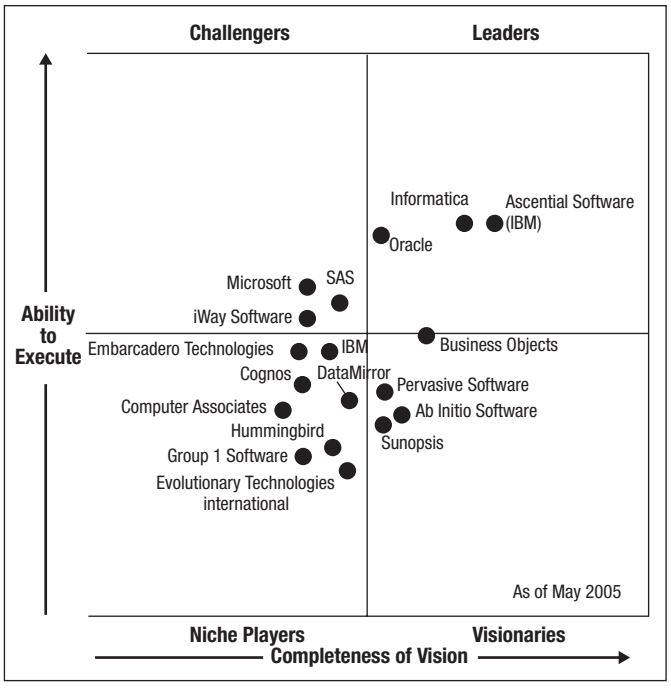
Van datawarehouse firewalls tot on-demand data integration

De ETL-matrix Reloaded 2005

Daan van Beek

Het grote nieuws in de ETL-markt is niet dat Microsoft binnenkort de markt gaat betreden met een compleet nieuwe versie van DTS, Integration Services genaamd, en mogelijk een dreiging zal vormen voor veel andere ETL-leveranciers. Hoewel het product qua functionaliteit opeens elf andere ETL tools achter zich laat, zijn er andere zaken die minstens van even groot belang zijn. Het gaat dan om on-demand data integration, out-of-the-box data auditing, datawarehouse firewalls en de verschuivingen ten opzichte van vorig jaar.

Het ETL Survey 2005 is in opdracht van DB/M door Passionned dit jaar voor de tweede keer uitgevoerd onder zeventien ETL tools. Naast het bijwerken van de bestaande vijftig kenmerken, zijn er in 2005 ruim twintig nieuwe kenmerken toegevoegd. Ook is de leverancier DataMirror nieuw binnengekomen in het onderzoek met het product Dynamic ODS[1]. Uitgangspunt voor opname in het ETL Survey 2005 was het ETL Magic Quadrant van Gartner, hier weergegeven voor de situatie van mei 2005.
Van de in dit artikel besproken ETL tools van Cognos en Sunopsis zijn rond het verschijnen van dit nummer nieuwe versies beschikbaar, respectievelijk DecisionStream 8 en Sunopsis V4.



Afbeelding 1: Het ETL Magic Quadrant van Gartner van mei 2005.

Beide leveranciers konden ons helaas nog geen kijkje in de keuken geven. De scores van deze leveranciers geven daarom waarschijnlijk een licht vertekenend beeld.
Een aantal opmerkelijke ontwikkelingen die verband houden met data-integratie, datawarehousing en Enterprise Application Integration (EAI) wordt hierna besproken.

On-demand data integration maakt het ODS overbodig

Slapende gegevens (gegevens die niet of nauwelijks worden gebruikt) in het datawarehouse zijn uit den boze. Deze nemen onnodig ruimte in beslag en vertragen dikwijls ook de responstijd. Daarom nemen we doorgaans niet alle gegevens die we beschikbaar hebben in de bronsystemen op. Vaak beperken we ons tot die gegevens die echt van belang zijn voor besluitvorming, een integraal klantbeeld en prestatieverbetering.

Toch bestaat de wens om na de analyse- en rapportagefase detailgegevens bij de hand te hebben, zodat snel actie kan worden genomen. Een voorbeeld: een groothandel ziet zich door concurrentie-ontwikkelingen genoodzaakt om het kapitaalbeslag van goederen op voorraad te verminderen. Na gedegen omzetanalyse besluit de logistieke directie het assortimentsbeleid aan te passen en artikelen die minder dan vijf maal op jaarbasis geraakt worden, niet meer op de plank te leggen. De leveranciers van desbetreffende artikelen zullen geïnformeerd moeten worden. Dan is het van belang om de artikelnummers van de leveranciers, de adresgegevens en overige gedetailleerde productinformatie ter beschikking te hebben. Deze informatie is niet volledig beschikbaar in het datawarehouse, met als gevolg dat men deze gegevens weer in het operationele systeem moet zien op te duikelen. Lastige conversies, uploads naar het operationele systeem en tijdverlies zijn het gevolg.

Met on-demand data integration worden alle relevante gegevens van de organisatie vastgelegd in een metadata-laag, bijvoorbeeld van de ETL tool. Voor elk gegeven stelt men vervolgens vast hoe deze getransformeerd moet worden en of deze opgenomen moet worden in het datawarehouse. Na realisatie en vulling van het datawarehouse, de metadata-laag en de ETL-processen, worden alle gegevens door middel van web services, een rapportage-oplossing of Enterprise Information Portal, aan de gebruiker aangeboden.
Bij selectie van gegevens zorgt de webservice ervoor dat de ETL tool de niet in het datawarehouse opgeslagen gegevens real-time ophaalt uit het bronsysteem. Er wordt dan op de achtergrond een ETL-proces gestart. Sommige organisaties hebben voor gedetailleerde informatie, zoals beschreven in bovenstaand voorbeeld, een Operational Data Store (ODS) in het leven geroepen. Met on-demand data integration is er nauwelijks meer noodzaak om een relatief kostbaar ODS te bouwen of te onderhouden. Voorwaarde is wel dat het operationele systeem niet teveel belast wordt en de verschillende componenten van deze architectuur goed samenwerken. Een on-demand architectuur ondersteunt het principe van closed-loop en operational Business Intelligence, waarbij de medewerkers in één moeite door op basis van informatie en kennis snel actie kunnen ondernemen om tot slot deze acties weer te evalueren en te monitoren. Momenteel bieden zeven ETL tools, waaronder Informatica PowerCenter, IBM Websphere Data Stage en SAS ETL Studio, deze mogelijkheid. Van supergeavanceerd tot eenvoudig.

