

Vier categorieën ETL-tools in update van de ETL-matrix

ETL Onderzoek 2007

Norman Manley

Zo'n slordige anderhalf jaar geleden heeft Daan van Beek in opdracht van Database Magazine een poging gedaan de markt van ETL-tools in kaart te brengen. De uitkomsten werden gepubliceerd als de ETL-matrix in DB/M 6 uit 2005. Nodig tijd voor een update van deze matrix. Norman Manley, collega van Daan van Beek, trok de markt in en inventariseerde, categoriseerde en classificeerde het aanbod.

Er zijn veel verschillen tussen het onderzoek van februari 2007 en dat van 18 maanden geleden. Om ervoor te zorgen dat wij geen appels met peren vergelijken zijn de producten in vier categorieën ingedeeld:

1. Pure ETL. Deze producten staan los van de leveranciers van zowel de database als de BI-producten die eventueel gebruikt worden;
2. Database-geïntegreerd. ETL-producten van een database-leverancier waar een deel van de functionaliteit in het ETL-product zelf zit en een deel in de (meegeleverde) database;
3. BI-geïntegreerd. ETL-producten van een BI-leverancier, waarbij beoogd wordt om waarde toe te voegen aan het BI-product;
4. Niche ETL. Spelers, die producten ontwikkeld hebben voor één bepaald doel waar zij erg goed in zijn, bijvoorbeeld het converteren van gegevens. De producten zijn niet bedoeld om breder ingezet te worden.

Zelfs rekening houdend met bovenstaande indeling, valt een aantal producten in meer dan één categorie, wat het vergelijken moeilijk maakt.

ETL uit de kraan

Ten opzichte van het vorige onderzoek is er een aantal producten en leveranciers bijgekomen; verdwenen zijn enkele die naar aanleiding van de resultaten van voorgaande jaren niet meer wensten mee te doen. Twee leveranciers, Ab Initio (zoals altijd) en Hummingbird, zijn volledig onbereikbaar geweest. Het Nederlandse kantoor van Information Builders gaf te kennen niet te willen meewerken, maar vanuit hun hoofdkantoor werd volledige medewerking verleend.

Een aantal vendors voert meer dan één product. Voor zover wij de producten apart konden beoordelen hebben wij dat gedaan. Waar het onduidelijk was of het antwoord betrekking had op product 1

of product 2 hebben wij de producten in het geheel niet opgenomen in de matrix.

Per functionaliteit is een punt aan elk product toegekend. Het aantal punten voor functionaliteit is ten opzichte van het vorige onderzoek gemiddeld met 18 procent gestegen.

