Strijd der giganten lijkt gestreden en het is dringen aan de top

De ETL-matrix: product-kenmerken en leveranciers

Daan van Beek

In DB/M 2 zijn naar aanleiding van een onderzoek op de markt voor ETL-tools, verschillende trends en ontwikkelingen besproken. Daaruit bleek onder meer dat ETL en EAI nog lang niet zo snel naar elkaar toegroeien als wel wordt gedacht en dat de laadtijden van gegevenspakhuizen tot wel 20 procent korter kunnen, als de leveranciers slimmer omgaan met het beschikbare geheugen en de volgorde van de laadprocessen.

Verrassend was tevens dat veel leveranciers inmiddels herbruikbaarheid en foutopsporing hebben omarmd en ietwat teleurstellend was hoe weinig ETL-tools 'slowly changing dimensions' ondersteunen. Reden genoeg om wat dieper op de producten zelf, de markt en de verschillende kenmerken van de ETL-tools in te gaan.

In dit artikel wordt stilgestaan bij de ETL-Matrix, hierin komen de meest belangrijke kenmerken van nagenoeg alle ETL-tools aan bod. Daarnaast wordt een rangschikking aangebracht voor de diverse facetten van het onderzoek: de basisfunctionaliteiten, de gebruiksvriendelijkheid, de connectiviteit, de platformondersteuning en de aanschafprijs. Hoewel de marktleiders in het Magic ETL Quadrant van Gartner (zie afbeelding 1) ook in dit onderzoek zeker niet altijd slecht presteren, is er een aantal interessante verschillen op te merken als we kijken naar de gebruiksvriendelijkheid, connectiviteit en platformondersteuning. In de afbeeldingen 2 tot en met 7 is dat weergegeven.

Onderzoeksopzet

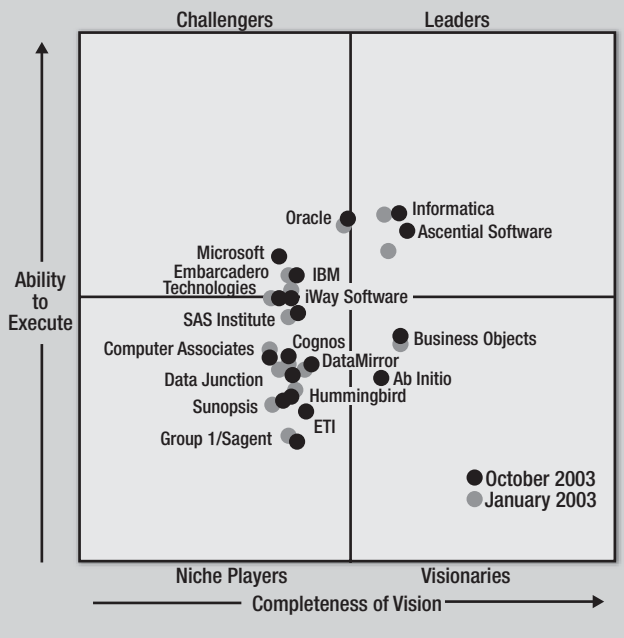
Het onderzoek is gehouden onder zestien leveranciers[1] die vermeld staan in het Gartner Magic ETL Quadrant van oktober 2003. Van die leveranciers hebben de volgende meegewerkt aan het onderzoek: Pervasive met Dj Cosmos, Sunopsis, Ascential met DataStage, IBM met Warehouse Manager, Cognos met DecisionStream, Group 1 Software met Data Flow Manager[2], Informatica met PowerCenter, Information Builders met ETL Manager[3], Business Objects met Data Integrator, SAS met ETL

ETL Winter Survey 2004 (2): de uitkomsten

Daan van Beek van passionned 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 DB/M 2 zijn de ontwikkelingen rond ETL-tools uit de doeken gedaan; in dit nummer wordt ingezoomd op de kenmerken, overeenkomsten en verschillen van de individuele producten en wordt de ETL-matrix gepresenteerd.

