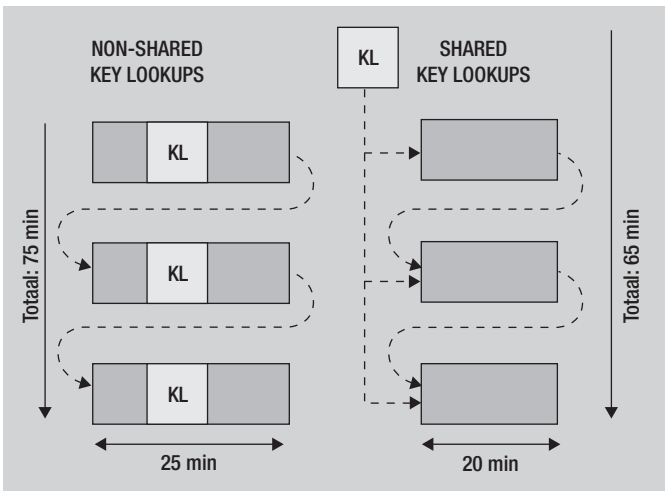
Afbeelding 4: Herbruikbaarheid en decompositie.

industrie. Grote hoeveelheden data, kenmerkend voor veel ETL-processen, zouden een belemmering vormen is een veel gehoord tegenargument, maar met kleine representatieve test-sets van gegevens is het niet moeilijk te verhelpen.

Foutopsporing en WYSIWYG is nog lang geen gemeengoed.
Het principe van WYSIWYG is voor tekstverwerkers een niet meer weg te denken fenomeen. De tijd dat we met allerlei control-codes letters moesten benadrukken of onderstrepen ligt ver achter ons. Dit principe is ook denkbaar én toepasbaar voor ETL. Bron- en doelgegevens zijn dan direct zichtbaar op het scherm en uiteraard ook het resultaat van de tussenliggende transformaties. Op dit moment omarmt slechts een klein aantal leveranciers van ETL-tools foutopsporing (25 procent) en het WYSIWYG-principe (19 procent), de rest werkt nog volledig op basis van metadata of men moet teveel handelingen verrichten om het resultaat te kunnen zien. Sommige leveranciers vinden dat zij het WYSIWYG-principe goed hebben toegepast door de code (bijvoorbeeld het SQL) te laten zien dat zij genereren. Het uiteindelijke resultaat, dat wat de extractie-, transformatie- en laadprocessen opleveren, is dan natuurlijk niet zichtbaar en het draagt dan ook niet echt bij aan een versnelling van de ontwikkeltijd van ETL-processen, noch aan de verbetering van de kwaliteit ervan. In sommige gevallen was het zelfs helemaal niet mogelijk om binnen de omgeving de data te bekijken tijdens het ontwikkelen.

Herbruikbaarheid, versiebeheer en decompositie breken door.
Veel leveranciers hebben versiebeheer, decompositie, herbruikbaarheid van componenten geïmplementeerd. Versiebeheer, dat in 56 procent van de ETL-tools is ingebouwd, is belangrijk voor het in productie nemen (en weer terug als het misloopt) van nieuwe processen, stappen, plannen enzovoort. Tevens zijn check-in- en check-out-mogelijkheden van groot belang in een multi user-omgeving. Herbruikbaarheid van logica en business rules wordt zo langzamerhand gemeengoed. Meer dan de helft van de ETL-tools (56 procent) heeft nu deze mogelijkheid in zich. Herbruikbaarheid is een zeer belangrijk fenomeen bij het ontwikkelen van software; het zorgt dat logica kan worden hergebruikt al dan niet met gebruik van parameters.
Hoe vaak komt het immers niet voor in een datawarehouse-omgeving, dat een ETL-proces een betekenisloze sleutel moet ophalen voor een product of dat de omzet tegen kostprijs én de voorraadwaarde tegen kostprijs uitgerekend moeten worden? Het is handig wanneer die logica eenmalig gedefinieerd kan worden en meerdere keren kan worden gebruikt. Define once, use multiple. Het scheelt tijd en fouten, hoewel de onderlinge afhankelijkheid van de ETL-processen dan wel toeneemt. Decompositie, het opdelen van het ETL-proces in verschillende stapjes, treffen we ook steeds vaker aan binnen ETL-tools (50 procent). Het gaat logisch goed samen met herbruikbaarheid. Decompositie maakt het ETL-proces inzichtelijk en vergemakkelijkt het onderhoud, doordat snel het bewuste stukje ETL-proces gevonden kan worden. Afbeelding 4 verduidelijkt de principes van herbruikbaarheid en decompositie.

