Anders dan het gebruikelijke ETL

Oracle EL-T

Jaap Jan Bakker

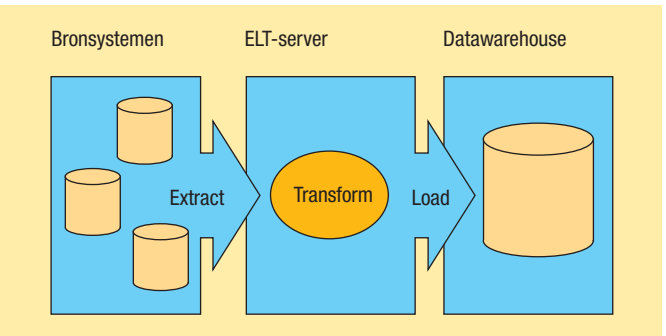
Oracle presenteerde op 12 februari de Oracle Integration Suite, een verzameling tools voor de integratie van data uit verschillende bronnen binnen een SOA.

Een prominent onderdeel van deze suite is de Oracle Data Integrator (ODI), een ETL-tool welke in 2006 aan de Oracle portfolio werd toegevoegd door de overname van Sunopsis, een van de talrijke overnames die Oracle gedaan heeft in de afgelopen periode. Voordien stond de ODI bekend onder naam Data Conductor. In het Technical Whitepaper over de Data Integrator (Oracle, december 2006) wordt door Oracle een aantal kenmerken belicht die het product in gunstige zin zouden onderscheiden van de concurrentie. Een van deze kenmerken is dat de ODI gebaseerd is op een EL-T architectuur, in tegenstelling tot de gebruikelijke ETL architectuur. De claim is dat met EL-T een betere performance is te realiseren tegen lagere kosten in vergelijking met een conventionele ETL werkwijze. Deze claim is op zich niet nieuw: ook bij de marketing van Sunopsis werd dit al nadrukkelijk genoemd, zij het dat het streepje toen nog na de E werd geplaatst: E-LT. In dit artikel wordt het verschil tussen traditionele ETL en het door Oracle gepropageerde EL-T beschreven.

ETL

Met de afkorting ETL wordt meestal in algemene zin het proces van data-overdracht en -verwerking beschreven binnen een datawarehouse-omgeving: data worden uit een of meerdere bronsystemen uitgelezen (Extract), ondergaan vervolgens een aantal bewerkingen zoals integratie van gegevens uit verschillende bronnen, opschoning en aggregatie (Transform), en worden tot slot in het doelsysteem ingeladen (Load).
In het beeld dat Oracle schetst van de tegenstelling tussen ETL versus EL-T, wordt de term ETL in een meer specifieke zin gebruikt: in dit verband gaat het om een bepaalde fysieke inrichting van het proces, waarbij voor de transformatie gebruik wordt gemaakt van een speciaal voor dit doel ingerichte server, uitgerust met een ETL tool van een bepaalde leverancier. Als voorbeelden van deze traditionele ETL aanpak worden in het Technical Whitepaper Informatica's Powercenter en IBM's DataStage met name genoemd.
Bij deze ETL aanpak, zie afbeelding 1, worden de verschillende stappen volgordelijk uitgevoerd: eerst worden de data uit de verschillende operationele systemen verzameld op de ETL server (Extract). Vervolgens vinden daar de noodzakelijke bewerkingen plaats die de ruwe brongegevens omvormen naar informatie die als basis kan dienen voor de rapportages en analyse. Na deze transformatiestap worden de data tot slot verplaatst van de ETL servers naar het datawarehouse.

Er worden door Oracle twee nadelen genoemd die met deze conventionele ETL opzet gepaard gaan: een gebrekkige performance en overbodig datatransport. Als eerste en belangrijkste nadeel wordt op de aard van de verwerking op de ETL server gewezen: gegevens worden daar op een individuele, rij-voor-rij basis verwerkt. Deze inefficiënte manier van werken heeft, zeker bij grote datavolumes, een zeer nadelige invloed op de performance. Ten tweede wordt gewezen op het dubbele transport van data: eerst moeten de gegevens immers verplaatst worden van de bronsystemen naar de ETL-server, en na verwerking moeten de data weer overgebracht worden naar het uiteindelijk datawarehouse. Dat betekent een dubbele belasting van het netwerk. Daar komt nog bij dat het transformatieproces op de ETL server niet op zichzelf staat: bij de bewerking zijn immers vaak gegevens nodig uit het datawarehouse. Zo zullen referentiegegevens zoals bijvoorbeeld codetabellen op de een of andere manier naar de ETL server moeten worden gehaald, om daar te worden gecombineerd met de nieuw verzamelde brondata. Het gevolg is een toename van het netwerkverkeer.
Vaak worden om redenen van performance een of meerdere stappen niet meer binnen de ETL omgeving uitgevoerd, maar verplaatst naar het datawarehouse. De implicatie daarvan is



Afbeelding 1: Traditionele ETL aanpak.

