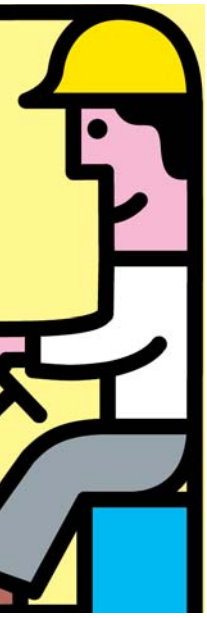


Twee echt serieuze spelers springen eruit

Open Source ETL

Jos van Dongen

Het blijft verbazingwekkend om te zien hoe actief de open source wereld is. Zelfs op het gebied van gespecialiseerde software voor Extractie, Transformatie en Laden (ETL) is een groot aantal projecten actief. In dit artikel wordt dieper ingegaan op de meest gangbare open source ETL-tools en wordt ook gepoogd een antwoord te vinden op de vraag of deze pakketten de vergelijking met hun vaak dure closed source concurrenten kunnen doorstaan.

Zoals in eerdere artikelen is geschreven is Java dé taal voor de ontwikkeling van veel open source tools. Logisch eigenlijk, als men bedenkt dat hiermee gewaarborgd wordt dat de pakketten op een veelheid aan platforms kunnen draaien.

Marktoverzicht

De markt voor OS ETL-tools laat dan ook vrijwel alleen maar Java-gebaseerde software zien. De basisarchitectuur van deze tools is gemeenschappelijk en heeft ook overeenkomsten met de closed source concurrentie. Simpel gezegd zijn er vier componenten die overal terug komen: een engine die het feitelijke ETL-werk verricht, connectoren naar bron- en doelsystemen, een set van (standaard) transformaties en een metadata repository van waaruit alles wordt aangestuurd.

Een eerste belangrijk verschil wordt gevormd door de wijze waarop ETL-jobs worden gedefinieerd. Er is slechts een klein aantal tools dat beschikt over een 'drag and drop' GUI zoals we die ook in de commerciële pakketten aantreffen, en een groot aantal tools waarbij jobs met behulp van XML dienen te worden gedefinieerd. Deze laatste tools zijn met name gericht op software-ontwikkelaars, en minder op gebruik door BI consultants.

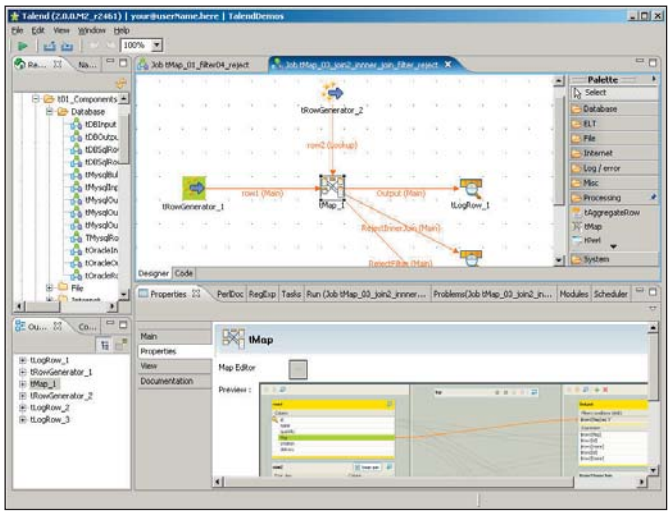
Het tweede belangrijke verschil is of de tool stand-alone werkt of onderdeel uitmaakt van een breder platform.

Het eerste dat opvalt bij de complete ETL-tools is dat de consolidatie die in de closed source wereld gaande is niet ongemerkt aan de open source partijen voorbij gaat. Het Belgische K.E.T.T.L.E. (recursief acroniem voor Kettle Extract, Transport, Transform and Load Engine) is sinds een jaar onderdeel van het eerder besproken Pentaho, en Talend is in januari ingelijfd door Jaspersoft. Beide producten worden verderop in dit artikel uitvoerig behandeld.

Wanneer we kijken naar de 'GUI-loze' stand-alone ETL-oplossingen zijn de meest bekende en gebruikte tools Ketl (niet te verwarren met K.E.T.T.L.E.) en CloverETL. Ketl van KineticNetworks was tot voor kort een closed source product, maar ook in dit geval is onderkend dat een snellere adoptie en ontwikkeling van software van een kleinere partij beter gewaarborgd is bij de open source community. Voor Clover is sinds enige tijd een GUI beschikbaar. Hoewel deze gratis is voor niet-commercieel gebruik is het geen open source product. Erg duur is CloverGUI trouwens niet: voor 2790,- euro koopt u twee jaar support plus een jaar recht op updates.