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



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

	Pervasive Dj Cosmos	Sunopsis	Ascential DataStage	IBM Warehouse Manager	Cognos DecisionStream	Group 1 Software / Sagent	Informatica PowerCenter	Information Builders - ETL Manager	Business Objects - Data Integrator	SAS ETL Studio
Bedrijf										
Gestart met verkoop	1985	1998	1996	1993	1993	1996	1996	1996	1998	1996
Klanten WW	30000	250	3000	disclosed	1000	1500	1800	800	300	900
Installaties WW	30000	1300ds	3000	disclosed	1000	1800	4000	800	300	1500
Installaties NL	unknown	0ds	25	disclosed	25	15	150	22	20	50
Vestiging in Benelux	Brussel	ja	ja	ja	ja	ja	ja	ja	ja	ja
Tool										
Platforms (7 voor Java, running on all platforms)	7	7	6	5	5	4	5	5	3	6
Versie	8.0	v3	7.0	8.1	7.2	4.5i	7	5.2.3	6.1	v8
Engine-based (eb) / code-generator (cg)	eb	cg	eb	cg	eb	eb	eb	cg	eb	cg
Type	map	map	proces	proces	proces	proces	map/proces	map/proces	proces	proces
Prijsindicatie per cpu / per seat (vanaf)	$ 7.000	€ 30.000	€ 35.000	€ 13.405	€ 78.335	€ 15.000	$ 60.000	€ 9.600	€ 32.480	€ 125.000
- in Euro's (koers: 1,21 dollar voor één Euro)	€ 5.785	€ 30.000	€ 35.000	€ 13.405	€ 78.335	€ 15.000	€ 49.587	€ 9.600	€ 32.480	€ 125.000
Gebruiksvriendelijkheid										
- Gebruiksgemak	-	+	+	+	+	++	0	0	++	++
- WYSIWYG	+	0	+	0	-	+	-	0	+	0
- Schermontwerp	0	+	+	+	+	+	+	+	++	++
- Taakcompatibiliteit ETL / EAI	-	+	+	+	+	+	+	0	+	+
Overzichtelijkheid en herbruikbaarheid										
- Herbruikbaarheid componenten*	ja	ja	ja	ja	nee	nee	ja	nee	ja	ja
- Decompositie*	ja	nee	ja	ja	ja	nee	ja	nee	ja	ja
- User defined functies*	ja	ja	ja	ja, db2	ja	nee	ja	nee	ja	ja
- Commentaar selectie van objecten*	nee	nee	ja	nee	nee	nee	nee	nee	nee	nee
Foutopsporing										
- Step by step running*	ja	nee	ja	nee	nee	nee	ja	nee	nee	nee
- Breakpoints*	nee	nee	ja	nee	nee	nee	ja	nee	nee	nee
- Watches*	nee	nee	ja	nee	nee	nee	ja	nee	nee	nee
Realtime ETL/EAI										
- Integratie batch / real-time*	ja	ja	ja	nee	nee	nee	ja	nee	ja	ja
- Mechanismen	mq	mq+trig	mq	n.v.t.	n.v.t.	n.v.t.	mq+log	n.v.t.	mq	mq
Functionaliteit										
- Splitting datastreams / multiple targets*	ja	nee	ja	ja	ja	ja	ja	ja, m15	ja	ja
- Conditional splitting*	ja	nee	ja	ja	ja	ja	ja	ja	ja	ja
- Pivoting*	ja	nee	ja	ja	ja	ja	ja	nee	ja	ja
- Key lookup's in memory*	ja	ja	ja	ja, db2	ja	ja	ja	nee	ja	ja
- Key lookup's herbruikbaar over proces*	ja	ja	nee	ja, db2	nee	nee	nee	nee	nee	ja
- Slowly changing dimensions*	hm	auto	wizard	hm	auto	hm	wizard	hm	auto	template
- Scheduler*	ja	ja	ja	ja, db2	ja	ja	ja	ja	ja	ja
- Impact analysis*	ja	ja	ja	ja	ja	ja	ja	ja	ja	ja
- Changed data capture*	nee	ja	ja	nee	nee	ja	ja	ja	ja	ja
Data sources/targets										
- Support voor joined tables als bron*	ja	ja	ja	ja	nee	ja	ja	ja	ja	ja
- Ingebouwde functies voor datakwaliteit*	nee	ja	ja	nee	ja	ja	ja	nee	nee	ja
- Native connections (-ODBC -flatfiles) (c)	+100	21	41	3	6	6	16	83	6	23
- Packages / enterprise applications (c)	1	7	6	0	1	6	5	5	5	6
- Real-time connecties (c)	3	6	8	0	0	0	7	1	8	3
Overig										
- Server-grid-technologie*	nee	nvt	ja	nvt	nee	nee	ja	nvt	nee	ja
- End-to-end BI-infrastructuur*	nee	nee	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
Berekeningen										
Aantal basisfunctionaliteiten	15	13	21	14	13	10	21	9	17	18
Gebruiksvriendelijkheid	-1	3	4	3	2	5	1	1	6	5
Aantal ondersteunde platforms	7	7	6	5	5	4	5	5	3	6
Aantal ondersteunde sources / targets	104	34	55	3	7	12	28	89	19	32
Aantal jaren op de markt	19	6	8	11	11	8	8	8	6	8