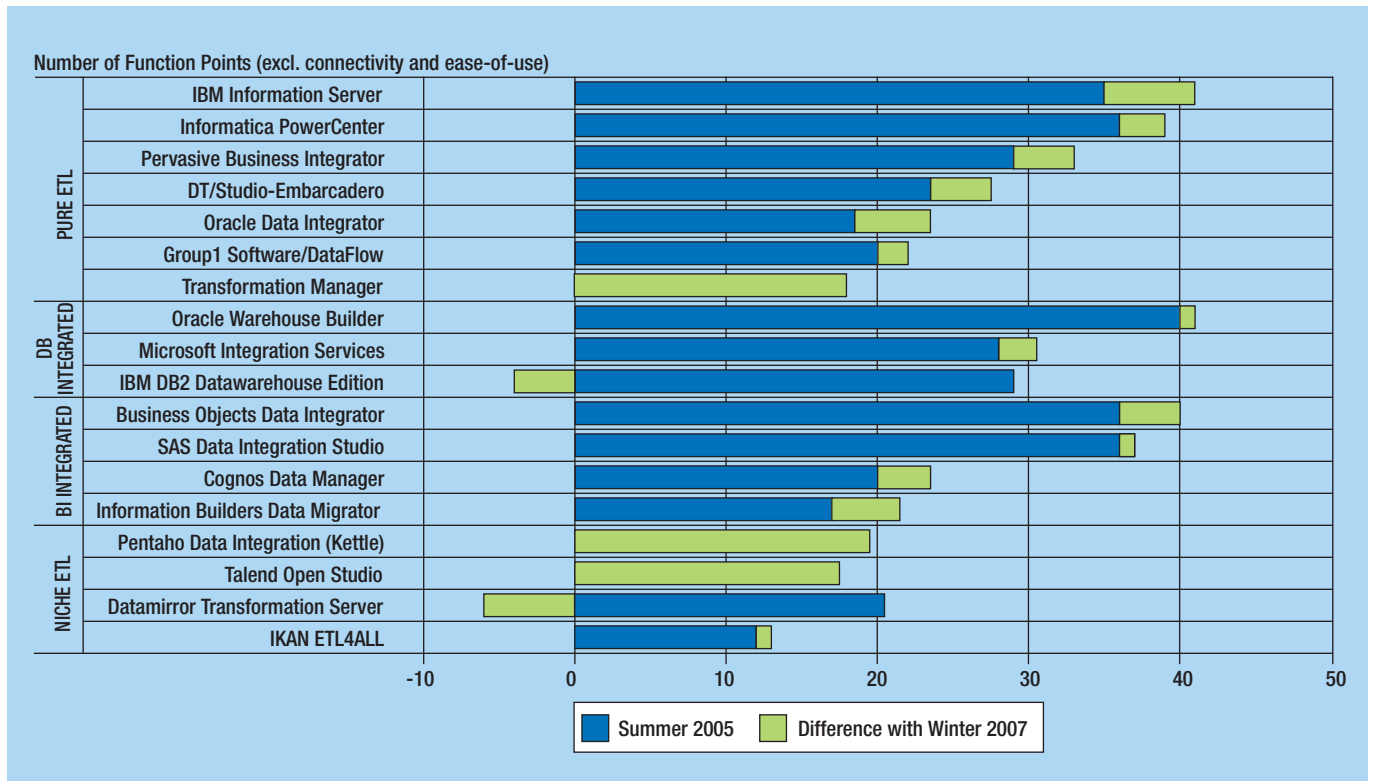
De prijzen zijn in veel gevallen nog onduidelijker geworden dan in de vorige editie al het geval was: het is erg afhankelijk van welke software men reeds van de leverancier heeft en op hoeveel machines het moet draaien. Sommige leveranciers bieden 'ETL uit de kraan' aan, ook wel Software as a Service (SaaS) genoemd. In dat model is de prijs gebaseerd op het recht van gebruik en het daadwerkelijke 'verbruik'. Dit alles maakte het vrijwel onmogelijk om de prijzen per functionaliteit te berekenen.

Motieven

Uit gesprekken met de leveranciers van de verschillende categorieën is het helder geworden dat men heel verschillende motieven heeft om een ETL-product op de markt te brengen. Bij de eerste categorie is het helder: men bouwt ETL-producten omdat dat hun

The ETL Survey 2007

De ETL-matrix die in deze editie van Database Magazine staat gepubliceerd is een uittreksel uit The ETL Survey 2007. Op onze website www.dbm.nl kunt u de gehele ETL-matrix inzien als doorzoekbare database, met de mogelijkheid tot matchmaking: door het ingeven of fijnstellen van selectiecriteria kunt u een shortlist samenstellen van producten die het meest beantwoorden aan uw requirements. Het diepgaande en uitgebreide Engelstalige rapport The ETL Survey 2007 kunt u aanschaffen via de website www.etltool.com van Passionned.



Afbeelding 1: Score op functionaliteiten.

business is, de producten zijn veel gebruikt, hebben veel functionaliteit, ook veel flexibiliteit. Omdat de producten in deze categorie minstens de helft van de totale markt in handen hebben, is er veel ervaring beschikbaar onder de consultants; hetgeen veel zekerheid geeft bij het plannen en uitvoeren van een project. Over het algemeen zijn deze producten niet goedkoop maar wel erg effectief. De tweede categorie, de database-leveranciers, heeft een heel andere doelstelling: meer gebruik van hun eigen database en zoveel mogelijk voorkomen dat andere databases van andere venders via een zijdeur (zoals bij datawarehousing) bij hun klanten binnenkomen. Ook deze producten zijn rijk aan functionaliteit, anders zouden zij niet kunnen concurreren, maar over het algemeen maken zij gebruik van de specifieke functionaliteit van de onderliggende database. En dat is de bedoeling ook: de klanten vastklinken aan de database. In deze categorie zijn de prijzen heel onduidelijk. De basisfunctionaliteit is vaak gratis (inbegrepen in de prijs van de database) maar extra functionaliteit is wel beschikbaar tegen een meerprijs. Als men besluit hier aparte computers voor op te zetten, wat vaak het geval is met datawarehousing, dan moet de database software voor de nieuwe machines ook aangeschaft worden – logisch, maar niet goedkoop. De derde categorie, de BI-leveranciers, is op zoek naar aan BI-gerelateerde producten die omzet genereren, die aan de bestaande klanten verkocht kunnen worden. Het probleem bij de meeste BI-leveranciers is dat hun producten dusdanig belangrijk geworden zijn, dat veel grote leveranciers (zoals Oracle, Microsoft en SAP) hun markt beschouwen als een interessante groeimarkt. Het eerste resultaat van deze extra concurrentie is druk op de prijzen