> De consolidatie in de closed source wereld gaat niet ongemerkt aan de open source partijen voorbij

Bij de ETL-oplossingen die onderdeel uitmaken van een groter geheel is een aantal interessante ontwikkelingen gaande, die wel eens richtinggevend zouden kunnen worden voor ETL-tools in het algemeen. Allereerst is daar LucidDB, een verticaal gepartitioneerde open source database die specifiek is ontwikkeld voor BI-doeleinden. LucidDB wil een compleet dataplatform bieden voor BI- en DWH-doeleinden en heeft daartoe de ETL-voorzieningen in de database engine geïntegreerd. Het product leent zich ook voor 'virtual datawarehouses' waarbij met behulp van 'wrappers' elke databron als een LucidDB schema gebruikt kan worden. Denk hierbij niet alleen aan csv-bestanden, maar ook aan de Salesforce.com web service, waarbij elk Salesforce-object als een tabel benaderd kan worden. Alle ETL-transformaties kunnen vervolgens met SQL worden gedefinieerd.



Afbeelding 1: Talend interface.

De Enhydra-organisatie benadert ETL meer als middleware en biedt Octopus aan als onderdeel van een compleet applicatieplatform. Een vergelijkbare ontwikkeling is te vinden binnen het Glassfish project dat een nieuwe open source Java EE 5 applicatieserver aan het bouwen is. De 'ETL Integrator' bestaat uit een ETL service engine en een grafische ETL Editor. Met de editor worden ETL-projecten gemaakt die vervolgens als JAR file uitgevoerd worden door de service engine. Op dit moment is de *wish list* nog beduidend langer dan de beschikbare features (slowly changing dimensions worden bijvoorbeeld nog niet ondersteund), dus voor serieus gebruik is dit product nog niet geschikt.

Talend

Dit bedrijf is relatief nieuw en heeft pas in oktober 2006, na drie jaar ontwikkelen, versie 1 van zijn ETL-tool op de markt gebracht. Talend heeft de zaken meteen groots aangepakt en laat dat op zijn Amerikaans graag weten ook, hoewel het een Frans bedrijf is.
Er zit voor enkele miljoenen durfkapitaal in het bedrijf, dus er is blijkbaar steeds meer vertrouwen in het open source businessmodel. Talend onderscheidt zich op een aantal vlakken van de meeste open source ETL-tools. Ten eerste is niet Java, maar Perl (Practical Extraction and Report Language) de primaire code voor de transformaties, hoewel er vanaf versie 2.0 wel voor Java gekozen kan worden. Ten tweede heeft men aardig wat werk gemaakt van de on line documentatie en tutorials om potentiële gebruikers snel op weg te helpen. Ook de interface ziet er, zoals alle op Eclipse gebaseerde tools, erg strak uit. Het installeren gaat wat moeizamer, omdat er ook een Perl interpreter gedownload en aan de praat gekregen moet worden. Bij het installeren kan zowaar gekozen worden voor Nederlands als taal. Helaas geldt dit alleen voor de installatieroutine en niet voor de uiteindelijke installatie. Nu ken ik geen enkele consultant of ontwikkelaar die het prettig vindt om in een Nederlandstalige interface te werken, maar het zou wel een onderscheidende feature zijn.
Bijzonder is de mogelijkheid om een 'business-model' te maken. Dit is een hoog-niveau grafische weergave van de te ontwikkelen oplossing. Leuk, maar zo lang er nog geen directe relatie is met de job designs die het werk uitvoeren is dit niet echt nuttig. Als gekeken wordt naar het 'palette' met daarin alle beschikbare Talend-componenten ziet het er in eerste instantie redelijk compleet uit. Wat verder kijkend vallen enkele punten op. Er is in de huidige productieversie 1.1 bijvoorbeeld slechts een beperkte set database-connecties beschikbaar: naast ODBC worden alleen Oracle, MySQL en PostgreSQL ondersteund. Ook de beschikbare transformaties zijn een stuk minder uitgebreid dan bij bijvoorbeeld Kettle. Het werkpaard is de 'tMap' component waarmee de meest voorkomende joins, lookups, merges en conditionele splitsing van resultaatsets opgezet kunnen worden (zie afbeelding 1). Toch lijkt Talend met name een tool voor (Perl) ontwikkelaars. De tool genereert Perl script en de ontbrekende functionaliteit zal in veel gevallen gecompenseerd worden door custom 'routines' op te nemen in de repository, of gebruik te maken van de Perl-componenten in de jobs. Het opbouwen van complexe routines vereist wat denkwerk vooraf omdat alles in 'jobs' ondergebracht moet worden. Jobs kunnen vervolgens wel weer andere jobs aanroepen, etcetera. Gelukkig kan de repository vrij in mappen en submappen worden ingedeeld om het overzicht te bewaren. Het achterliggende programmeerparadigma en de Eclipse-basis heeft echter ook zijn voordelen. Conditionele uitvoering, debugging, breakpoints, 'step by step' verwerking, logging en error handling: het zit er allemaal in. En er wordt hard gewerkt om nog meer standaard componenten onder te brengen.

