

Kiezen in een omgeving met een dynamische markt

# ETL-tool of.. Kalido?

Paul van der Linden



**Bij het ontwikkelen van de wat grotere, complexere datawarehouses wordt vandaag de dag niet lang stil gestaan bij de aankoop van een ETL-tool. Waar de bewerkingen die gedaan moeten worden op de brondata een niet-triviaal karakter hebben en het vermoeden bestaat dat tussentijds ook nog sprake zal zijn van veranderingen in bewerkingen of gegevensbronnen, is de keuze voor een ETL-tool al gauw gemaakt.**

Een aantal jaren geleden werd over zo'n aankoopbeslissing nog gewikt en gewogen. Zeker, de noodzaak werd ingezien, maar anderzijds ging het toch om een aanzienlijk bedrag van honderd-duizenden euro's. Geconstateerd kan worden dat ETL-tools in de praktijk zodanig zijn bevallen, dat over zo'n aanzienlijk investering niet meer lang wordt nagedacht. Met enige overdrijving zou je zelfs kunnen stellen dat zo'n aankoopbeslissing bijna 'gedachte-loos' wordt gemaakt.

Toch legt de aankoop van een ETL-tool beslag op een belangrijk deel van het beschikbare IT-budget. De voordelen van zo'n tool staan zoals gezegd buiten kijf. Maar inmiddels is er sprake van een andere vraag die beantwoord dient te worden. Deze luidt: koop ik een ETL-tool of wordt het... Kalido?

## NIEUWE AFWEGING

Kalido? Dankzij de "onderwater marketing" van Kalido is het eerst noodzakelijk om enige aandacht te besteden aan het product Kalido, om dan vervolgens de afweging tussen een ETL-tool en Kalido centraal te stellen. Op de functionaliteit die ETL-tools bieden zal niet expliciet worden ingegaan omdat deze in globale termen bekend zal zijn. Daarnaast gaat het hier ook niet om een functioneel vergelijk tussen ETL-tools en Kalido.

Kalido Dynamic Information Warehouse is de volledige naam voor een product dat vanaf 1986 is ontwikkeld door het voormalige Shell Services International (SSI). Als grote multinational had Shell voortdurend te maken met het bij elkaar voegen en uit elkaar

trekken van informatie als gevolg van organisationele veranderingen, overnames of fusies. Die werden op hun beurt weer veroorzaakt door de veranderende markten waarin Shell opereert. De problemen die ontstaan in termen van informatievoorziening zullen bekend zijn: datastructuren die niet te combineren zijn, verschillende coderingen gebruikt voor dezelfde kenmerken, verschillende tijdsindelingen (kalenders) en valuta's in gebruik, verschillende tijdzones en verschillen in de historie (wijzigingen) die wordt vastgehouden.

In de loop van de tijd werd het SSI duidelijk dat alleen door de veranderingen die door de business werden ingegeven als uitgangspunt te nemen, voorkomen kon worden dat telkens weer dezelfde soort wijzigingen moesten worden doorgevoerd op de databasestructuren. Kalido kun je zien als een data-integratie-product dat als het ware op een database geklikt kan worden. Momenteel worden Oracle, MS SQL Server en DB2 ondersteund. Het inmiddels verzelfstandigde SSI (= Kalido) biedt hiermee de mogelijkheid om veel soepeler te voldoen aan veranderende vragen uit de business. Hoe gebeurt dit?

## GENERIEK DATA MODELLEREN

Kalido maakt gebruik van *generic data modeling*. Dit betekent dat data op een hoog niveau worden gemodelleerd, waarbij de structuur van de data eveneens als data wordt opgeslagen. Dit in tegenstelling tot de reguliere manier van modellering, waarin de datastructuur wordt doorgevoerd in de fysieke structuur van de

*Kalido kun je zien als een data-integratie-product*

database. Verandert nu de structuur van de data, dan betekent dat in het geval van Kalido dat een nieuwe entry plaatsvindt in een tabel. Het generieke datamodel blijft derhalve intact en verandert

niet. In de reguliere (non-Kalido) manier van werken leidt zo'n datastructuurwijziging tot een wijziging in het fysieke database-model; een veel ingrijpendere operatie.