en dat betekent dat men andere omzetbronnen moet vinden om de kosten te kunnen blijven betalen. Het tweede resultaat van meer concurrentie is dat men de klanten beter van dienst wil zijn, en het leveren van het ETL-proces is een logische stap in de dienstverlening. Omdat men hier concurreert met de eerste twee categorieën, moeten de producten van goeden huize komen om een kans te maken – en dat is niet altijd het geval. Zowel Business Objects als SAS hebben heel veel functionaliteit in hun producten zitten en zijn een concurrent voor de andere categorieën; een aantal andere BI-vendors heeft er aanmerkelijk meer moeite mee gehad. Hun producten zijn bedoeld om omzet te genereren en dat is ook te zien in de prijzen. De producten zijn weliswaar goed geïntegreerd met hun BI front-ends maar over het algemeen vrij duur. Wel hebben wij begrepen dat enige flexibiliteit in de prijs mogelijk is, met name als men overweegt om een product van een van de database-leveranciers aan te schaffen. De vierde categorie is wat moeilijker te beschrijven. Soms betreft het conversie-tools, soms echte ETL-tools maar voor een beperkte omgeving, soms zijn het producten die eigenlijk bedoeld zijn voor iets anders, maar hebben ze ETL-functionaliteit in zich. Als breed inzetbaar ETL-product kunnen zij meestal niet gebruikt worden, maar in specifieke gevallen zijn ze goed inzetbaar.

Van ETL naar EIM

Naar aanleiding van het onderzoek in 2005 is de discussie ontstaan of de traditionele batch-achtige ETL-functionaliteit wel of niet gecombineerd zou moeten worden met de real-time verwerking van wat men toen EAI (Enterprise Application Integration)