De spiegel van de bron: data auditing

Een belangrijk punt van aandacht bij het bouwen van een datawarehouse is om ervoor te zorgen dat de gegevens in het datawarehouse exact overeenkomen met de gegevens uit de bronsystemen. Bij het transformeren van gegevens kan immers genoeg mis gaan, in het bijzonder als bestanden aan elkaar worden gekoppeld. Om te controleren of de gegevens kloppen, gebruikt men vaak 'hashing' tabellen. Tijdens de extractie sommeren we dan – van de rijen die in het datawarehouse opgenomen moeten worden – een aantal numerieke attributen bijvoorbeeld per klant of per product. Deze totalen hevelen we dan – met de gegevens zelf – over naar de datawarehouse-omgeving. Nadat het datawarehouse is geladen, berekenen we de totalen opnieuw, maar nu in het datawarehouse zelf. Deze totalen vergelijken we met de tijdens de extractie berekende totalen en bij verschillen zoeken we een verklaring. Bij het ontwerpen, bouwen en testen van deze constructie was altijd veel tijd gemoeid. Daarom bieden sommige ETL tools – waaronder Data Integrator – de mogelijkheid tot 'data auditing', waarbij vrijwel automatisch controle-totalen worden berekend en vergeleken met de datawarehouse-totalen.

De Datawarehouse Firewall: 'ongure types' blijven buiten

Organisaties dienen idealiter, voordat men een datawarehouse bouwt, inzicht te krijgen in de kwaliteit van de gegevens. Geen inzicht hierin kan leiden tot vervelende vertragingen tijdens het bouwproces. Gelukkig bieden steeds meer ETL tools *out-of-the-box* (en soms zonder meerkosten) de mogelijkheid om de brongegevens van te voren te inspecteren op dubbele waarden, aantallen, uitersten, gemiddelden, relaties, afwijkingen en andere doublures. Dit noemen we data profiling, waarbij gaandeweg een profiel ontstaat van de data. Mede op basis van het profiel kunnen we het datawarehouse gaan inrichten, het gegevensmodel vaststellen, business rules definiëren en het ETL-proces ontwerpen. De business rules kunnen bij sommige ETL tools automatisch worden gegenereerd en vormen dan een zogenaamde Datawarehouse Firewall: alle data waaraan iets mankeert blijven buiten de deur. Uiteraard is het dan wel zaak om de afgekeurde gegevens nader te onderzoeken en indien nodig actie te ondernemen, bijvoorbeeld opschonen en herladen.

Huwelijk ETL-EAI? Zelfs nog geen verkering!

Vorig jaar bleek al dat nog weinig ETL tools echte faciliteiten bevatten om EAI-projecten uit te voeren. Hoewel inmiddels veel ETL tools mogelijkheden bieden voor real-time data-integratie, is er nog lang geen sprake van de zogenaamde 'convergence' tussen ETL tools en EAI tools. En een huwelijk? De partijen hebben elkaar amper leren kennen en hebben niet eens verkering. Laat staan dat organisaties EAI-projecten met een ETL tool gaan uitvoeren of andersom. Of nog erger: alles op één hoop gooien wat betreft data-integratie.
Natuurlijk gaan de tools beter samenwerken, zo kan Informatica's ETL tool gegevens van WebMethods lezen, en biedt IBM Websphere DataStage de mogelijkheid om gebruik te maken van faciliteiten van DataStage TX. Maar de tools zullen ieder op zich blijven bestaan en zeker niet in elkaar opgaan. Wel zullen we zien dat deze tools steeds meer gebruik gaan maken van gemeenschappelijke metadata, transformaties en connectoren.
Tot slot, de aard van ETL en BI verschilt enorm met die van EAI. Op technisch vlak: bij ETL is er sprake van bulk, integratie en convergentie, bij EAI is er sprake van grote hoeveelheden transacties met weinig data, distributie en divergentie. Op het vlak van de bedrijfsvoering: ETL gebruiken we voor datamigratie, data-integratie voor besluitvorming en een integraal klantbeeld, EAI gebruiken we voor het optimaliseren van werkstromen en eenmalige gegevensinvoer. Kortom; het huwelijksbootje moet nog gemaakt worden en de rivier zelfs nog gegraven. Nu een EAI tool adviseren voor ETL-taken is hetzelfde als zeggen dat goede managementinformatie rechtstreeks uit het bronsysteem kan komen en dat datawarehouses overbodig zijn. ETL tools en EAI tools moeten we meer zien als broer en zus en die trouwen doorgaans niet met elkaar. Ze gaan als ze volwassen zijn juist ieder hun weg, maar houden doorgaans goed contact met elkaar.

De ETL-matrix

In dit artikel presenteren we de ETL-matrix, met verticaal alle productkenmerken, en horizontaal de verschillende ETL tools. De matrix bevat een aantal onderdelen: Bedrijf, Tool, Gebruiks-

vriendelijkheid, Overzichtelijkheid en herbruikbaarheid, Foutopsporing, Real-time ETL/EAI/Webservices, Functionaliteit, Data sources/targets, Architectuur en infrastructuur, en Berekeningen. In dat laatste onderdeel wordt de puntentelling bepaald die in de grafieken is gebruikt. Hierna worden grafieken getoond en besproken die de ETL tools rangschikken naar compleetheid, gebruiksvriendelijkheid, connectiviteit, platformondersteuning, groeipotentie en prijs per functionaliteit. Voor iedere grafiek is de formule voor de berekening van de rangschikking opgenomen.