> Talend lijkt met name een tool voor (Perl) ontwikkelaars

Meer informatie

Op internet kunt u meer informatie vinden over de genoemde open source ETL-tools.
CloverETL: http://cloveretl.berlios.de/index
CloverGUI: www.clovergui.net
Glassfish project:
www.glassfishwiki.org/jbiwiki/Wiki.jsp?page=ETLSE
Ketl: www.ketl.org
Kettle: www.ibridge.be
LucidDB: www.luciddb.org
Octopus: www.enhydra.org/tech/octopus/index.html
Talend: www.talend.com
Bent u geïnteresseerd in de kleinere spelers die vaak niet verder gaan dan het verwerken van XML documenten, dan biedt www.manageability.org/blog/stuff/open-source-etl/view een compleet overzicht van tools.

In gesprek met Matt Casters

Matt is 38 jaar, woonachtig in Brussel en inmiddels zo'n 12 jaar bezig met BI en Datawarehousing. In 2001 heeft hij de stap gemaakt naar het zelfstandig ondernemerschap en wilde zichzelf onderscheiden van de massa en de marktwaarde van zijn bedrijf vergroten door een eigen ETL-tool te ontwikkelen. Na een jaar van brainstormen en ontwerpen heeft Matt met enkele collega's een eerste versie gebouwd.

Gedurende de eerste drie jaar van het bestaan van Kettle was dit voornamelijk een éénmans hobbyproject, dat vooral in de avonden en weekends verder is ontwikkeld. Pas eind 2005 heeft Matt de beslissing genomen om het product te 'open sourcen' om hiermee een bredere acceptatie mogelijk te maken. Wat ook meespeelde was het besef dat het zeer moeilijk bleek om een kosteneffectief commercieel alternatief voor bijvoorbeeld Informatica te bieden, met name vanwege de hoge marketing-inspanningen die dit zou vergen. Met 'open' bedoelt Matt ook écht open. Hij heeft bewust gekozen voor een oplossing die door middel van plug-ins makkelijk uitbreidbaar is. Door de keuze voor de 'Lesser' GPL-licenctie zijn ook commerciële bedrijven in de gelegenheid om proprietary plug-ins te ontwikkelen, en deze samen met Kettle te verkopen zonder dat de code ook 'open' gemaakt dient te worden. Begin 2006 is via één van deze plug-in ontwikkelaars het contact met Pentaho ontstaan en daarna ging het snel. Matt heeft het eigendomsrecht op de Kettle software verkocht aan Pentaho en is nu zelf in dienst bij dit bedrijf als Chief Data Integration Architect.

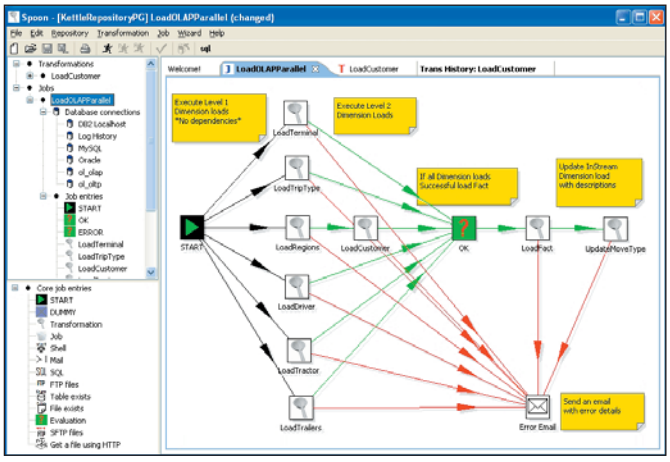
Zijn werkgebied strekt zich sindsdien wat breder uit, zo is hij ook verantwoordelijk voor de metadata-laag waardoor ad hoc reporting inmiddels mogelijk is binnen Pentaho. Ambitie is er nog voldoende: versie 3 die eind dit jaar uit moet komen zal diverse nieuwe componenten bevatten die nu nog alleen in commerciële high-end producten beschikbaar zijn. Denk hierbij aan shareable metadata, impact-analyse, data lineage, data profiling, integratie met data mining en voorzieningen voor versioning, team development en life cycle management. Voor de nog langere termijn staat het punt 'supermetadata' op de wensenlijst, dat wil zeggen het genereren van de metadata voor de transformaties uit business rules, en niet meer op basis van de techniek.