	IKAN - ETL4ALL	ETI Solution	Oracle Warehouse Builder	Hummingbird ETL	DT/Studio - Embarcadero	Microsoft DTS
	2003	1993	1998	1996	2002	1997
	25	370	disclosed	400	75	onbekend
	4	5000ds	disclosed	400	75	onbekend
	2	15ds	disclosed	5	0	onbekend
	ja	nee	ja	ja	ja	ja
	7	12	6	3	7	1
	2.3	5.1	9.2.03	5.03	2.1	2000
	eb	cg	cg	eb	eb	eb
	proces	map	proces	proces	proces	proces
	€ 1.799	NULL	€ 4.196	$ 50.000	€ 25.000	€ 0
	€ 1.799	NULL	€ 4.196	€ 41.322	€ 25.000	€ 0
	+	0	+	+	+	0
	0	-	0	-	-	0
	++	+	+	+	+	0
	-	0	+	+	+	0
	ja	nee	nee	ja	nee	nee
	ja	nee	nee	ja	nee	ja
	nee	nee	ja	ja	nee	ja
	nee	nee	nee	nee	nee	nee
	nee	nee	ja	nee	ja	nee
	nee	nee	ja	nee	ja	nee
	nee	nee	ja	nee	ja	nee
	nee	ja	ja	nee	nee	nee
	n.v.t.	mq	mq+log+trig	n.v.t.	n.v.t.	n.v.t.
	nee	ja	ja	ja	ja	nee
	nee	ja	ja	ja	ja	nee
	nee	nee	ja	ja	ja	nee
	nee	ja	ja,db	ja	nee	ja
	nee	nee	ja,db	nee	nee	nee
	hm	hm	hm	hm	hm	hm
	nee	nee	ja	ja	ja	Agent
	nee	ja	ja	ja	nee	nee
	nee	ja	ja,db	ja	ja	nee
	nee	ja	ja	ja	nee	ja
	nee	ja	ja	nee	nee	nee
	0	12	7	18	4	1
	1	2	1	2	0	0
	0	2	3	2	1	0
	nee	nvt	nvt	nee	nee	nee
	nee	nee	ja	nee	nee	nee
	nee	nee	ja	nee	nee	nee
	nee	ja	ja	ja	nee	nee
	2	10	19	12	8	5
	2	0	3	2	2	0
	7	12	6	3	7	1
	1	16	11	22	5	1
	1	11	6	8	2	7

Legenda afbeelding 2: De ETL-matrix

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 codegeneratoren het aantal developerseats (ds).
- Installaties NL: het aantal installaties in Nederland. Voor codegeneratoren het aantal developer seats (ds).
- Vestiging in Benelux: is er een vestiging in de Benelux?

Tool
- Platforms: het aantal platforms dat wordt ondersteund.
- Versie: de versie van het product dat is geëvalueerd.
- Engine-based (eb) of code-generator (cg): is het product Engine-based of een code-generator?
- Type: wanneer een onbeperkt aantal processtappen tussen source en target kan worden gedefinieerd dan is het een 'Proces' anders 'Map'.
- Prijsindicatie per server-cpu: de instapprijs voor een enkele cpu. Voor code-generatoren: de prijs van een developer seat.

Gebruiksvriendelijkheid

++ zeer goed + goed 0 neutraal - slecht -- zeer slecht
- Gebruiksgemak: wat is het gebruiksgemak van het product. Is het snel te leren, is het makkelijk in gebruik?
- WYSIWYG: kan de ontwikkelaar tijdens het ontwikkelen van een ETL-proces de gegevens en het resultaat van de transformaties gemakkelijk bekijken?
- Schermontwerp: ziet het scherm er rustig en evenwichtig uit. Gelet is op onder meer symmetrie, consistentie, kleurgebruik, fontgebruik.
- Taakcompatibiliteit ETL/EAI: ondersteunt de tool de taken (en de volgorde daarin) van de ETL-ontwikkelaar? Vaak zal een ETL ontwikkelaar eerst de bronnen in kaart brengen en targets definiëren, waarna pas wordt gestart met het ontwikkelen van de transformaties.

