Chef-koks presenteren dimensionele modellen als gerechten

Kimball: ETL tools worden steeds belangrijker

Marc Houtkooper

In het kader van dit themanummer over Datawarehousing heeft Marc Houtkooper een interview gehouden met Ralph Kimball. Daarin heeft ETL centraal gestaan. De relatie tussen ETL en Datawarehousing komt ook aan bod.

Nederland mag trots zijn. Het is dit jaar het enige land buiten de Verenigde Staten waaraan Dr. Ralph Kimball een bezoek brengt om een van zijn cursussen te geven. Tijdens zijn korte verblijf in Amsterdam in september jongstleden, nam hij de tijd om over zijn nieuwe boek te praten en een tiental vragen te beantwoorden.

Als een toonaangevende visionair op het gebied van datawarehousing en stermodellering, waarom deze plotselinge interesse in ETL?

Kimball: "Ik ben altijd al geïnteresseerd geweest in de meer praktische benadering van een datawarehouse en daarbij ontdekte ik dat er vaak weinig aandacht besteed wordt aan de bouw van de ETL-basis. Mijn studenten en cursisten laten me veelvuldig weten dat ze naast modellering van een datawarehouse, juist ook hulp nodig hebben om de achterkant van het datawarehouse te bouwen.

Ik realiseerde me toen dat iedereen wel begrijpt waar de letters E, T en L voor staan, maar dat niemand weet wat het in de praktijk betekent. Blijkbaar is er weinig ervaring op dat gebied bij mijn studenten. Daarnaast houd ik van de uitdaging om listige problemen op te lossen en die in algemene disciplines op te schrijven in een boek over ETL."

Denkt u dat de ETL-markt langzaam verdwijnt onder druk van EBIS (Enterprise Business Intelligence Suite) en AA (Analytical Applications) en hun 'out of the box' ETL-oplossingen? Met andere woorden: Moeten organisaties nog steeds investeren in dure ETL tools?

"Business Intelligence en Analytical Applications moet je zien als het eindresultaat van een goede datawarehouse-omgeving. Ik ben blij dat er de laatste tijd zoveel nadruk gelegd wordt op Business Intelligence en Analytical Applications omdat men daarmee de nadruk legt op de waarde van een datawarehouse en niet op de kosten. Business Intelligence, Analytical Applications, Customer Relationship Management, Business Process Mangement en datamining zijn allemaal verschillende vormen van het aloude DSS-concept (Decision Support System) dat al zo'n dertig jaar meegaat.
Elk van deze systemen heeft een solide basis nodig van kwalitatief hoogstaande en geïntegreerde data en geen van deze zonder zorgvuldige extractie, transformatie en laden van data. Ook al heeft ETL geen *sex appeal*, het vertegenwoordigt minstens 70 procent van de totale ontwikkelinspanning. Dus het antwoord op de vraag is: Nee, de ETL-markt verdwijnt niet; ETL tools worden steeds belangrijker.

> Gegeven de groeiende vraag naar geïntegreerde data zal de ETL-markt groot genoeg blijven

Tijdens het onderzoek voor mijn nieuwe boek, The Data Warehouse ETL Toolkit, welke in september is gepubliceerd, raakte ik meer en meer overtuigd dat de verantwoordelijkheid om een heel ETL-fundament van een Enterprise Data Warehouse te bouwen, eigenlijk te complex is om uit het blote hoofd te doen. Zelfs nu nog worden de meeste ETL-installaties gedaan met handgeschreven scripts en/of programmacode. Toch is de toekomst aan de geautomatiseerde integratie-tools, waar de logica gewaarborgd is en waar de vereiste metadata-tabellen automatisch gegenereerd en beheerd worden."

Analisten beweren dat Microsoft langzaam maar zeker de ETL-markt aan het overnemen is. Hoe denkt u daarover?