Hoewel Matt veel vanuit huis kan werken zullen zijn dagen dus voorlopig goed gevuld blijven, al was het alleen maar vanwege de 150 e-mails dagelijks, de conference calls in de avonduren, regelmatige bezoekjes aan Orlando en het actief deelnemen aan de forums. Wilt u op de hoogte blijven van alle Kettle-ontwikkelingen en Matt's belevenissen, dan is een regelmatig bezoekje aan Matt's blog op www.ibridge.be onontbeerlijk.

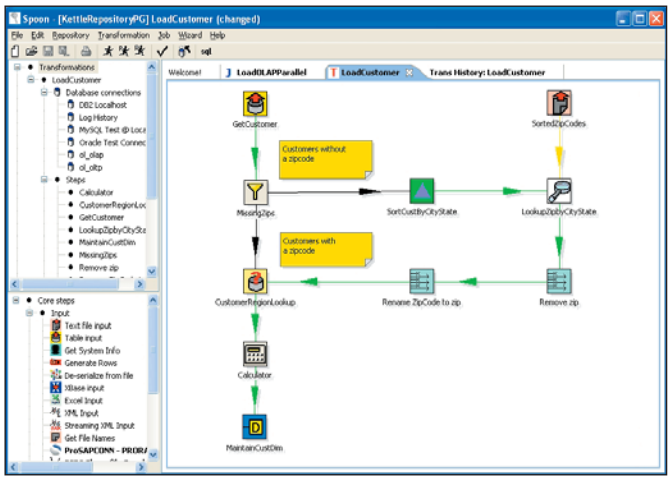
Versie 2 gaat onder andere koppelingen met webservices, file comparison, diverse native database-koppelingen, een grafische SQL builder en Java-ondersteuning bieden. Het is ook mogelijk om eigen componenten toe te voegen, dus er zullen ongetwijfeld ook de nodige third party componenten gaan verschijnen.

K.E.T.T.L.E.

De officiële naam wordt meestal zonder puntjes gewoon als Kettle geschreven (niet te verwarren met het eerder genoemde Ketl). Aan de wieg van deze Java-tool staat voormalig BI-consultant Matt Casters die eind 2001 is gestart met het ontwikkelen van Kettle, zie het kader. Kettle is na ruim vijf jaar aanbeland bij versie 2.4 en biedt een uitgebreid pakket aan functionaliteit waarmee alle voorkomende datatransformaties kunnen worden gerealiseerd. Kettle bestaat uit een aantal onderdelen die vernoemd zijn naar zaken die bij een 'ketel' horen, zoals 'Spoon' voor het maken van jobs en transformaties, 'Pan' en 'Kitchen' voor het uitvoeren van (scheduled) transformaties en jobs, en 'Carte' voor het remote uitvoeren van deze jobs. Voorheen was er ook nog een 'Chef' om meerdere jobs te kunnen integreren, maar deze functionaliteit is inmiddels opgenomen in 'Spoon'.
De basis wordt gevormd door de centrale repository waarin alle metadata worden vastgelegd, maar er kan ook zonder repository gewerkt worden. Het opzetten van jobs en transformaties wijst zich eigenlijk vanzelf, maar de user interface bevat nog wel wat ruwe kantjes. De helpfunctie is bijvoorbeeld niet vanuit het menu aan te roepen (laat staan contextgevoelig), selecteren en aanpassen van sommige onderdelen werkt niet altijd even intuïtief (maar dat went snel) en omdat een query builder ontbreekt is het ofwel SQL krassen, ofwel een tweede tool ernaast gebruiken voor dit doel. Ook op sommige high-end features als debugging, data profiling, team development en versioning zult u nog even moeten wachten, hoewel er voor de laatste drie prima open source producten voorhanden zijn.
Verder bevat Kettle zo ongeveer alles wat u van een ETL-tool mag verwachten, en meer. Vooral in combinatie met een MySQL target database zijn zaken beschikbaar die alleen in de duurdere



Afbeelding 2: Kettle job.



Afbeelding 3: Kettle transformatie.