Overzichtelijkheid en herbruikbaarheid
- Herbruikbaarheid componenten*: zijn componenten en vooral transformaties, of delen ervan, herbruikbaar en parametriseerbaar?
- Decompositie*: kunnen processen opgedeeld worden in kleine benoembare aanroepbare blokjes (modulair programmeren)?
- User defined functions*: is het mogelijk om binnen de tool user defined functions te definiëren en aan te roepen?
- Commentaar selectie van objecten*: kan er commentaar op een selectie van objecten worden gegeven, zodat het commentaar vast is verbonden met de objecten?

Foutopsporing
- Step by step running*: is het mogelijk om stap voor stap (en/of rij voor rij) de processen uit te voeren en kijken wat het resultaat is?
- Breakpoints*: is het mogelijk om breakpoints te zetten op een transformatie of een rij gegevens?
- Watches*: biedt de tool de mogelijkheid om watch points definiëren?

Real-time ETL/EAI
- Integration batch/real-time*: kunnen binnen de ETL-tool gegevens real-time én in batch worden verwerkt?
- Mechanismen: hoe worden wijzigingen in de bron gedetecteerd en vervolgens doorgegeven? (mq = message queing; logging = database logs of journals; trigger = database triggers).

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 conditioneel, dat wil zeggen als omzet > 1000; schrijf dan naar tabel 1, anders naar tabel 2.
- Pivoting*: is het mogelijk om niet-genormaliseerde gegevens die kolomsgewijs zijn gestructureerd en zo worden aangeleverd om te zetten naar rijen?
- 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 key lookup's herbruikbaar over de verschillende laadprocessen heen, dat wil zeggen dat ze slechts een maal in het geheugen worden geladen en voor meerdere feittabellen gebruikt kunnen worden? Voor meer uitleg zie het artikel in DB/M2.
- Slowly changing dimensions*: is er ondersteuning voor slowly changing dimensions? (hm = handmatig; wizard = wizard-gedreven; auto = ingebakken).
- Scheduler*: is er een scheduler aanwezig die ook afhankelijkheden ondersteunt?
- Impact analysis*: is het mogelijk om een impact-analyse te maken van voorgestelde wijzigingen (wanneer een attribuut of tabel moet wijzigen)?
- Changed data capture*: ondersteunt de ETL-tool het principe van Changed Data Capture (alleen de wijzigingen uit de database meenemen)?

Data sources/targets
- Support voor joined tables als bron*: kan men grafisch (bijvoorbeeld met drag en drop) aangeven dat twee tabellen door de database gejoined moeten worden in plaats van door de ETL-tool zelf (back-propagation van joins)?
- Ingebouwde functies voor datakwaliteit*: zijn er functies beschikbaar die tijdens het uitvoeren van een ETL-proces de kwaliteit van de gegevens controleren (bijvoorbeeld een matching-transformatie of een address cleanser)?
- Native connections (c): hoeveel en welke native connecties ondersteunt de ETL-tool? (ODBC, OLE DB, flat files uitgesloten).
- Packages/enterprise applications (c) : kan de ETL-tool metadata lezen van package/enterprise applications (bijvoorbeeld SAP, Siebel, Peoplesoft), zo ja hoeveel?
- Real-time connections (c): hoeveel en welk type real-time gegevenswachtrijen/berichten leest en schrijft de ETL-tool?

Overig
- Server-grid-technologie*: ondersteunt de ETL-tool grid-technologie, dat wil zeggen kunnen ETL-processen automatisch worden uitgesmeerd over meerdere servers afhankelijk van hun capaciteit?
- End-to-end BI-infrastructuur*: wisselt de ETL-tool metadata uit (bijvoorbeeld sterschema's) met OLAP of reporting tools hetzij van eigen makelij of van derden?
- CWM-compliant*: is de ETL-tool CWM-compliant (ondersteunt het het Common Warehouse Meta Model)?
- Versiebeheer*: is er een mogelijkheid voor versiebeheer met check-in en check-out faciliteiten?