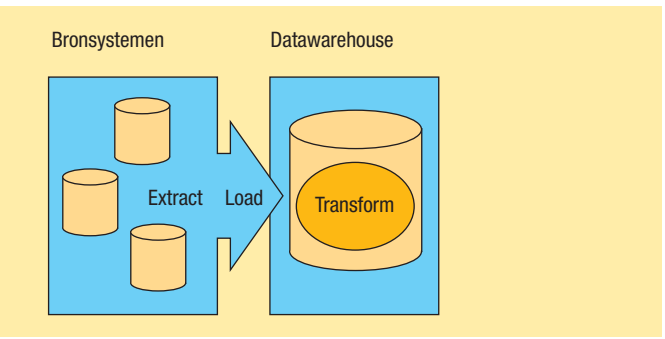
echter dat er code buiten de ETL tool om ontwikkeld wordt, waardoor de bruikbaarheid van de binnen de tool opgeslagen metadata op losse schroeven komt te staan.

EL-T

Met de EL-T werkwijze verandert zowel het *waar* als het *hoe* van de transformatie. Er wordt hier geen gebruik meer gemaakt van een speciaal voor dit doel ingerichte server, maar de transformaties worden na het laden uitgevoerd in het datawarehouse, zie afbeelding 2. Op deze manier worden de eerdergenoemde twee nadelen van de klassieke ETL werkwijze teniet gedaan.

> Er is geen apart datatransport meer nodig van ETL server naar datawarehouse

Ten eerste kan voor het transformatieproces in het datawarehouse optimaal gebruikt worden van de mogelijkheden die de database engine biedt. De belangrijkste daarvan is de mogelijkheid van set-bewerkingen die binnen RDBMS engines standaard voorhanden zijn: in plaats van een enkele rij kan met een bewerking een grote hoeveelheid records tegelijk verwerkt worden. Hiermee zijn grote voordelen in de performance te halen in vergelijking met de rij-voor-rij verwerking op de separate ETL server. Daar komt bij dat databases de laatste jaren meer en meer uitgerust zijn met extra technologieën die gebruikt kunnen worden bij het inlezen, de opslag, de verwerking en het bevragen van grote hoeveelheden data. Daarbij kunnen we denken aan zaken als XML ondersteuning, partitionering, compressie van data en indexen, materialized views, bulkverwerking, paralelle verwerking, analytische functies binnen SQL etcetera. Met een EL-T opzet kunnen deze features ook ingezet worden voor een zo optimaal mogelijke transformatiestap.
Ten tweede is er geen apart datatransport meer nodig van ETL server naar het datawarehouse, de brongegevens worden direct naar hun uiteindelijke bestemming verplaatst. Ook de benodigde aanvullende gegevens zijn binnen het datawarehouse direct voorhanden. Zo kunnen bijvoorbeeld de nieuwe data direct gecombineerd worden met de lookup-tabellen in het datawarehouse. Datatransport en netwerkbelasting worden zo beperkt tot het minimaal noodzakelijke.
De tegenstelling is overigens niet zo strikt als door Oracle wordt voorgesteld: ook 'traditionele' ETL tools bieden tegenwoordig de optie om bepaalde bewerkingen niet op de ETL server uit te voeren, maar in de database. Zo kent bijvoorbeeld Informatica de zogenaamde 'pushdown' optie waarmee onderdelen van de transformatie naar de bron- of doel-database kunnen worden verplaatst. Ook IBM's DataStage heeft deze mogelijkheid om onderdelen van de verwerking naar de database te verplaatsen. In afbeelding 3 wordt een aantal kenmerken van beide benaderingen naast elkaar gezet.



Afbeelding 2: EL-T aanpak.

EL-T	ETL
transformatieproces in datawarehouse	verwerking op aparte ETL server
set-verwerking	rij-voor-rij verwerking
extra belasting datawarehouse	load balancing ETL server en datawarehouse
minder datatransport, alle noodzakelijke data bij elkaar	transport tussen ETL server en database
verificatie en schoning na laden	filtering mogelijk voor laden
geen aparte server nodig	aparte ETL server + software

Afbeelding 3: Beide aanpakken naast elkaar.

Conclusie

De transformatie binnen het ETL proces is uit oogpunt van doorlooptijd en verwerkingscapaciteit veruit de meest intensieve stap. Hier ligt de bottleneck in het ETL traject. Het verbeteren van de performance door gebruik te maken van de kracht van de database server is de belangrijkste claim van de EL-T aanhangers. Met de overname van Sunopsis is de term EL-T ook in het Oracle woordenboek bijgeschreven. Dat is eigenlijk merkwaardig, omdat er al veel langer een Oracle toepassing beschikbaar is die ook volgens dit principe werkt: de Oracle Warehouse Builder. In de context van dat product is echter altijd gewoon gesproken over ETL. De Data Integrator biedt volgens Oracle de voordelen in performance van een EL-T opzet met behoud van de voordelen die het gebruik van een ETL tool biedt: een geïntegreerde grafische ontwikkelomgeving, die is gebaseerd op een repository met metadata. Omdat er geen aparte ETL server meer nodig is, zijn de initiële kosten en onderhoudskosten significant lager. Met de EL-T werkwijze wordt immers gebruik gemaakt van de al bestaande database servers. Wel zullen deze voldoende capaciteit moeten hebben om de extra belasting aan te kunnen, zowel qua verwerking als opslag.

Jaap Jan Bakker is Oracle Database Consultant bij Atos Origin.