"De twee grote spelers in deze markt zijn (of worden) Microsoft en Oracle. De meer gespecialiseerde spelers zijn Informatica, Ascential en Ab Initio (deze laatste voornamelijk in de VS). Cognos en Business Objects hebben hun eigen (optionele) ETL tool, respectievelijk DecisionStream en Data Integrator. ETL is zo belangrijk en fundamenteel, dat mijns inziens de eerste vijf tools in mijn lijstje winstgevend en daarmee succesvol blijven. Gegeven

de groeiende vraag naar geïntegreerde data zal de ETL-markt groot genoeg zijn en blijven voor alle huidige spelers. Microsoft is soms moeilijk te beoordelen. Ze hebben fantastische software en natuurlijk een enorm bereik op alle desktops. Het probleem is dat ze geen inspanning verrichten in de verkoop van ICT-oplossingen aan bedrijven. Oracle doet het daar veel beter. Zowel Microsoft als Oracle bezitten sterke ETL-producten en de strijd tussen die twee zal nog wel enige jaren aanhouden."

De werelden van ETL en Enterprise Application Integration (EAI) groeien steeds dichter naar elkaar toe. Wat denkt u dat er gaat gebeuren met deze twee omgevingen?

"Mijn verwachting is dat ze apart blijven van elkaar. Op dit moment hebben de ETL-leveranciers weinig impact en ervaring op de markt van EAI. Andersom hebben de EAI-leveranciers geen goede architectuur om de grote bulk aan datatransport te realiseren, die nodig is voor het vullen van een datawarehouse. Voor de nabije toekomst zullen de twee omgevingen samen gaan werken, maar wel complementair blijven."

Rapporten tonen aan dat de markt voor data-integratie gezond is en groeiende. Twee voorbeelden hiervan met een grote impact op ETL zijn Radio Frequency Identification (RFID) in de wereld van de logistiek en IAS en IFRS (International Accounting and Reporting Standards) in de financiële wereld.
Wat is uw mening hierover?

"Zowel RFID als 'compliance' hebben een significante impact op ETL. In de Verenigde Staten is compliance momenteel erg actueel. De nieuwe regelgeving vereist dat. Waar het gaat om omzet- en winstcijfers, moet men aantonen hoe deze getallen tot stand zijn gekomen. De ETL-processen worden hierbij geanalyseerd om hun details en afhankelijkheden in kaart te brengen. Met een goede ETL-toepassing is dat natuurlijk geen enkel probleem. Als het gaat om RFID, dan zal de impact op ETL wat langer duren. Ik zie RFID als de nieuwe barcode en daarmee brengt het niets nieuws voor de ETL-wereld. Wel zal er op termijn steeds meer behoefte zijn om de verschillende RFID-systemen te kunnen koppelen om het traceren en analyseren van individuele producten mogelijk te maken."

Wat zijn de belangrijkste criteria om een ETL-proces te kwalificeren? Heeft u wederom twintig selectiecriteria opgesteld, net als bij de kwalificatie van een dimensioneel datawarehouse?

"Mijn mening is dat de wetenschap rond ETL niet zo gestructureerd is, dat we dat zo maar kunnen doen. Misschien over een aantal jaren. In hoofdlijnen kunnen we stellen dat een effectief ETL-systeem te meten is aan: 1. de datakwaliteit; 2. de tijdigheid van de aanleverende systemen; 3. de relevantie van de data voor de afnemers (DSS); en 4. de totale kosten, inclusief de ontwikkel- en beheerkosten."