Studio, IKAN met ETL4ALL[4], ETI met ETI Solution, Oracle met Oracle Warehouse Builder[5], Hummingbird met Hummingbird ETL, DT/Studio met Embarcadero[6] en ten slotte Microsoft met DTS[7]. De interviews, waarin de verschillende leveranciers hun producten mochten presenteren, zijn in bovenstaande volgorde gehouden in de periode december 2003 tot en met januari 2004. Er is vooral gelet op de basisfunctionaliteiten en -kenmerken die de snelheid van het ontwikkelen in data-integratieprojecten, al dan niet real-time, kunnen versnellen en het onderhoud van datawarehouses vergemakkelijken. De functionaliteiten zijn dus niet volledig, maar vormen echter wel een goede dwarsdoorsnede van de meest gewenste functionaliteiten voor dat soort projecten.

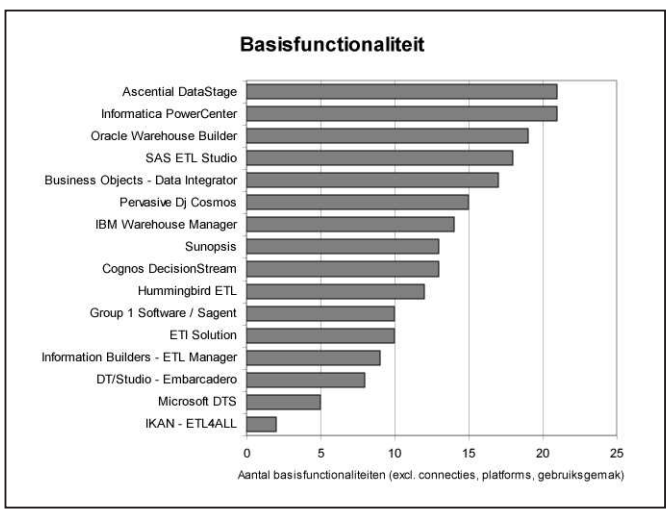
De ETL-matrix

Afbeelding 2 presenteert de ETL-matrix, met verticaal alle productkenmerken en horizontaal de verschillende ETL-tools. De matrix bevat een aantal onderdelen: Bedrijf, Tool, Gebruiksvriendelijkheid, Overzichtelijkheid en herbruikbaarheid, Foutopsporing, Real-time ETL/EAI, Functionaliteit, Data sources/ targets, Overig en Berekeningen. In dat laatste onderdeel wordt de puntentelling bepaald die in de grafieken voor de rangschikking is gebruikt.

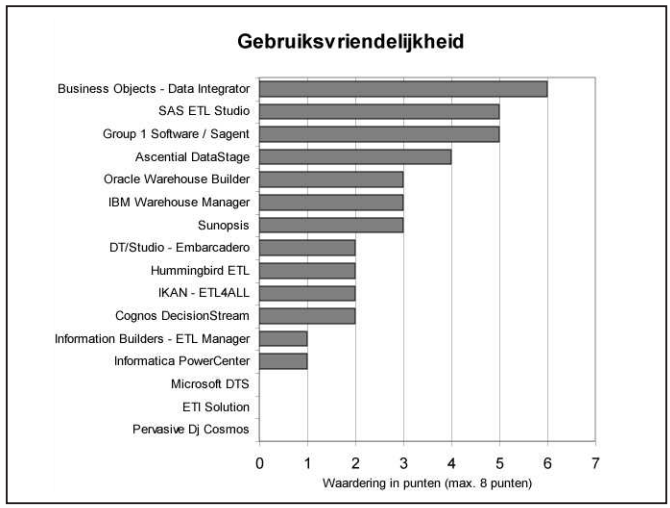
Vervolgens worden grafieken getoond die de ETL-tools rangschikken naar compleetheid, gebruiksvriendelijkheid, connectiviteit, platformondersteuning, groeipotentie en aanschafprijs.
Voor iedere grafiek is de formule voor de berekening van de rangschikking opgenomen.

Compleetheid

In afbeelding 3 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 dan krijgen ze twee punten.



Afbeelding 3: De rangschikking naar basisfunctionaliteit, Ascential DataStage en Informatica PowerCenter hebben de meeste functionaliteiten en IKAN met ETL4ALL de minste functionaliteiten.