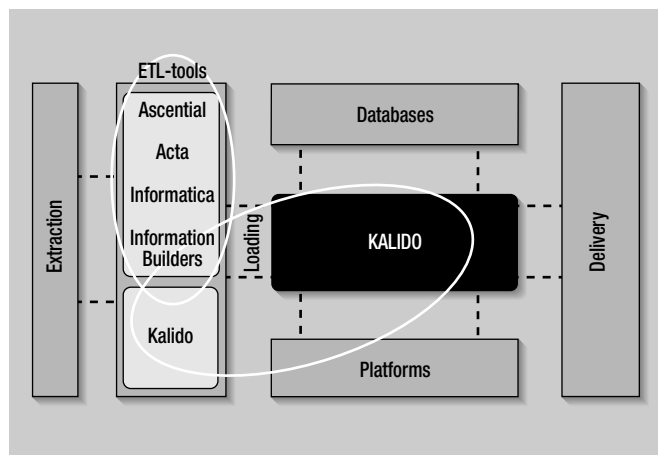
Aspecten van het generieke datamodel zijn de ondersteuning van meerdere hiërarchieën, meerdere kalenders, ondersteuning van hiërarchieën met variabele diepte en het herkennen van structuurwijzigingen, die zich in de loop van de tijd hebben voorgedaan.

*Zoals wel vaker liet Inmon door deze uitspraak zien een aantal jaren achter te lopen*

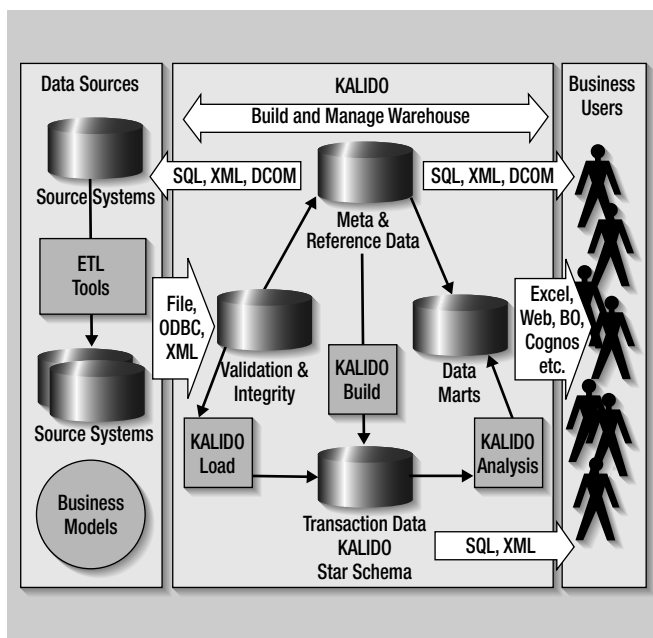
Data worden in Kalido zoveel mogelijk opgeslagen in twee tabellen. Ook dit is een geheel andere wijze om met data om te gaan. In een reguliere situatie leidt bijna elke entiteit tot een aparte database-tabel. Kalido maakt overigens onderscheid in drie soorten gegevens: transactionele data, business entiteiten en metadata. De manier waarop Kalido werkt is dat eerst bepaald dient te worden om wat voor soort gegevens het gaat. Vervolgens kan dan het generieke schema worden gegenereerd. Dat generieke schema heeft de vorm van een sterschema. Dit heeft te maken met de performance.

Het terrein dat door Kalido wordt afgedekt betreft het modelleren, vullen, beheren en onderhouden van de datawarehouse database. Maar ook business modellering, metadata management, alsmede data-aggregatie, datamart generatie en het archiveren van data behoren tot de functionaliteit die door Kalido wordt geboden. Daarnaast beschikt Kalido ook over beperkte ETL-functionaliteit. Hierbij kan gedacht worden aan het controleren van een data feed en het opslaan van foutieve records. Aan de voorkant is het mogelijk om op basis van doorgezette metadata bijvoorbeeld een BO Universe aan te laten maken dan wel een MS Analysis Services of PowerPlay kubus.

Omdat Kalido vertrekt vanuit de aanname dat de bedrijfscontext voortdurend verandert en de mogelijkheden biedt om daar soepel mee om te gaan, kan het met recht het eerste business



AFBEELDING 1: PLAATS VAN KALIDO EN ETL-TOOLS.