Compleetheid

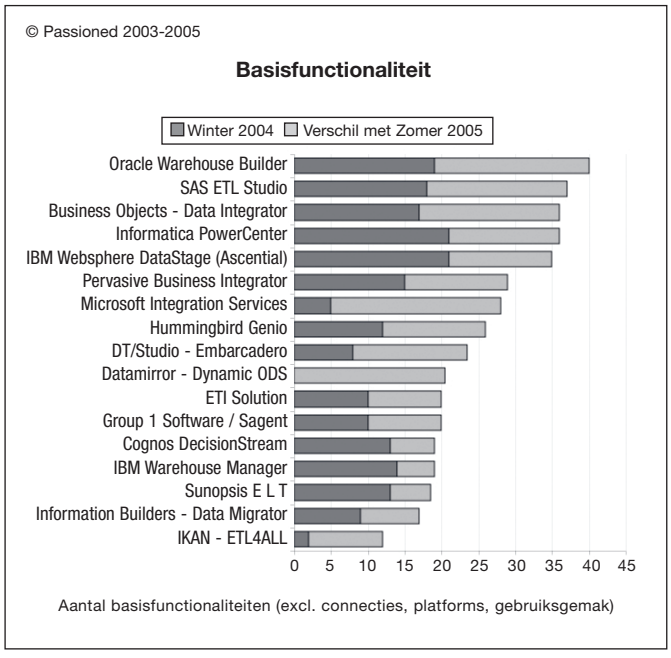
In afbeelding 2 staan de ETL tools gerangschikt op compleetheid. Dit is berekend door ieder kenmerk met een positief antwoord ('Ja') één punt toe te kennen en vervolgens een optelling te maken. Het kenmerk 'Slowly changing dimensions' is daarop een uitzondering, als ETL tools dat handmatig ondersteunen krijgen ze nul punten, ondersteunen ze het met een *wizard* dan krijgen ze één punt, is die functionaliteit ingebakken (out-of-the-box) dan krijgen ze twee punten. In sommige gevallen worden halve punten weggeven; bij 'half' is het kenmerk dan niet volledig beschikbaar of 'ja, db' betreft dan functionaliteit die alleen in combinatie werkt met de database, zoals het geval is bij Oracle Warehouse Builder en IBM Warehouse Manager. Het maximale aantal punten dat kan worden behaald is 45. De bijbehorende kenmerken zijn in de legenda en matrix gemarkeerd met een sterretje.

Gebruiksvriendelijkheid

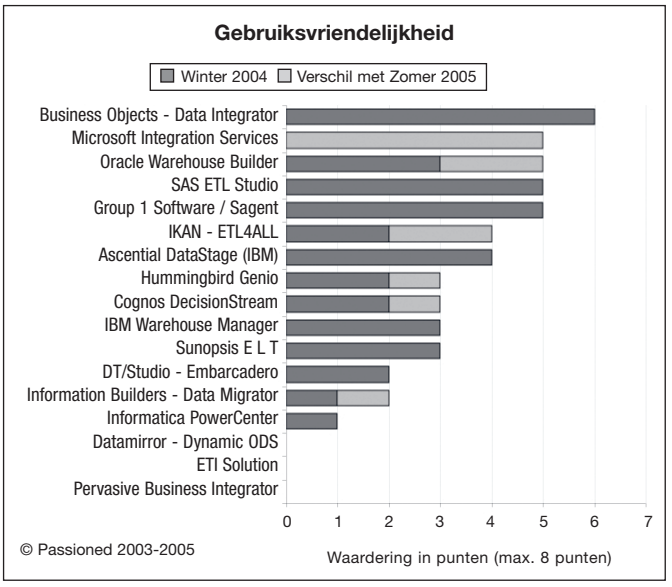
Het meten van gebruiksvriendelijkheid is niet altijd objectief vast te stellen. Toch hebben we gezien het belang van gebruiksvriendelijkheid bij de vaak toch al complexe ETL-processen een aanzet gegeven om dat zo objectief mogelijk te meten. Daarbij is gelet op hoe gemakkelijk een ETL-proces kan worden ontwikkeld, vormgegeven en onderhouden, wat de inleertijd is van de tool, of de gebruikersinterface uitnodigt tot verkennen, of het schermontwerp rustig is en in balans (symmetrie en de zogenaamde leeslijnen), of een gebruiker dezelfde handelingen steeds moet herhalen, of de data tijdens het ontwikkelen kunnen worden ingezien en of de tool de taken van de ETL-ontwikkelaar in de juiste volgorde ondersteunt. Het resultaat daarvan is weergegeven in afbeelding 3. Het maximale aantal punten dat een ETL tool op dit onderdeel kon behalen was acht, voor ieder van de in totaal vier kenmerken maximaal twee punten, aangegeven met twee plusjes (++). Voor een negatief oordeel werden punten afgetrokken, bijvoorbeeld als bij gebruiksgemak, WYSIWYG en taakcompatibiliteit ETL/EAI een plusje staat en bij schermontwerp een minnetje, krijgt de tool twee punten (drie punten voor ieder plusje minus één punt voor het minnetje).

Connectiviteit

In afbeelding 4 staan de ETL tools gerangschikt naar connectiviteit, de mate waarin zij verschillende soorten bronnen native, dus zonder tussenkomst van ODBC of OLE-DB, kunnen lezen en/of schrijven. Er is gekeken naar het lezen en schrijven van verschillende typen bronnen zoals databases, XML-documenten, wachtrijgegevens en Enterprise Applications zoals Siebel, SAP en dergelijke. De connectiviteitscore is een optelling van het aantal bronnen, het aantal enterprise applications van waaruit de metadata kunnen worden ingelezen en het aantal realtime gegevenswachtrijproducten dat kan worden gelezen.



Afbeelding 2: Oracle Warehouse Builder (OWB), SAS ETL Studio, Data Integrator en PowerCenter bieden de meeste basisfunctionaliteiten aan en ETL4ALL de minste basisfunctionaliteiten. OWB en SAS hebben Ascential en Informatica van de eerste respectievelijk tweede plaats verdrongen.



Afbeelding 3: De rangschikking naar gebruiksvriendelijkheid. Business Objects gooit hier voor het tweede jaar de hoogste ogen. Microsoft Integration Services heeft ten opzichte van DTS subliem werk verricht en komt op een tweede plaats te staan. DataMirror, ETI Solution en Pervasive behalen geen punten voor dit onderdeel, grotendeels veroorzaakt door hun programmeerachtige omgeving.

Platformondersteuning

De rangschikking naar platformondersteuning is weergegeven in afbeelding 5. Alle verschillende versies van Windows zijn bij elkaar geteld. De diverse smaken van Unix zijn afzonderlijk gewaardeerd. Meestal vergt het meer inspanning om, vooral op het gebied van het beheer, de verschillende smaken van Unix te ondersteunen dan de verschillende versies van Windows.