Het laden van datawarehouses duurt onnodig lang.
ETL-tools danken voor een groot deel hun succes aan het snel kunnen laden en transformeren van gegevens vanuit meerdere bronnen. Een bekend instrument om snelle laadtijden te garanderen is het toepassen van zogenaamde key lookups waarbij delen van tabellen in het geheugen worden geladen. Hierdoor zijn bijvoorbeeld klant-, productsleutels of prijzen snel op te zoeken. Dit vermindert het aantal joins, versnelt de doorlooptijd en maakt het ETL-proces veel minder complex. Veel datawarehouse-omgevingen bevatten vaak meer dan één feittabel waar bijvoorbeeld een productdimensie aan vastzit. Dan is het handig wanneer de productdimensie vóór het laden van alle feittabellen in het geheugen kan worden gezet. Uiteraard nadat deze is bijgewerkt met eventuele productwijzigingen uit het bronsysteem.
Afbeelding 5 verduidelijkt dit principe. Het principe van shared



Afbeelding 5: Het verschil in laadtijd van een datawarehouse met normale key lookups of met zogenaamde shared key lookups kan wel oplopen tot 20 procent.

key lookups kan de laadtijd van een 'run' met wel 10 tot 20 procent inkorten. Toch zijn er op dit moment slechts twee ETL-tools die dit principe ondersteunen. Laadtijden van veel datawarehouses duren dus 'onnodig' lang.

Plannen voor modelgedreven ontwikkeling van ETL-processen. CASE, Computer Aided Software Engineering, genereert op basis van business-specificaties computer-programma's die bedrijfsprocessen ondersteunen. Dit wordt ook wel 5GL genoemd. Modelgedreven ontwikkeling van ETL-processen doet in principe hetzelfde. Het gaat hier dan niet om traditionele informatiesystemen of applicaties maar om datawarehouses en management-informatiesystemen of een deel ervan. Slechts één leverancier uit het onderzoek had concrete plannen om op die manier ETL-processen te genereren. Die plannen komen dicht in de buurt van de functionaliteit van Kalido, een product van een leverancier die de modelgedreven benadering voor datawarehouses heeft omarmt maar geen echte ETL-tool is. Kalido benodigd namelijk nog steeds een ETL-tool om de gegevens te verzamelen en te integreren. Het is om die reden dat het niet is opgenomen in het Gartner-Magic ETL-Quadrant.

> ETL-transformaties laten het licht schijnen op opmerkelijke trends 'verborgen' in de bedrijfsgegevens

Het modelgedreven ontwikkelen van ETL-processen kan grote voordelen hebben, net als bij CASE. In grote lijnen werkt het als volgt: definiëer de kernprestatie-indicatoren (KPI's) van een onderneming en bepaal daarbij de relevante dimensies en koppel die twee zaken aan de tabellen en attributen van de bronsystemen. Met wat aanvullende aanwijzingen is het dan mogelijk om op basis van specifieke kennis van datawarehouse-modellering om de extracties, transformaties, starschema's, aggregaties en laadprocessen in ruwe vorm te genereren. Complexe transformaties, business rules en kernprestatie-indicatoren zullen dan wellicht nog wel handmatig moeten worden aangepast en uitgebreid. Dit hangt voor het grootste gedeelte af van de complexiteit van de kernprestatie-indicatoren en de opzet van de structuur van de bronsystemen.