ETL-tools voorhanden zijn, zoals clustering en partitioning. Het lijstje in de ETL-matrix betreffende 'functionaliteit' bevat weinig zaken die niet ondersteund worden. En voor zover functies niet standaard geleverd worden zijn ze altijd te maken: alles wat met een database, SQL of Javascript kan, is in principe voorhanden. Documentatie wordt niet automatisch gegenereerd, maar het staat iedereen natuurlijk vrij om vanuit de repository rapportages te genereren. Qua connectiviteit zal niet snel misgegrepen worden, omdat er default van JDBC gebruik gemaakt wordt, terwijl voor Oracle ook de native OCI-connectie gebruikt kan worden. De lijst met ondersteunde databases is indrukwekkend (22 stuks), en wat meteen opvalt is dat zowel met typische low-end (Dbase III, MS Access) als high-end producten (Netezza en Teradata) en alles daar tussenin gewerkt kan worden.

> Kettle biedt een uitgebreid pakket aan functionaliteit

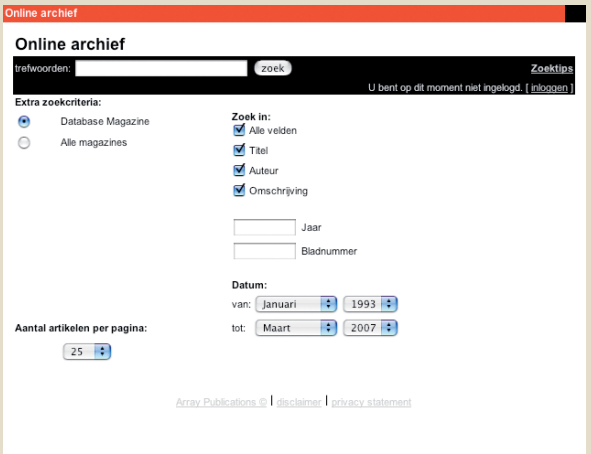
Het voert wat ver om in dit artikel alle functionaliteit te beschrijven, maar zaken als slowly changing dimensions, pivot/de-pivot, merge joins, table comparisons, database & stream lookups en field splitters zijn allemaal beschikbaar. Een compleet voorbeeld van het opbouwen van een datamart is te zien in afbeelding 2 en 3. Wat ook opvalt is de 'installatie' procedure. Het volstaat om het Kettle zip-bestand te downloaden en uit te pakken. Als er al een Java Virtual Machine aanwezig is kunt u eenvoudigweg Spoon.bat starten en aan het werk, in dit geval zelfs in het Nederlands mocht u dat willen.
Voor degene die twijfelt aan de volwassenheid van dit product nog het volgende. Klant van het eerste uur is het Vlaamse Verkeerscentrum dat Kettle inzet voor de verwerking van alle informatie die de detectielussen in 1500 weggedeeltes genereren. Kettle draait hiervoor 24 uur per dag, 7 dagen per week elk kwartier een job die de nieuwe data ophaalt en verwerkt. Het gaat hier om een database van meer dan 500 GB groot en bijna 3 miljard rijen. Zelfs grote banken en verzekeraars gebruiken Kettle voor bedrijfskritische toepassingen, maar zijn wat huiverig om deze informatie naar buiten te brengen. Ook partijen als Netezza en NCR hebben actief meegewerkt om Kettle geschikt te maken voor koppelingen met hun producten.

Conclusie

Hoewel er een grote diversiteit is aan ETL-achtige oplossingen in de open source wereld zijn er twee echt serieuze spelers. Talend is de 'new kid on the block' en moet zich nog bewijzen in de praktijk. Dit zal ongetwijfeld gaan lukken als ETL-onderdeel van Jaspersoft en met een partner als Capgemini. Kettle is van deze twee de meest bewezen oplossing en heeft ondanks de wat minder strakke interface de meeste functionaliteit nu al standaard aan boord.
Het zijn echter wel twee verschillende oplossingen: zoekt u een programmeerachtige code-generator met uitgebreide debug-functionaliteit, dan lijkt Talend een logische keuze, maar zoekt u een complete engine based ETL-tool, dan is Kettle op dit moment nog niet te kloppen.

Jos van Dongen
Jos van Dongen (jvdongen@tholis.com) is Senior Consultant bij Tholis Consulting.

Online archief Database Magazine



Database Magazine-lezer opgelet! Artikelen over onderwerpen als Datawarehousing, SQL, ETL, Business Intelligence, Relationele databases, modellering en nog veel meer vindt u in het Online Archief van Array Publications. Vaktijdschriften als Storage Magazine, Database Magazine, IT Service Magazine, Java Magazine en ons Oracle vakblad Optimize hebben hun artikelenarchief online gezet. Met een Google-achtige zoekstructuur vindt u snel wat u zoekt op www.dbm.nl