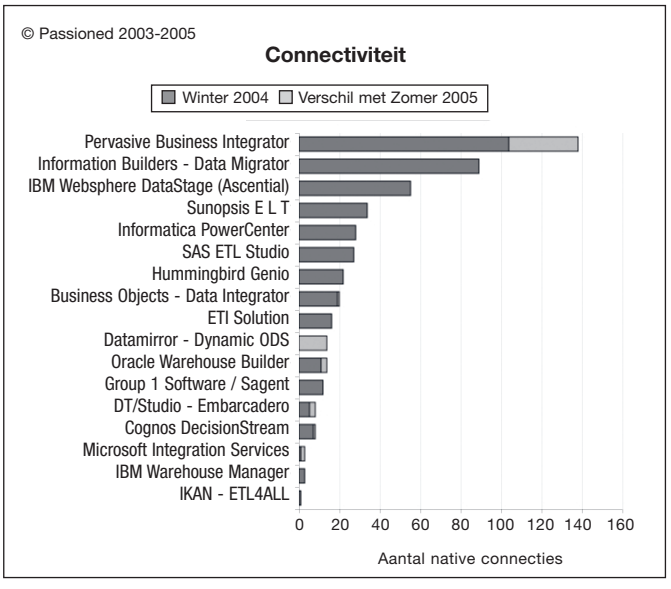
Groeipotentie

Kijken we naar de verhouding tussen het aantal jaren op de markt en het totale aantal functionaliteiten (inclusief gebruiksvriendelijkheid, connectiviteit en platformondersteuning), dan kunnen we dat vertalen naar de groeipotentie van de tool. Welke tools zullen de komende jaren de markt gaan of blijven domineren? De tools DT/Studio en ETL4ALL zijn hier niet in meegenomen, omdat ze nog maar 'net' op de markt zijn en dat zou hen wellicht onterecht bovenaan plaatsen.

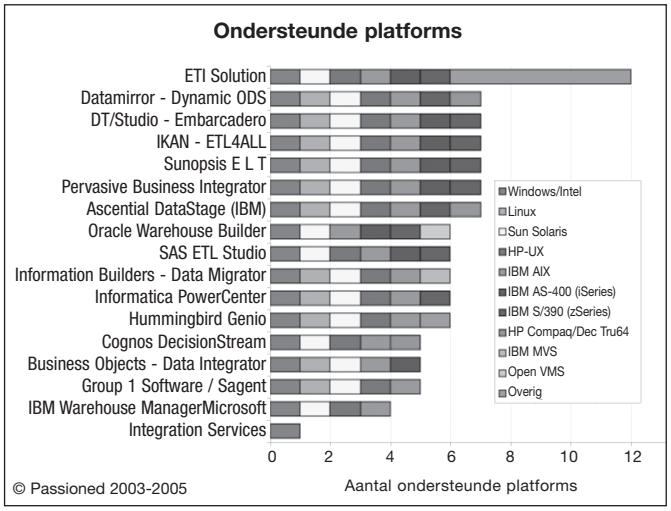
Veel nieuwe functionaliteiten, een hoge gebruiksvriendelijkheid, een redelijke prijs, een hoge connectiviteit en een goede ondersteuning van de verschillende platforms, zouden moeten resulteren in marktdominantie. Wie deze positie de komende jaren zou kunnen bereiken is uitgewerkt in de conclusie.

Verhouding prijs-functionaliteit

In afbeelding 7 is de verhouding weergegeven tussen de prijs van twee configuraties (zie legenda) en de totale functionaliteit. Ook hier hebben we de gebruiksvriendelijkheid, de connectiviteit en de platformondersteuning meegerekend. Leveranciers spannen zich immers niet alleen in om louter functionaliteit te leveren,



Afbeelding 4: De ETL tools gerangschikt naar connectiviteit. Pervasive met Business Integrator en Information Builders met Data Migrator (voorheen ETL Manager) behalen veel punten. Microsoft met Integration Services, IBM met WarehouseManager en IKAN met ETL4ALL zijn minder geschikt voor heterogene omgevingen met veel verschillende soorten gegevensbronnen. Ten opzichte van vorig jaar zijn er nauwelijks wijzigingen.



Afbeelding 5: De rangschikking van de ETL tools naar platformondersteuning. ETI ondersteunt de meeste platforms, Microsoft met Integration Services het minste aantal. Wat opvalt is dat de meerderheid Linux ondersteunt.

maar natuurlijk ook om deze makkelijk te kunnen bedienen onder verschillende infrastructuren en in heterogene IT-omgevingen.

Analyse ETL-markt: stilte voor de storm?

De markt voor ETL is continu in beweging maar doet toch nog redelijk vertrouwd aan. Ascential is dit jaar weliswaar overgenomen door IBM, maar voor de rest is het vrij rustig aan het front. Na een avontuur met PowerAnalyzer concentreert Informatica zich weer op (enterprise) data-integratie en Microsoft is al vijf jaar aan het werk voor de nieuwe DTS. Is het stilte voor de storm? Kan Microsoft net als met Analysis Services binnen enkele jaren een marktaandeel behalen van meer dan twintig procent? Is het succes van Analysis Services te kopiëren?

Als we kijken naar de geboden functionaliteit van Integration Services, de gebruiksvriendelijkheid en integratie met Visual Studio, dan moeten we tot de conclusie komen dat men een sterke troef in handen heeft. Microsoft kan straks een Enterprise Business Intelligence Suite (EBIS) leveren, waardoor klanten *one-stop-shopping* kunnen doen voor relatief weinig geld. Het zal op termijn gevolgen hebben en niet alleen voor de onderkant van de ETL-markt. Toch zal Integration Services voor het overgrote deel van de ETL-producten niet al te grote gevolgen hebben, voorlopig althans. De BI- en ETL-markt groeien nog stevig, veel ETL servers draaien op Unix of Linux en veel ETL-producten maken inmiddels onderdeel uit van een Enterprise Business Intelligence Suite.