AFBEELDING 2: DE KALIDO DATA ARCHITECTUUR.

driven datawarehouse genoemd worden. Binnen Shell is inmiddels sprake van zo'n honderd implementaties in 85 landen. In Nederland zijn daarnaast Philips en Unilever gebruikers van Kalido.

**DE AFWEGING ETL-TOOL OF KALIDO**

Een tool als Informatica (als voorbeeld van een ETL-tool) en Kalido proberen allebei een bijdrage te leveren aan het zo goed mogelijk omgaan met veranderingen. Dat betekent zowel de initiële afhandeling daarvan als het verdere beheer. Natuurlijk doen ze dat beide op hun eigen terrein.

Bij ETL-tools gaat het om het traject van het ontsluiten van de verschillende gegevensbronnen tot en met het wegschrijven van de uiteindelijke resultaten vanuit de staging area naar de datawarehouse database. Met daar tussenin handelingen zoals data-transport, data-validatie en -manipulatie, data-combinatie en data recovery. Kalido richt zich op het modelleringsaspect.

Veranderingen worden aangegeven in een businessmodel en uiteindelijk vertaald naar een fysiek datamodel. Dat businessmodel kan, mag en zal voortdurend veranderen. De opzet van Kalido is dat het effect hiervan op het fysieke datamodel zo goed als afwezig is. Het zal duidelijk zijn dat deze vertaalslag (logica) de kern vormt van wat Kalido is.

Aangezien ETL-tools en Kalido beide een stuk stabiliteit bieden en in orde van grootte hetzelfde, niet geringe, beslag leggen op het beschikbare budget (300.000 euro) ligt het voor de hand dat het niet realistisch is om aan te nemen dat Informatica en Kalido tegelijkertijd aangeschaft kunnen worden. De vraag rijst dan in welke situatie je de voorkeur zou moeten geven aan de een, dan wel de ander.

## AANZET TOT EEN ANTWOORD

Functioneel gezien bestaat er slechts een geringe overlap tussen een tool als Informatica en Kalido. De inschatting geldt dus meer de situatie waarin het datawarehouse wordt gebouwd. Een keuze voor Kalido past in een omgeving waarin sprake is van een dynamische markt, die ertoe leidt dat de logische relatie tussen gegevens voortdurend wijzigt. Ook als de gegevens die we willen vastleggen of de historie die we willen bewaren, nogal eens verandert is Kalido een goede keuze. Daarnaast past Kalido goed bij multinationals die gegevens, die van overal vandaan komen,

### Daarnaast beschikt Kalido ook over beperkte ETL-functionaliteit

moeten consolideren en weer naar de verschillende landen-organisaties toe in (ook) lokaal formaat terug rapporteren.

In een situatie waarin deze aspecten niet of nauwelijks aan de orde zijn, maar wel veel verschillende gegevensbronnen het datawarehouse voeden, is een keuze voor een ETL-tool logischer. Zeker als de bewerkingen die op de aangeleverde data moeten plaatsvinden, een complex karakter hebben. In de tabel onder aan de pagina is een aantal overwegingen opgenomen.

Gegeven het beschikbare budget zal het duidelijk zijn dat de organisatie die voor Kalido kiest, het ETL-traject (voorlopig?) invult door gebruik te maken van de ETL-functionaliteit die een database vaak biedt, of teruggrijpt naar handmatig gemaakte programma's. Gezien de betrekkelijke eenvoud van ETL in deze situatie, zou dit weinig problemen moeten opleveren.

Waar gekozen wordt voor een ETL-tool zal de datawarehouse database op een klassieke manier worden vormgegeven. Dat betekent dat business-veranderingen meestal zullen leiden tot

noodzakelijk ingrijpen bij de fysieke databasestructuur. Maar de inschatting is hier dat het aantal keren dat zoiets aan de orde is redelijk beperkt zal zijn en daarmee hanteerbaar.

## NADELEN VAN KALIDO

Kalido heeft hoog ingezet en ziet de Fortune 2000 als haar prospectlijst. Daar hoort een hoge prijs bij, maar klaarblijkelijk ook "stealth marketing". En dat is jammer omdat Kalido een interessant product is dat ook goed past bij organisaties die niet op deze lijst staan.