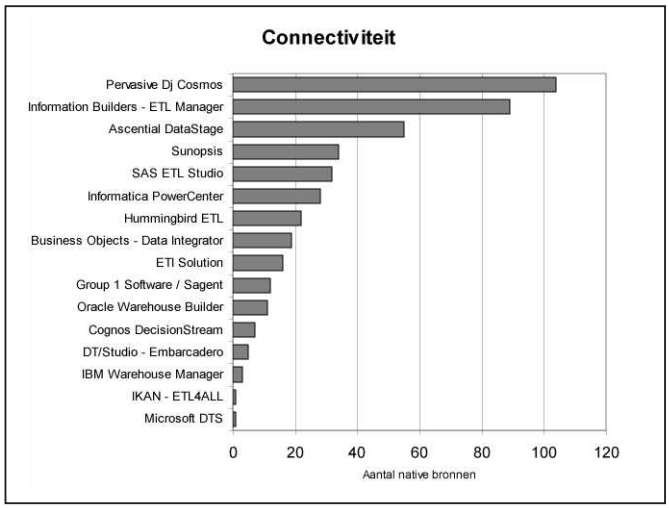
Afbeelding 4: De rangschikking naar gebruiksvriendelijkheid, Business Objects gooit hier hoge ogen, Microsoft DTS, ETI Solution en Pervasive Dj Cosmos behalen geen punten.

Het maximale aantal punten dat kan worden behaald is 24, verdeeld over 23 kenmerken. Deze zijn in de legenda gemarkeerd met een sterretje (*).

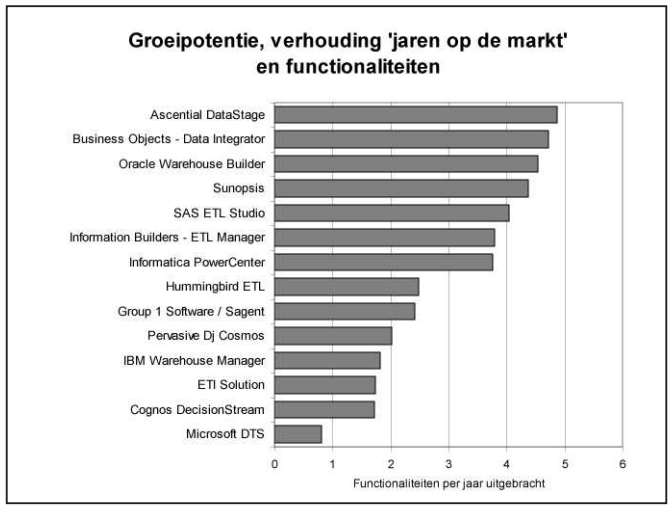
Gebruiksvriendelijkheid

Het meten van gebruiksvriendelijkheid is niet gemakkelijk en niet altijd objectief vast te stellen. Toch is gezien het belang van gebruiksvriendelijkheid bij de vaak toch al complexe ETL-processen, een eerste aanzet gegeven om dat zo objectief mogelijk te meten. Daarbij is tijdens de presentatie gelet op hoe gemakkelijk een ETL-proces kan worden ontwikkeld, vormgegeven en onderhouden, wat de inleertijd is, of de gebruikers-interface 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 4.

Het maximale aantal punten dat een ETL-tool 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, dan krijgt de tool twee punten (drie punten voor ieder plusje minus één punt voor het minnetje).



Afbeelding 5: De ETL-tools gerangschikt naar connectiviteit, Pervasive met Dj Cosmos en Information Builders met ETL Manager behalen veel punten, Microsoft DTS, IBM Warehouse Manager en IKAN met ETL4ALL zijn nauwelijks geschikt voor heterogene omgevingen met veel verschillende soorten gegevensbronnen.

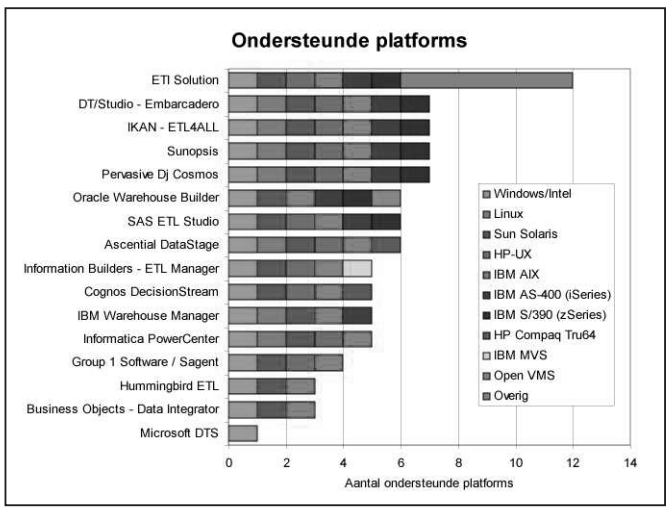


Afbeelding 7: De volgende ETL-tools hebben groeipotentie: DataStage, Data Integrator, Oracle Warehouse Builder en Sunopsis gooien hoge ogen en brengen gemiddeld vier tot vijf nieuwe functionaliteiten per jaar op de markt. Microsoft DTS en DecisionStream van Cognos brengen slechts gemiddeld anderhalve functionaliteit uit per jaar.