Moet Informatica zich dan zorgen maken? Of Sunopsis? Als er op korte termijn al leveranciers last krijgen van Microsoft dan denk ik eerder aan Embarcadero (ontbreken EBIS, hoewel deze wel een sterke marktpositie heeft in databeheer- en datamodelleer-tools), IKAN (net op de markt en relatief weinig functionaliteit), Group 1 Software (geen BI-speler en verhouding prijs-functionaliteit),

	Pervasive Business Integrator	Sunopsis E L T	IBM Websphere DataStage (Ascential)	IBM Warehouse Manager	Cognos DecisionStream	Group 1 Software / Sagent	Informatica PowerCenter	Information Builders - Data Migrator	Business Objects - Data Integrator	SAS ETL Studio
Bedrijf										
Gestart met verkoop	1985	1998	1996	1993	1993	1996	1996	1996	1998	1996
Klanten WW	35000	420	3500	disclosed	2000	1700	2250	400	925	1200
Installaties WW	50000	1900	>3500	disclosed	3100	2100	5000	800	>925	1950
Installaties NL	10	3	45	disclosed	38	18	200	48	>30	55
Vestiging in Benelux	Brussel	ja	ja	ja	ja	ja	ja	ja	ja	ja
Tool										
Platforms	7	7	7	5	5	4	6	6	5	6
Versie	8.6	v3	7.5.x	8.2	7.2	5.5	7.1.2	7.1	11	v3.3
Engine-based / codegenerator	eb	cg	eb	cg	eb	eb	eb	cg	eb	cg
Type	map	map	proces	proces	proces	proces	map/proces	map/proces	proces	proces
Prijs configuratie I (x 1.000 euro)	€ 36	€ 44	€ 93	€ 60	€ 99	€ 105	€ 115	€ 104	€ 68	€ 218
Prijs configuratie II (x 1.000 euro)	€ 72	€ 89	€ 141 (schatting)	€ 80	€ 187	€ 285	€ 131 (schatting)	€ 206	€ 91	€ 342
Gebruiksvriendelijkheid										
- gebruiksgemak	-	+	+	+	+	++	0	+	++	++
- WYSIWYG	+	0	+	0	0	+	-	0	+	0
- schermontwerp	0	+	+	+	+	+	+	+	++	++
- taakcompatibiliteit ETL / EAI	-	+	+	+	+	+	+	0	+	+
[subscore]	-1	3	4	3	3	5	1	2	6	5
Overzichtelijkheid en herbruikbaarheid										
- herbruikbaarheid componenten*	ja	ja	ja	ja	half	ja	ja	ja	ja	ja
- decompositie*	ja	nee	ja	ja	ja	ja	ja	nee	ja	ja
- user-defined functies*	ja	ja	ja	ja, db	ja	ja	ja	nee	ja	ja
- commentaar selectie van objecten*	nee	nee	ja	nee	nee	nee	nee	nee	nee	nee
[subscore]	3	2	4	2,5	2,5	3	3	1	3	3
Foutopsporing										
- step by step running*	ja	nee	ja	nee	nee	nee	ja	nee	ja	nee
- row-by-row running*	ja	nee	ja	nee	nee	nee	ja	nee	ja	nee
- breakpoints*	ja	nee	ja	nee	nee	nee	ja	nee	ja	nee
- watches*	nee	nee	ja	nee	nee	nee	ja	nee	ja	nee
- compiler/validate*	ja	ja	ja	nee	ja	nee	ja	nee	ja	ja
[subscore]	4	1	5	0	1	0	5	0	5	1
Realtime ETL/EAI/Web services										
- integratie batch - realtime*	ja	ja	ja	nee	nee	nee	ja	nee	ja	ja
- mechanismen*	mq+trig	mq+trig	mq	n.v.t.	n.v.t.	n.v.t.	mq+log	n.v.t.	mq	mq
- on-demand data integration*	ja	nee	ja	nee	nee	nee	ja	ja	ja	ja
[subscore]	4	3	3	0	0	0	4	1	3	3
Functionaliteit										
- splitting datastreams/multiple targets*	ja	nee	ja	ja	ja	ja	ja	ja	ja	ja
- conditional splitting*	ja	nee	ja	ja	ja	ja	ja	ja	ja	ja
- union*	nee	nee	ja	ja	ja	ja	ja	nee	ja	ja
- pivoting*	ja	nee	ja	ja	ja	ja	ja	nee	ja	ja
- depivoting	ja	nee	nee	nee	nee	ja	nee	nee	ja	nee
- key lookup's in memory*	ja	ja	ja	ja, db	ja	ja	ja	nee	ja	ja
- key lookup's herbruikbaar over proces*	ja	ja	nee	ja, db	nee	nee	nee	nee	nee	ja
- slowly changing dimensions*	hm	auto	hm	hm	auto	hm	wizard	auto	auto	auto
- scheduler*	ja	ja	ja	ja, db	ja	ja	ja	ja	ja	ja
- status afhandeling binnen job*	ja	nee	nee	ja	nee	nee	nee	nee	ja	ja
- impact analysis*	ja	ja	ja	ja	ja	ja	ja	ja	ja	ja
- data lineage*	nee	nee	ja	ja	half	nee	ja	ja	ja	ja
- automatische documentatie*	ja	ja, pdf	ja	nee	ja	nee	ja	nee	nee	ja
- support voor data-miningmodels*	nee	nee	nee	ja, db	nee	nee	ja	nee	nee	ja
- support voor analytische functies*	nee	nee	nee	ja, db	nee	ja	nee	nee	nee	ja
[subscore]	10	6,5	9	9,5	10,5	9	11	7	12	15
Data sources/targets										
- support voor joined tables als bron*	ja	ja	ja	ja	nee	ja	ja	ja	ja	ja
- ingebouwde functies voor data kwaliteit*	nee	ja	ja	nee	ja	ja	ja	nee	ja	ja
- ingebouwde functies voor data validatie*	ja	nee	ja	nee	ja	nee	ja	nee	ja	ja
- data profiling*	ja	nee	ja	nee	nee	nee	ja	nee	ja	ja
- changed data capture*	nee	ja	ja	nee	nee	ja	ja	ja	ja	ja
[subscore]	3	3	5	1	2	3	5	2	5	5
- native connections (-ODBC -flatfiles)	130	21	41	3	7	6	16	83	6	18
- packages / enterprise applications	3	7	6	0	1	6	5	5	5	6
- real-time connecties	5	6	8	0	0	0	7	1	9	3
Architectuur en infrastructuur										
Parallel processing										
- Symmetric Multiprocessing (SMP)*	ja	ja	ja	ja, db	nee	ja	ja	ja	ja	ja
- Massively parallel processing (MPP)*	nee	ja, db	ja	ja, db	nee	nee	ja	nee	nee	ja
- Cluster Aware*	nee	nee	ja	ja, db	nee	nee	nee	ja	ja	ja
- Grid*	nee	nee	ja	ja, db	nee	nee	ja	nee	nee	ja
Schaalbaarheid										
- Job distributie*	ja	ja	ja	ja	ja	ja	ja	ja	ja	ja
- Data pipelining*	nee	nee	ja	ja, db	nee	ja	nee	nee	ja	ja
- Partitionering*	ja	nee	ja	ja, db	nee	ja	ja	ja	ja	ja
Basisarchitectuur	h	h+d	h+m	h+d	h	h	h+m	h+d+m	h	h+d
- end-to-end BI infrastructuur*	nee	half	nee	ja	ja	ja	ja	ja	ja	ja
- CWM-ondersteuning*	ja	nee	ja	ja	nee	nee	ja	nee	ja	ja
- versiebeheer*	ja	nee	ja	nee	ja	nee	ja	ja	ja	ja
[subscore]	5	3	9	6	3	5	8	6	8	10
Berekeningen										
Aantal basisfunctionaliteiten	29	18,5	35	19	19	20	36	17	36	37
Gebruiksvriendelijkheid	-1	3	4	3	3	5	1	2	6	5
Aantal ondersteunde platforms	7	7	7	5	5	4	6	6	5	6
Aantal ondersteunde sources/targets	138	34	55	3	8	12	28	89	20	27
Aantal jaren op de markt	20	7	9	12	12	9	9	9	7	9