ETL-matrix	PURE ETL							IBM DB2 Datawarehouse Edition
	DT/Studio - Embarcadero	Group 1 Software / Data Flow	IBM Information Server	Informatica PowerCenter	Oracle Data Integrator (Sunopsis)	Pervasive Business Integrator	Transformation Manager	
Company								
Start Date Company	1993	1981	1911	1993	1996	1986	2002	1911
Where is the Head Office	USA	USA	USA	USA	USA	USA	UK	USA
Sales started in	2002	1996	1996	1996	1998	1986	2000	1993
Customers WW	200	2000	5000+	2700	420	15000	30	100+
Installations WW	150	>2000	-	20000	1900	50000	100	100+
Installations NL	9	2	100+	200	3	100	0	5
Office in the Benelux	yes	yes	yes	yes	yes	yes	no	yes
Tool								
Platforms	5	4	5	6	7	5	10	4
Current Version	3.1	6	8.0.1	8.1.1	4.1	8.12	4.01	9.1
Stand-alone or Integrated	Stand-alone	Stand-alone	Stand-alone	Stand-alone	Stand-alone	Stand-alone	Stand-alone	Integrated
Engine-based / codegenerator	eb	eb	eb	eb	cg	eb	cg	cg
Type	process	process	process	process	map	map	map	process
Ease of Use								
- user friendly	+	++	+	0	+	-	-	+
- WYSIWYG	-	+	+	0	0	+	0	0
- screen design	+	+	+	+	+	0	+	+
- compatibility ETL / EAI	+	+	++	+	+	-	-	+
Total score userfriendliness	2	5	5	2	3	-1	-1	3
Clarity and re-usability								
- re-usability of components	yes	yes	yes	yes	yes	yes	yes	yes
- modularity	yes	yes	yes	yes	no	yes	yes	no
- user-defined functions	yes	no	yes	yes	yes	yes	yes	yes, db
- comments on selected objects	no	no	yes	no	no	no	no	no
Number of functions in this section	3	2	4	3	2	3	3	1,5
Debugging								
- step by step running	yes	no	yes	yes	no	yes	yes	no
- row-by-row running	yes	yes	yes	yes	no	yes	yes	no
- breakpoints	yes	no	yes	yes	no	yes	yes	no
- software watchpoints	yes	yes	yes	yes	no	yes	yes	no
- compiler / validate	yes	no	yes	half	yes	yes	yes	no
Number of functions in this section	5	2	5	4,5	1	5	5	0
Realtime ETL/EAI/Web services								
- integration batch / real-time	yes	no	yes	yes	yes	yes	no	no
- changed data mechanisms**	log	no	mq+log+trig	mq+log+trig	mq+trig	mq+trig	no	mq+log+trig
- on-demand data integration	no	no	yes	yes	no	yes	no	no
- use the same metadata	yes	no	yes	no	yes	yes	no	no
Number of functions in this section	3	0	6	5	4	5	0	3
Functionality								
- splitting datastreams / multiple targets	yes	yes	yes	yes	no	yes	yes	yes
- conditional splitting	yes	yes	yes	yes	no	yes	yes	yes
- union	yes	yes	yes	yes	no	yes	yes	yes
- pivoting	yes	yes	yes	yes	no	yes	no	yes
- depivoting	yes	yes	no	yes	no	yes	no	no
- key lookup's in memory	no	yes	yes	yes	yes	yes	yes	yes, db
- key lookup's re-usability in different processes	no	no	half	yes	yes	yes	no	yes, db
- slowly changing dimensions*	bh	bh	auto	wizard	auto	bh	bh	bh
- scheduler	yes	yes	yes	yes	yes	yes	no	yes, db
- error handling within job	no	yes	half	no	no	yes	yes	yes
- impact analysis	yes	yes	yes	yes	yes	yes	yes	yes
- data lineage	no	yes	yes	yes	no	no	yes	yes
- automatic documentation	yes	no	yes	yes	yes, pdf	yes	yes	yes
- support voor data-mining models	no	no	no	yes	no	no	no	yes, db
- support voor analytical functions	no	yes	no	yes	no	yes	no	yes, db
Number of functions in this section	8	11	12	13	6,5	12	8	10,5
Data sources/targets								
- support for joined tables as source	yes	yes	yes	yes	yes	yes	no	yes
- built-in functions for data quality	no	yes	yes	yes	yes	no	no	no
- built-in functions for data validation	yes	yes	yes	yes	yes	yes	no	yes
- data profiling	yes	no	yes	yes	no	yes	no	yes
- changed data capture	yes	no	yes	yes	yes	no	no	yes
Number of functions in this section	4	3	5	5	4	3	0	4
- native connections (-ODBC -flat files)	4	7	41	9	21	80	0	1
- packages / enterprise applications	1	5	6	3	7	10	1	0
- real-time connections	1	0	3	4	6	4	0	1
Architecture and infrastructure								
Parallel processing								
- Symmetric Multiprocessing (SMP)	no	yes	yes	yes	yes	yes	no	yes, db
- Massively parallel processing (MPP)	no	no	yes	no	yes, db	no	no	yes, db
- Cluster Aware	yes	no	yes	half	no	yes	no	yes, db
- Grid	no	no	yes	yes	no	no	no	yes, db
Scalability								
- Job distribution	yes	yes	yes	yes	yes	yes	yes	yes
- Data pipelining	no	no	yes	yes	yes	no	no	yes, db
- Partitioning	yes	yes	yes	yes	yes	no	no	yes, db
Base architecture***	m	h	h+d+m	h+d+m	h+d	h+m	h	d
- end-to-end BI infrastructure	half	yes	no	yes	half	no	no	yes
- CWM-support	yes	no	yes	yes	no	yes	yes	yes
- version management	no	no	yes	yes	yes	yes	yes	no
Number of functions in this section	4,5	4	9	8,5	6	5	2	6
Calculations								
Total number of basic functions	27,5	22	41	39	23,5	33	18	25
Userfriendliness	2	5	5	2	3	-1	-1	3
Number of supported platforms 0, 1, 2 or 3	1	1	1	2	2	1	3	1
Number of supported sources / targets 0, 1 or 2	0	1	2	1	2	2	2	0
Number of supported packages	0,5	2,5	3	1,5	3,5	5	0,5	0
Messaging 0 or 1	1	0	1	1	1	1	1	1
Total number of points	32	31,5	53	46,5	35	41	23,5	30

* bh = by hand
 ** mq = message queueing log = database logging or journals trig = database triggers
 *** h = hub & spoke d = distributed m = multihub/spoke

Functionality Points
 half = 0,5 pnts yes = 1 pnts
 (incl. yes, but...) auto = 2 pnts

Vier categoriën tools

1. Pure ETL.

- DT/Studio-Embarcadero
- Group I Software/Data Flow
- IBM Information Server
- Informatica PowerCenter
- Oracle Data Integrator (Sunopsis)
- Pervasive Business Integrator

2. Database integrated.

- IBM DB2 Data Warehouse Edition
- Microsoft Integration Services
- Oracle Warehouse Builder

3. BI integrated.

- Business Objects Data Integrator
- Cognos Data Manager
- Information Builders Data Migrator
- SAS Data Integration Suite

4. Niche ETL.

- Datamirror Transformation Server
- IKAN ETL4ALL

noemde. Bij het laatste onderzoek is het in ieder geval zichtbaar geworden dat de perceptie van de leveranciers is dat dit een must is. Eén vendor van BI-software wist ons te vertellen dat ETL niet meer bestaat – alles is EIM (Enterprise Information Management). Vanuit zijn standpunt gezien heeft hij waarschijnlijk gelijk, maar de klant is nog steeds geïnteresseerd in ETL. De klant heeft nu de keus om zijn data te transformeren en te laden wanneer hij dat ook maar wil, zowel real-time als in batch, als in een combinatie van beide.