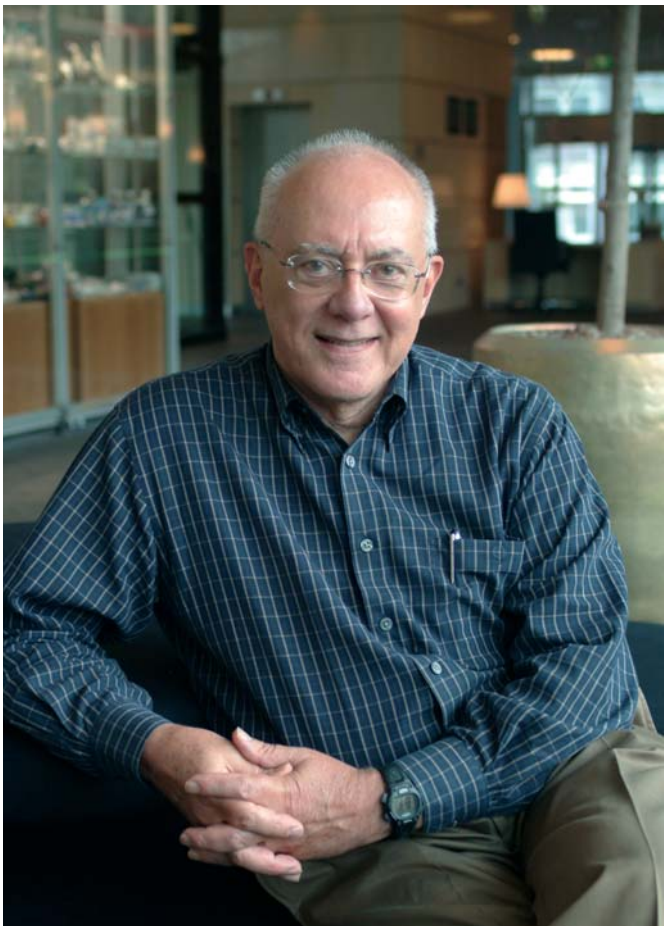

Foto Harry Otto

Dr. Ralph Kimball: "De ETL-markt verdwijnt niet".

Hoe belangrijk zijn metadata voor een ETL-proces?

"Metadata zijn enorm belangrijk. Metadata bepalen welke data geselecteerd moeten worden (extractie), hoe de data geschoond moeten worden, wanneer de data klaar zijn, wat de structuur is van de data, waar men de data naar toe moet (laden) en welke beveiligingsbeperkingen er gelden voor de data."

ETL en datawarehousing lijken onafscheidelijk. Denkt u dat ETL ook voor andere doeleinden gebruikt zou moeten worden?

"ETL is eigenlijk hetzelfde als 'data-verplaatsing' en 'data-integratie'. Dit is ook de reden dat de pure ETL-leveranciers als Informatica en Ascential hierover praten en niet meer over ETL."

In uw boek spreekt u van Extract, Clean, Conform and Deliver (ECCD) in plaats van het traditionele Extract, Transform and Load (ETL). Waarom?

"Zodra je verder kijkt dan de letters E, T en L, dan is er niemand die de details kan benoemen. In mijn boek probeer ik dat wel te doen. De T van Transformatie kan men onderverdelen in: Schoning en Conformeren.

Schoning betekent data profileren, domeincorrecties, tabel-correcties, data-integriteitscorrecties en het bekrachtigen van de business-regels.
Conformeren betekent dat naast het technisch afstemmen, juist ook overeenstemming plaats moet vinden op het gebied van business-entiteiten en -waarden. Conformeren van facts en conformeren van dimensies zijn trouwens totaal verschillende taken. Een volledig geconformeerde data-omgeving geldt als voorwaarde voor een succesvolle implementatie van een datawarehouse.
Zodra er meerdere onderwerpen bij betrokken zijn is conformeren een noodzaak, zelfs al komt alles in een bronsysteem. Dus je ziet, er zit een hele interessante wereld verscholen achter de letter 'T' van ETL.
In mijn boek beschrijf ik het ETL-proces als een keuken van een restaurant. De laatste handeling is, als de chef-kok het gerecht op het bord presenteert, vlak voordat het opgediend wordt. De D van 'Deliver' in ECCD. In mijn boek en tijdens mijn cursussen leer ik de ETL chef-koks hoe gerechten opgediend moeten worden: de dimensionele modellen."

> Conformeren van facts en conformeren van dimensies zijn totaal verschillende taken

Tegenwoordig zijn datawarehouses en BI-omgevingen steeds meer succesvol. Forse investeringen in het verleden lijken zich terug te betalen; analisten durven steeds meer te praten over Return on Investment (ROI). En daarmee doet zich een nieuwe tendens voor: niet alleen managers willen hun afgeleide data gerapporteerd zien, maar ook moet deze informatie gepresenteerd worden in een call center-omgeving of gepubliceerd worden op het Internet. Kan een oud en vertrouwd batch-georiënteerd ETL-proces geconverteerd worden naar een real-time ETL-systeem?