Zoals altijd geldt ook hier dat een voordeel ook een nadeel is. Kalido houdt alle wijzigingen van alle kenmerken (attributen) bij. In een situatie dat een organisatie beschikt over enorme aantallen klanten, waarvan men ook nog eens een heleboel weet (leeftijd, inkomensgroep, woonplaats, bestedingsgedrag, gezinssamenstelling, betalingsgedrag – oftewel veel occurrences met veel kolommen) wordt wel heel veel data opgeslagen. In termen van opslagruimte en responsnelheid is dat een uitdaging, maar wel oplosbaar.

De samenwerking tussen Informatica's PowerCenter en Kalido blijkt in de huidige versies van de producten nog voor verbetering vatbaar te zijn. Hier wordt aan gewerkt; hetgeen ook verwacht mag worden van beide leveranciers.

## CONCLUSIE

Toen Bill Inmon een tijdje geleden Nederland aandeed, vertelde hij dat Kalido een geweldig product is waar hij erg enthousiast over is. Om er vervolgens aan toe te voegen dat er volgens hem maar een handje vol bedrijven is die behoefte zouden hebben aan Kalido. Hiermee doelde hij op het beperkt aantal 'echte' multinationals. Zoals wel vaker liet Inmon door deze uitspraak zien een aantal jaren achter te lopen op de rest van de wereld. De functionaliteit die Kalido biedt is immers interessant voor alle organisaties die frequent in de fysieke databasestructuur moeten ingrijpen, om te kunnen voldoen aan de veranderende eisen uit de business. En dan hebben we het niet uitsluitend over multinationals.

ETL-tools en Kalido bieden beide de functionaliteit om met veranderingen om te gaan. Hiermee leveren ze een potentieel belangrijke bijdrage aan de ontwikkeling, het beheer en het onderhoud van een datawarehouse. Ze richten zich echter op verschillende soorten van veranderingen. ETL-tools grijpen aan bij veranderingen in de IT-sfeer, met name veranderende gegevensbronnen of veranderingen in de bewerkingen die op gegevens moeten plaatsvinden. Kalido gaat uit van veranderingen in de business zelf. Beide tools kunnen dus heel goed aanvullend optreden, maar aangezien ze nogal prijzig zijn is het alleen de hele grote ondernemingen gegeven om ze beide aan te kunnen schaffen. Dat betekent dat een afweging gedaan moet worden of (eerst) een ETL-tool wordt aangeschaft dan wel Kalido.

Situatie	ETL-tool of Kalido ?
Complexe gegevensbewerkingen	ETL-tool
Voortdurende veranderingen in de wijzigingen (historie) die bijgehouden moeten worden	Kalido
Gegevensbronnen die worden gebruikt veranderen constant	ETL-tool
Vergelijkbare gegevens, ook na veranderde invulling van dimensies (bijvoorbeeld: verkoopdistrict)	Kalido
Hanteren van verschillende kalenders, valuta's, tijdzones naast elkaar	Kalido

## Een voorbeeld uit de praktijk: Philips CE

Philips Consumer Electronics (CE), onderdeel van Philips, is één van de grootste leveranciers van consumentenelectronica. Philips CE is op zichzelf beschouwd al een multinational met vestigingen over de hele wereld. Ze produceert en levert een grote waaier van producten, van flatscreen-televisies tot DVD-recorders. Tot de subdivisies van Philips CE behoren mainstream (audio, tv, vcr, dvd), digitale netwerken (set-top boxes) en communicatieproducten (bijvoorbeeld gsm's). Philips CE is goed voor tussen de 30 en 40 procent van de totale omzet van Philips.

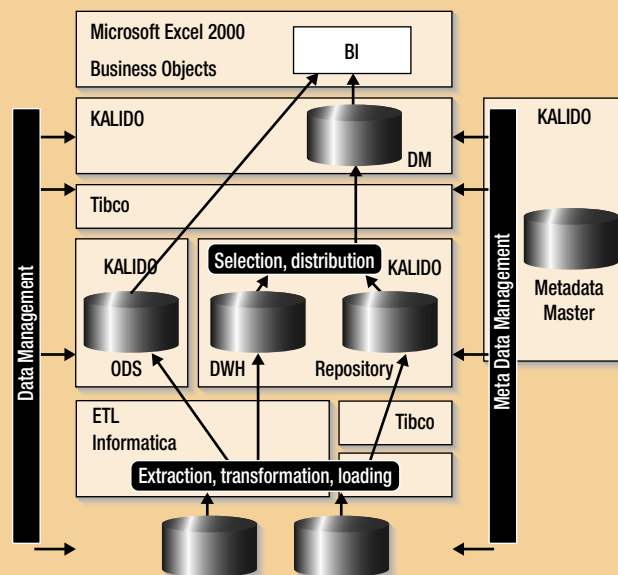
De locale operationele units en de business units van Philips CE verzorgden voorheen hun eigen datawarehousing, waarbij vaak lokaal geselecteerde tools werden gebruikt. Dit leidde ertoe dat dezelfde gegevens op meerdere plaatsen werden opgeslagen. Ook waren hierdoor de IT-kosten hoger dan noodzakelijk en nam het aantal interfaces exponentieel toe.