IKAN - ETL4ALL	ETI Solution	Oracle Warehouse Builder	Hummingbird Genio	DT/Studio - Embarcadero	Microsoft Integration Services	Datamirror - Dynamic ODS
2003	1993	1998	1996	2002	1997	1995
80	370	disclosed	400	200	unknown	50
250	5000	disclosed	400	200	unknown	200
12	15	disclosed	5	15	unknown	3
ja	nee	ja	ja	ja	ja	ja
7	12	6	6	7	1	7
3.0	5.2.2	10gR2	6	2.3.2	2005	3.7
eb	cg	cg	eb/cg	eb	eb/cg	cg
proces	map	proces	proces	proces	proces	map
€ 7	disclosed	€ 8	€ 60	€ 39	€ -	€ 83
€ 11	disclosed	€ 12	€ 140	€ 65	€ -	€ 101
+	0	++	+	+	+	0
0	-	0	0	-	++	-
++	+	++	+	+	+	0
+	0	+	+	+	+	0
4	0	5	3	2	5	-1
ja	nee	ja	ja	ja	ja	ja
ja	nee	ja	ja	ja	nee	nee
nee	nee	ja	ja	ja	ja	ja
nee	nee	nee	nee	nee	nee	nee
2	0	3	3	3	2	2
nee	nee	ja	nee	ja	ja	nee
nee	nee	ja	nee	ja	nee	nee
nee	nee	ja	nee	ja	ja	nee
nee	nee	ja	nee	ja	ja	nee
ja	nee	ja	ja	ja	ja	nee
1	0	5	1	5	4	0
nee	ja	ja	ja	ja	nee	ja
n.v.t.	mq	mq+log+trig	mq	mq	n.v.t.	mq+log+trig
nee	nee	ja	nee	nee	?	nee
0	2	5	2	2	0	4
ja	ja	ja	ja	ja	ja	ja
ja	ja	ja	ja	ja	ja	ja
ja	nee	ja	ja	ja	ja	ja
ja	nee	ja	nee	ja	ja	nee
ja	nee	ja	nee	ja	ja	nee
nee	ja	ja, db	ja	nee	ja	nee
nee	nee	ja, db	nee	nee	nee	nee
hm	hm	auto	hm	hm	auto	hm
nee	nee	ja	ja	ja	ja	ja
nee	nee	nee	ja	nee	ja	ja
nee	ja	ja	ja	ja	ja	ja
nee	ja	ja	nee	nee	ja	nee
nee	ja	ja	ja	ja	nee	ja, pdf
nee	nee	ja	nee	half	ja	nee
5	6	13,5	9	8,5	14	6,5
nee	ja	ja	ja	nee	ja	ja
nee	nee	ja	ja	nee	ja	nee
nee	ja	ja	nee	nee	nee	ja
nee	nee	ja	nee	nee	nee	nee
nee	ja	ja	ja	ja	nee	ja
0	3	5	3	1	2	3
0	12	7	18	6	1	5
1	2	4	2	1	1	6
0	2	3	2	1	1	3
ja	ja	ja	ja	ja	ja	ja
nee	ja	ja	nee	nee	nee	nee
ja	ja	ja	ja	nee	nee	nee
nee	ja	ja	nee	nee	nee	ja
ja	ja	ja, db	ja	ja	ja	ja
nee	ja	ja, db	ja	ja	ja	ja
nee	ja	ja, db	ja	nee	ja	ja
h	h+d	h+d	h+d	h	h	h+m
nee	nee	ja	ja	nee	ja	nee
ja	ja	ja	ja	ja	nee	nee
nee	ja	ja	ja	nee	ja	nee
4	9	8,5	8	4	6	5
12	20	40	26	23,5	28	20,5
4	0	5	3	2	5	-1
7	12	6	6	7	1	7
1	16	14	22	8	3	14
2	12	7	9	3	8	10

Bedrijf
- Gestart met verkoop: In welk jaar kwam het product voor het eerst op de markt?
- Klanten WW: Het aantal klanten wereldwijd
- Installaties WW: Het aantal installaties wereldwijd. Voor codegenerator het aantal developerseats
- Installaties NL: Het aantal installaties in Nederland. Voor codegeneratoren het aantal developerseats
- Vestiging in Benelux: Is er een vestiging in de Benelux?

Tool
- Platforms: Het aantal platforms dat het ETL tool ondersteunt (7 voor Java, draait vrijwel op ieder platform)
- Versie: De versie van het product
- Engine-based / codegenerator: "Is het Engine-based of een Code-generator (eg = engine based; cg = code generator)"
- Type: Wanneer een onbeperkt aantal processtappen tussen source en target kunnen worden gedefinieerd dan is het een 'Proces' anders 'Map'
- Prijs configuratie I (x 1.000 euro): WINTEL, 2 processoren, Oracle, MS SQL Server, 2 ontwikkelaars en een licentie voor ontwikkel en testdoeleinden (1 processor)
- Prijs configuratie II (x 1.000 euro): UNIX, 4 processoren, Oracle, DB/400, SAP, 3 ontwikkelaars en een licentie voor ontwikkel- en testdoeleinden (2 processoren)

Gebruiksvriendelijkheid
- gebruiksgemak: Wat is het gebruiksgemak van het product. Is het snel te leren en is het makkelijk in (dagelijks) gebruik?
- WYSIWYG: Wordt het What You See Is What You Get principe toegepast op DATA?
- schermontwerp: Ziet het scherm er rustig en evenwichtig uit?
- taakcompatibiliteit ETL / EAI: Ondersteunt de tool de taken (en de volgorde daarin) van de ETL-ontwikkelaar?