Kimball tot slot: "In het algemeen kan gesteld worden dat een batch-georiënteerde ETL-pijplijn opnieuw gebouwd moet worden om een continue stroom (streaming) aan data te kunnen verwerken. Feitelijk moet alles veranderen. Het originele extractieprogramma kan niet meer zitten wachten op een bestand dat midden in de nacht klaar is. Het nieuwe extractieprogramma moet het EAI-verkeer of de incrementele log-bestanden voortdurend lezen. Tegelijkertijd, midden in het ETL proces, moet bijvoorbeeld een sort routine opnieuw geprogrammeerd worden om ook streaming data te kunnen verwerken. Tot slot zullen ook de query- en rapportage-tools moeten veranderen, omdat de data volgens een push-mechanisme aangeleverd zullen worden, in plaats van de 'pull' van de eindgebruiker.
ETL-leveranciers praten tegenwoordig over hoe batch-georiënteerd ETL-systemen direct overgezet kunnen worden naar real-time streaming ETL, maar zoals gesteld is dit ingewikkelde materie waarbij iedere situatie zorgvuldig bestudeerd moet worden. Een interessante en complexe ontwerpuitdaging."

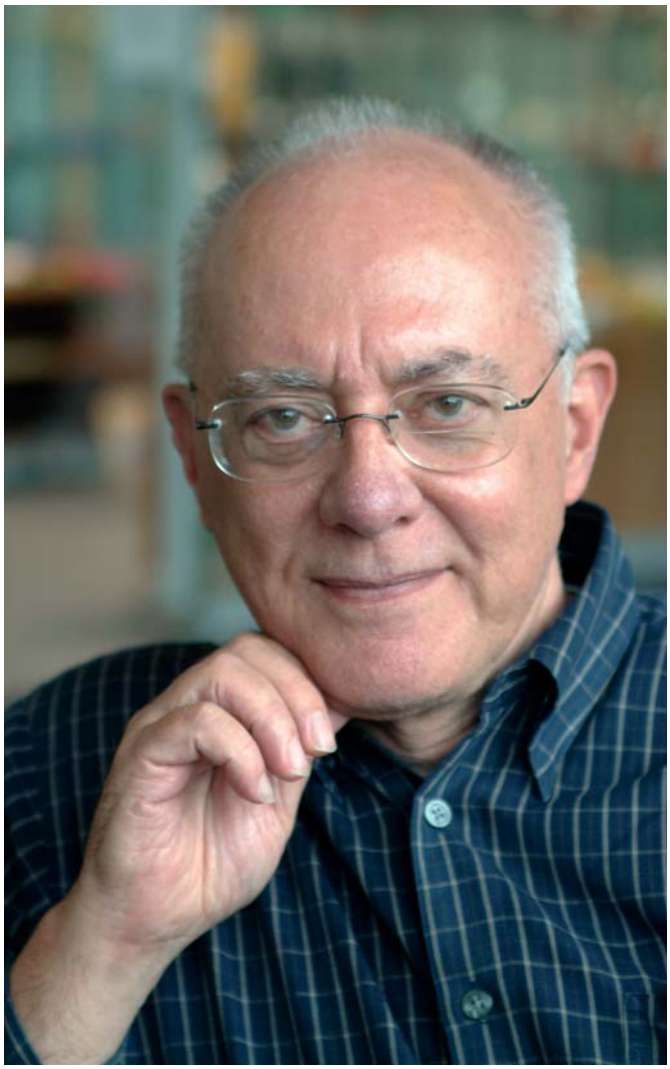

Foto Harry Otto

Kimball: "Interessante en complexe ontwerpuitdagingen".

Conclusies

Uit het interview met Ralph Kimball kan men concluderen dat ETL-processen en hun tools aan belangrijkheid eerder toe- dan afnemen. De ETL-markt is volop in ontwikkeling. Niet alleen worden de ETL tools meer geavanceerd, ook maken ze steeds vaker onderdeel uit van totale Business Intelligence-oplossingen en Analytical Applications. De toenemende vraag naar data-integratieproducten zal de oven in de ETL-keuken nog lang brandend houden.

Marc Houtkooper (marc.houtkooper@newcom.nl) is Senior Consultant Datawarehousing/BI bij Newcom Information Systems