Datakwaliteit en ETL gaan steeds vaker samen. Gereedschappen die het ETL-proces ondersteunen, bevatten steeds meer transformaties en hulpmiddelen om de kwaliteit van de gegevens te verbeteren. Ongeveer 50 procent van de ETL-tools biedt de mogelijkheid om tijdens het laadproces gegevens op te schonen. Slechte kwaliteit van gegevens blijft een belangrijke oorzaak dat BI-projecten vertraging oplopen (en soms falen) en ETL-processen 'onnodig' complex maakt. Veel leveranciers van ETL-tools onderkennen dat en voorzien hun gereedschappen van allerlei transformaties en instrumenten om niet-juiste data te weren, te ontdubbelen, te routeren en daarover te rapporteren. De meest voorkomende transformaties zijn de Afwijzer, de Houder en de Ontdubbelaar. De Afwijzing wijst een rij af; de Houder zout een rij op, bijvoorbeeld totdat andere bijbehorende gegevens meekomen; de Ontdubbelaar *matched* de binnenkomende rij met de al aanwezige rijen en wijst in het geval er al een rij bestaat deze af. Omdat controle op datakwaliteit in sommige gevallen, zoals bij een Ontdubbelaar, een enorme aanslag pleegt op de doorlooptijd van het laadproces, bieden veel leveranciers ook een batchgewijze controlemogelijkheid aan. Eenmaal per dag of per uur worden dan de dubbele rijen verwijderd.

Overige opvallende onderzoeksresultaten

Server-grid-technologie: twee van de zestien leveranciers (13 procent) bieden server-grid-technologie aan. De doorlooptijden van ETL-processen kunnen hierdoor worden ingekort, IT-middelen worden op die manier beter benut en er vindt een betere verdeling plaats van de hoeveelheid werk. Dit lijkt op een gedistribueerde omgeving voor ETL, met dit verschil dat een grid met servers niet alleen maar ETL-processen draait maar ook ingezet kan worden voor andere taken. De vraag rijst of dit de stabiliteit en de betrouwbaarheid van ETL ten goede komt.
Slowly changing dimensions: het valt op dat nog zo weinig ETL-tools (31 procent) op dit moment expliciet slowly changing dimensions ondersteunen, omdat dit toch zo heel kenmerkend is voor data-integratie en datawarehouses. ETL-tools zijn toch juist hiervoor in het leven geroepen? ETL-tools die slowly changing dimensions expliciet ondersteunen hebben deze functie ingebakken (18 procent) of bieden die wizard-gestuurd (13 procent) aan.

Conclusie

De markt voor ETL-tools en de gereedschappen zelf worden zo langzamerhand volwassen, ook in technologisch opzicht.
De BI-industrie is serieus bezig om de markt zo goed mogelijk te ondersteunen bij het ontwikkelen van ETL-processen en data-integratie. Het meest opmerkelijke uit het onderzoek is wel dat ETL en EAI helemaal niet zo hard naar elkaar toegroeien als wel wordt gedacht. Real-time ETL is wel doorgebroken.
Daarnaast valt het op dat in het bijzonder de 'big players' van ETL hun producten laten evolueren tot volwaardige ontwikkelomgevingen voor data-integratie, waarvan real-time ETL, (ingebakken) afhandeling voor slowly changing dimensions, herbruikbaarheid en decompositie het meest in het oog springen. Tot slot kunnen nog veel leveranciers de laadtijden voor hun klanten verder inkorten door het principe van 'shared key lookups' in te bouwen en toe te passen.

Daan van Beek M.Sc (daanvanbeek@passionned.nl) is managing consultant voor passionned, een netwerk van project- en programmamanagers voor BI, data-integratie en -management, kennis-management en IT-strategie.