Interessant blijft de vraag wie de gebruiker is. En dan bedoel ik niet alleen wie het product koopt en vanuit welk budget, maar wie er achter het scherm zit om te specificeren welke data waar vandaan komen en waar naartoe gaan. In een gesprek met Business Objects over de gebruiksvriendelijkheid van hun product (die heel hoog is) werd duidelijk dat een deel van hun markt de business-analist is, niet iemand die getraind is als programmeur of systeem-beheerder, maar de eigenaar van de data. De verschillen in gebruiksvriendelijkheid bij de producten zijn vrij groot. Een aantal producten is bij uitstek bedoeld voor de professionele IT'er en zijn voor die doelgroep ook erg geschikt. Deze producten zijn echter niet of nauwelijks te gebruiken door een niet-automatiseerder.

Van flat files tot obscure bestanden

Andere grote verschillen zijn te vinden bij het aantal 'vreemde' typen bestanden die men lezen kan, buiten flat files en ODBC. Dit varieert van 'geen enkele' bij Microsoft die alles leest via OLE-DB, tot '83' bij Information Builders die in staat is de meest obscure bestanden te lezen. Erg belangrijk als je vanuit dergelijke bestanden in moet lezen, volkomen overbodig als je ze niet hebt. Het

aantal platforms waarop men draait gaat ook van 0 tot 7, ook daar is het alleen belangrijk om te weten dat jouw platform ertussen zit. Voor de eerste keer dit jaar hebben wij de leveranciers de vragen schriftelijk laten beantwoorden en hen ook laten tekenen om toestemming te geven voor publicatie van de resultaten. Ook hebben wij de vragen dit jaar in het Engels gesteld, om overleg met hoofdkantoren gemakkelijker te maken. Dit heeft, bij een aantal leveranciers, geleid tot andere antwoorden op vragen die wij ook in het verleden gesteld hadden. Functies die er vroeger wel waren zijn inmiddels verdwenen en het is onduidelijk of dit komt door interpretatieverschillen of door het feit dat het nu op papier moet.

Datakwaliteit

In vergelijking met het onderzoek van 2005 besteden de leveranciers veel meer aandacht aan 'Data Quality' en 'Data Cleansing'. Mede naar aanleiding van de grote schandalen bij Enron en Worldcom hoor je nu veel meer over Sarbanes-Oxley en Basel-II. Er wordt veel aandacht besteed aan de juistheid van de data. Veel producten hebben óf data cleansing routines in zich óf bieden de mogelijkheid om dit soort producten op te roepen om datavervuiling tegen te gaan. Experts in de markt roepen al jaren dat informatie alleen zin heeft als die accuraat is, en zelfs dat verkeerde informatie nog erger is dan geen informatie. Nu zien wij, voor de eerste keer op een breed front, dat dit probleem aangepakt wordt bij het bouwen en onderhouden van de databases en datawarehouses.

Performance

Een van de gebieden waar er nauwelijks vooruitgang is geboekt is performance. Op het gebied van parallel processing, job distribution en data pipelining hebben de database-leveranciers alle gewenste functionaliteit en, op enkele uitzonderingen na, de rest niet. Interessant genoeg vindt niemand dat een probleem; het wordt over het algemeen gezien als een functie van de database en niet van de ETL/integratie-tool. Van zowel leveranciers als klanten hebben wij gehoord dat performance niet zo veel problemen geeft; of we zetten niet genoeg in de datawarehouses, of we zetten genoeg machinekracht in zodat het werkt.

Conclusie

De conclusie is dat wij volgend jaar weer gaan kijken, maar dan met een aantal nieuwe vragen over data cleansing, gebruikersgemak voor eindgebruikers en online-faciliteiten. Het onderzoek zal in 2008 waarschijnlijk de naam EIM (Enterprise Information Management) dragen.

Norman Manley is managing partner van Passionned.

Discussiëren met Norman Manley over de ETL-matrix? Dat kan! Zie voor meer informatie de mededeling op pagina 47.