Connectiviteit

In afbeelding 5 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 connectiviteit-score is een optelling van het aantal bronnen, het aantal enterprise applications van waaruit de metadata kunnen worden ingelezen en het aantal real-time gegevenswachtrijproducten dat kan worden gelezen. Deze kenmerken zijn in de matrix en legenda aangegeven met de tekens (c).



Afbeelding 6: De rangschikking van de ETL-tools naar platformondersteuning, ETI ondersteunt de meeste platforms, Microsoft met DTS het minste aantal.

Platformondersteuning

De rangschikking naar platformondersteuning is weergegeven in afbeelding 6. Alle verschillende versies van Windows zijn op één hoop gegooid, de diverse varianten van Unix zijn apart genomen. 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

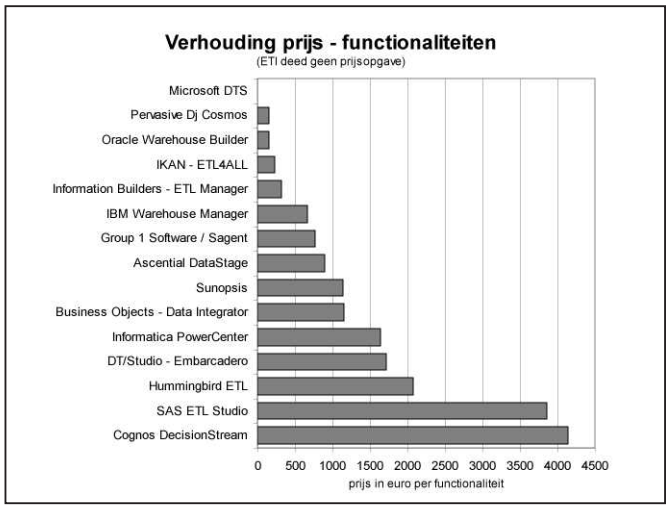
Kijken we naar de verhouding tussen het aantal jaren op de markt en het totale aantal functionaliteiten (inclusief gebruiksvriendelijkheid, connectiviteit en platformondersteuning) dan kunnen we dat vertalen naar de groeipotentie van de tool. Dit is weergegeven in afbeelding 7. Welke tools zullen de 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 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.

Verhouding prijs-functionaliteit

In afbeelding 8 is de verhouding weergegeven tussen de prijs en het aantal basisfunctionaliteiten. Ook hier is weer de gebruiksvriendelijkheid, de connectiviteit en de platformondersteuning meegerekend, immers leveranciers spannen zich niet alleen in om kale functionaliteit te leveren maar natuurlijk ook om deze makkelijk te kunnen bedienen onder verschillende infrastructuren en in heterogene IT-omgevingen.



Afbeelding 8: De verhouding tussen de prijs en de functionaliteiten. Microsoft levert met DTS het meeste waar voor zijn geld (het wordt gratis meegeleverd), SAS met ETL Studio en Cognos met DecisionStream zijn het duurst. Men betaalt voor deze laatste twee tools maar liefst plus minus € 4.000 per functionaliteit.

Welke tool is nu de beste?

Het is niet makkelijk om deze vraag direct te beantwoorden, want uiteindelijk draait het om het rendement dat met de ETL-tool bereikt kan worden en in welke (technische) omgeving de ETL-tool ingezet gaat worden. Werkt uw organisatie vooral met Oracle dan is Oracle Warehouse Builder een potentiële kandidaat, werkt uw organisatie in een zeer heterogene omgeving met een grote diversiteit aan bronnen, dan is Pervasive Dj Cosmos het overwegen waard. Let echter wel dat het product laag scoort op gebruiksvriendelijkheid, het is een echte programmeurs-tool. Werkt uw organisatie veel met externen dan is marktondersteuning erg belangrijk, u wilt dan snel een externe consultant inzetten die alle *ins en outs* kent van de tool. In dat geval is PowerCenter of DataStage aan te raden. Is uw organisatie verspreid over de gehele wereld en is de infrastructuur erg divers (veel verschillende platformen) dan komt men uit op ETI Solution. Wilt u batch én real-time ETL uitvoeren binnen één omgeving en applicatie-integratieprojecten gaan doen, dan is er inmiddels meer keus waaronder Dj Cosmos, Sunopsis, PowerCenter, DataStage enzovoort. Worstelt de organisatie vooral met het goed inrichten van een datawarehouse waarin historie een belangrijk fenomeen is, dan moet vooral gekeken worden of de ETL-tool slowly changing dimensions automatisch ondersteunt en zijn Sunopsis, Cognos DecisionStream of Data Integrator van Business Objects goede kandidaten.