Overzichtelijkheid en herbruikbaarheid
- herbruikbaarheid componenten*: Zijn componenten herbruikbaar en parametriseerbaar (hier wordt niet copy-paste bedoeld!)?
- decompositie*: Kunnen processen opgedeeld worden in kleine benoembare aanroepbare blokjes (modulair programmeren)?
- user-defined functies*: Is het mogelijk om binnen de tool user-defined functions te definiëren en aan te roepen?
- commentaar selectie van objecten*: Kan men commentaar op een selectie van objecten geven zodat het commentaar vast is verbonden met de objecten?

Foutopsporing
- step by step running*: Kan men stap voor stap de processtapjes uitvoeren?
- row-by-row running*: Kan men het proces rij voor rij het uitvoeren?
- breakpoints*: Kan men breakpoints zetten op een processtapje of een rij gegevens?
- watches*: Kan men watchpoints definiëren?
- compiler/validate*: Kan met 1 druk op de knop de 'code' worden gevalideerd en fouten worden gemarkeerd?

Realtime ETL/EAI/Web services
- integratie batch - realtime*: Kunnen binnen de ETL tool gegevens real-time én in batch worden verwerkt?
- mechanismen*: "Hoe worden veranderingen gedetecteerd of doorgegeven (mq = message queing; logging = database logs of journals; trigger = database triggers)?"
- on-demand data integration*: Kan men een ETL proces als een Web Service publiceren?

Functionaliteit
- splitting datastreams/multiple targets*: Kan een databron in 1 keer worden gelezen en weggeschreven worden naar twee of meer tabellen
- conditional splitting*: Idem, maar dan condtioneel d.w.z. als omzet > 1000 schrijf dan naar tabel 1 anders naar tabel 2
- union*: De rijen van twee tabellen met dezelfde structuur in één tabel plaatsen
- pivoting*: Is het mogelijk om niet-genormaliseerd gegevens die kolomsgewijs zijn gestructureerd om te zetten naar rijen
- depivoting: Is het mogelijk om genormaliseerde gegevens die rijgewijs zijn gestructureerd om te zetten naar kolommen (tegenovergestelde van pivoting)
- key lookup's in memory*: Kun je een tabel volledig in memory laden en daarop zoeken? (zonder te joinen)
- key lookup's herbruikbaar over proces*: Zijn deze herbruikbaar over verschillende laadprocessen heen, zodat key lookup tabel slechts 1 x in het geheugen hoeft te worden geladen?
- slowly changing dimensions*: "Is er ondersteuning voor slowly changing dimensions (hm = handmatig; wizard = wizard; auto = out-of-the-box)"
- scheduler*: Is er een scheduler aanwezig die ook afhankelijkheden ondersteunt?
- status afhandeling binnen job*: Kunnen binnen een job fouten worden gedetecteerd, en kan op basis daarvan een andere route worden gekozen?
- impact analysis*: Is het mogelijk om een impact-analyse te maken van voorgestelde wijzigingen (wanneer een attribuut of tabel moet wijzigen)
- data lineage*: Kan snel de bron van een attribuut/informatie-element worden achterhaald (reversed impact analysis)
- automatische documentatie*: Kan men een proces/transformatie automatisch publiceren/documenteren en er met een browser doorheen navigeren?
- support voor data-miningmodels*: Kunnen tijdens het laadproces resultaten van een datamining proces worden gebruikt?
- support voor analytische functies*: Kunnen allerlei analytische functies zoals forecasting, basket analyses, regressie tijdens laadproces worden aangeroepen?

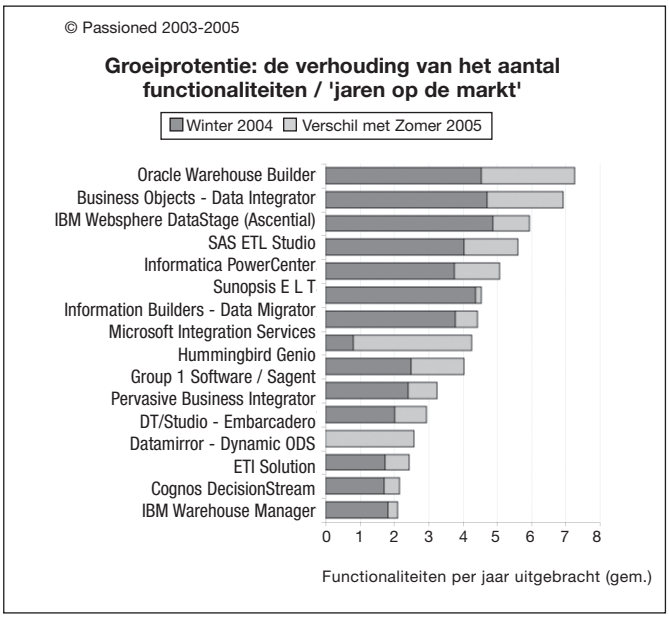
Data sources/targets
- support voor joined tables als bron*: Kun je grafisch (met drag en drop) aangeven dat 2 tabellen door de database gejoined moeten worden ipv door de ETL-tool zelf
- ingebouwde functies voor data kwaliteit*: Zijn er functies beschikbaar die tijdens het uitvoeren van een ETL-proces de kwaliteit van de gegevens controleren (bijv. een matchingtransformatie of een address cleanser) | fuzzy logic
- ingebouwde functies voor data validatie*: Zijn er data validatie functies beschikbaar zoals (verwijder duplicaten, missing values, invalid data) | 100% match
- data profiling*: Opties voor data profiling, unieke waarden, max, min, distributie van waarden, et cetera
- changed data capture*: Ondersteunt de ETL tool het principe van Changed Data Capture (alleen de wijzigingen uit de database selecteren)

- native connections (-ODBC -flatfiles): Hoeveel en welke native connecties ondersteunt de ETL tool? (ODBC, OLE DB, flat files uitgesloten)
- packages / enterprise applications: Bij hoeveel packages / enterprise applications kan met een enkele handeling meta data gelezen worden (bijv. SAP, Siebel, Peoplesoft)
- real-time connecties: Hoeveel en welk type real-time gegevenswachtrijen / berichten kunnen worden gelezen?