Besloten werd dat deze situatie verandering behoeft. Hiertoe werd gekozen voor een shared services model waarbij de CE Information Office optreedt als informatiemakelaar tussen de business vraag naar informatie en het aanbod van data. Hierdoor is de efficiëntie van datawarehousing binnen de organisatie inmiddels verbeterd. De aanpak die is gekozen is sterk pragmatisch van karakter.

Het pragmatische bestaat eruit dat de Data Warehouse groep de diversiteit in datastructuren die in de business voorkomt als uitgangspunt neemt. Er wordt dus niet geprobeerd om eerst een universeel datamodel met alle business units af te spreken. Evenmin werd ervoor gekozen om één enkelvoudig centraal datawarehouse te realiseren. Indien gesproken wordt over Enterprise Data Warehousing, dan gaat het in feite om een aantal verschillende datawarehouse implementaties, die op maat zijn gemaakt voor specifieke business behoeften. Hierbij wordt wel steeds van dezelfde architectuur en tooling uitgegaan.

Bij het opzetten van de architectuur en het kiezen van de tools, speelde de vraag of gekozen moest worden voor een ETL-tool (Informatica) dan wel voor Kalido. Uitgaande van de situatie van Philips CE en de markten waarin zij opereert, is die keuze uitgevallen ten gunste van Kalido. Overigens werd het in een later stadium mogelijk om ook Informatica aan te schaffen.

Als overweging kan het volgende gelden. Is er sprake van zeer complexe en uitgebreide bewerking van gegevens: kies dan een ETL-tool. Ook als de aard van de bewerkingen constant verandert of er steeds andere gegevensbronnen worden gebruikt dient eerder aan een ETL-tool gedacht te worden. Zit de dynamiek echter in de gegevensstructuur (veranderende relaties tussen gegevens, veranderende indelingen, veranderende behoefte aan de historie die je bijhoudt), dan ligt Kalido eerder voor de hand. En dan resteert natuurlijk altijd nog de 'koninklijke oplossing'; zorgen dat je



AFBEELDING 3: DE TECHNISCHE ARCHITECTUUR BIJ PHILIPS CE.

Han Slaats, programmamanager Enterprise Data Warehousing bij Philips CE: 'Doordat we werken vanuit een centraal service-model en gebruik maken van een consistente architectuur en zeer flexibele technologieën zoals Kalido en Informatica, krijgen we een snelle ROI op datawarehousing-projecten, een staande organisatie die continuïteit in ontwikkeling en support garandeert en een start van wat we een 'toegevoegde waarde-ladder' noemen. Hierbij voegt elk nieuw project waarde toe door gebruik te maken van de resultaten van het vorige project. De inschatting van Han Slaats is dat het gebruik van Kalido Philips CE zo'n 40 tot 50 procent van de datawarehouse ontwikkel- en onderhoudskosten bespaart.

Philips CE gebruikt Kalido voor meerdere projecten. Om welke specifieke projecten het gaat, wil men echter binnenskamers houden. Philips CE is uiterst tevreden over de besparingen die Kalido oplevert en ziet voorlopig nog geen eindpunt in het

voldoende budget hebt om ze beide aan te schaffen. Maar dat is niet iedereen gegeven.

*Met dank aan Han Slaats (Philips CE) en Ad Stam (Atos Origin) voor hun commentaar op een eerder concept van dit artikel.*

Paul van der Linden (Paul.PFH.vanderLinden@AtosOrigin.com) is senior consultant Data Warehousing/BI bij Atos Origin.