Waar het om gaat, is dat er aan de hand van organisatorische en IT-kenmerken wordt gekeken welk product het beste bij de organisatie en de ETL-ontwikkelaars past. Dit matching-proces wordt vergemakkelijkt wanneer een organisatie specifieke eisen heeft (bijvoorbeeld het moet draaien op Linux en het moet real-time data-integratie aankunnen) die de keuze al snel beperken. In het voorbeeld van Linux in combinatie met real-time blijven er nog vijf ETL-tools over. Stelt een organisatie daarnaast nog meer eisen dan blijft er wellicht geen tool meer over die de eisen kan invullen. De organisatie is dan min of meer genoodzaakt om met behulp van bijvoorbeeld SQL-scripts zelf een datawarehouse te ontwikkelen of kan haar eisen afzwakken. Het zelf ontwikkelen is echter bijna altijd af te raden omdat daarmee de ontwikkelsnelheid en de onderhoudbaarheid sterk achterblijven ten opzichte van ETL-tools. Gemiddeld kan men met een ETL-tool drie tot vijf keer sneller ontwikkelen en datawarehouses die het complete bedrijfsproces ondersteunen, kunnen met de huidige technologische stand van zaken met ETL-tools in 6 tot 9 maanden worden gebouwd. Mits er natuurlijk een juiste 'fit' is tussen de tool en de (IT) organisatie.

Conclusie

Het onderzoek wijst uit dat de ETL-tools onderling op nogal wat punten verschillen, maar ook dat er gemeenschappelijke ontwikkelingen zijn waar te nemen (zie het artikel in DB/M2). In dit onderzoek zijn alle belangrijke functionaliteiten en kenmerken op een rijtje gezet en gewaardeerd. De rangschikking die in de grafieken te zien is, is transparant en onderbouwd met 'harde' feiten die in de ETL-matrix zijn verzameld. Het samenstellen van een shortlist is dan geen erg moeilijke opgave meer.

Men kan constateren dat een aantal, ook wat minder bekende ETL-tools, nu een flinke stap maakt of gaat maken. Hieronder is Business Objects met Data Integrator en Sunopsis. De eerste zal dan wel moeten werken aan een hogere connectiviteit en betere platformondersteuning en de laatste zal de komende jaren in hetzelfde tempo als voorheen, essentiële functionaliteiten moeten blijven toevoegen.

Ongetwijfeld zal de nummer één van dit onderzoek, Ascential met DataStage, die de meeste functionaliteit biedt tegen een gemiddelde prijs, redelijk gebruiksvriendelijk is, een goede connectiviteit en platformondersteuning kent, zijn positie verder proberen uit te bouwen.

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.

Noten

1. *Door omstandigheden waren Computer Associates, Data Mirror en Ab Initio niet in staat om mee te doen. In plaats daarvan heeft IKAN meegedaan.*
2. *Versie 5, verwacht in het voorjaar van 2004, zal herbruikbaarheid, decompositie en user-defined functions ondersteunen.*
3. *Versie 5.5 scoort iets beter op gebruiksvriendelijkheid en er zijn meer mogelijkheden om de data te zien tijdens het ontwikkelen.*
4. *Versie 2.4 zal joined tables als bron, changed data capture en CWM gaan ondersteunen.*
5. *Herbruikbaarheid van componenten zal in versie 10g worden ondersteund.*
6. *De nieuwe versie, verwacht in juni 2004, zal server grid-technologie en versiebeheer gaan bevatten. In de loop van 2004 worden connecties geleverd naar SAP R/3, Peoplesoft, Siebel en JD Edwards.*
7. *De nieuwe versie van DTS, onderdeel van Yukon, zal extra functionaliteiten bevatten op het gebied van herbruikbaarheid, foutopsporing, slowly changing dimensions en versiebeheer. Deze versie zal in de eerste helft van 2005 uitkomen en integreert volledig met Visual Studio en .NET.*