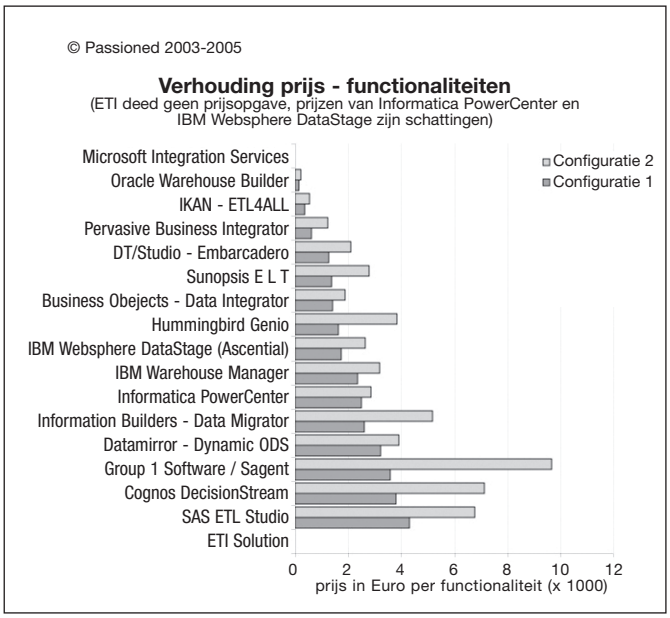
Architectuur en infrastructuur
Parallel processing
- SMP (Symmetric Multiprocessing)*: Wordt Symetric Multiprocessing ondersteunt? Windows NT en UNIX ondersteunen het standaard. Bij SMP delen processoren het intern en extern geheugen
- MPP (Massively parallel processing)*: Bij MPP heeft iedere processor zijn eigen in- en extern geheugen en database, waardoor hoge performance mogelijk wordt. Deze databases dienen echter wel gesynchroniseerd te worden.
- Cluster Aware*: Is de ETL server 'cluster aware' en ondersteunt het load balancing, fail-over en andere cluster mogelijkheden?
- Grid*: Kan men ETL processen draaien in een 'grid'?

Schaalbaarheid
- Job distributie*: Kan men ETL processen toewijzen aan verschillende machines of processoren?
- Data pipelining*: Kan men binnen een ETL proces de verschillende stapjes toewijzen aan verschillende machines of processoren?
- Partitionering*: Kan men op basis van bijvoorbeeld productcodes bepalen welke machine of processor welke data verwerkt?

Basisarchitectuur: h = hub & spoke (alles door één punt), d = distributed (meerdere lijnen tussen sources en targets) en m = multi hub/spoke
- end-to-end BI infrastructuur*: Wisselt de ETL tool meta data uit (bijv. star schemas) met OLAP of reporting tools hetzij van eigen makelij of van derden
- CWM-ondersteuning*: Is de ETL tool CWM-compliant (ondersteunt het het Common Warehouse Model)?
- versiebeheer*: Is er een mogelijkheid voor versiebeheer met check-in en check-out mogelijkheden?



Afbeelding 6: De groeipotentie van de ETL tools. Oracle Warehouse Builder, Business Objects Data Integrator en IBM Websphere DataStage gooien hoge ogen en krijgen gemiddeld zes tot zeven nieuwe functionaliteiten per jaar erbij. IBM met Warehouse Manager en Cognos brengen slechts gemiddeld twee functionaliteiten uit per jaar.



Afbeelding 7: De verhouding tussen de prijs en de geboden functionaliteit. Microsoft levert met Integration Services het meeste waar voor zijn geld (het wordt gratis meegeleverd met MS SQL Server). Group 1 Software met Sagent, Cognos met DecisionStream en SAS met ETL Studio zijn het duurst. Men betaalt voor deze laatste twee tools bijna € 4.000 per functionaliteit (configuratie 1). De prijzen van Informatica PowerCenter en IBM Websphere DataStage zijn schattingen. Bij Oracle Warehouse Builder, Microsoft Integration Services en IBM Warehouse Manager dient men rekening te houden met de aanschaf van de database indien men deze nog niet in huis heeft.

Hummingbird en wellicht Cognos (onder andere verhouding prijs-functionaliteit op dit moment).
De verwachting is dat – bij succesvolle introductie van IS en een stabiel product – de BI-markt de komende paar jaar een aantal verrassingen in petto heeft. Een mogelijk scenario is dat enkele EBIS-leveranciers nog wel hun ETL-producten proberen mee te leveren, maar zich meer en meer concentreren op de front-end, waardoor dan een heftige concurrentiestrijd kan ontstaan. Binnen drie tot vier jaar zou dat dan kunnen leiden tot een of meer overnames of faillissementen.

Conclusie

In dit artikel presenteerden we het ETL-onderzoek en de belangrijkste ontwikkelingen in de ETL-markt. Qua basisfunctionaliteit hebben Oracle Warehouse Builder en Business Objects Data Integrator, Ascential en Informatica van de eerste respectievelijk tweede plaats gedrongen. Het was duidelijk te zien dat de twee koplopers veel werk verzet hebben. Zij hebben niet stilgezeten en essentiële functionaliteiten toegevoegd aan hun product. Het wachten is op release 8 van Cognos DecisionStream. Wellicht dat zij met een aantal verrassingen komen, hetgeen niet zou verbazen omdat de introductie omgeven is met een prettige sfeer van spanning.
Microsoft is in ieder geval de hoogste stijger, mits ze het product natuurlijk nog dit jaar op de markt brengen. Integration Services kan de komende jaren (grote) impact hebben op de ETL-markt, en zelfs de BI-markt, maar moet zich nog wel eerst bewijzen. Sunopsis zal in versie 4 met essentiële functionaliteiten moeten komen voor data-integratietaken (waaronder union, splitting en pivotting) anders zal het de boot wellicht missen, ondanks de unieke propositie als database-onafhankelijke SQL code generator.

Noot

1. *Dit is een combinatie van twee producten namelijk Transformation Services en Constellar Hub.*

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).