



Idee van een single version of the truth wordt losgelaten

De volgende generatie EDW

Ronald Damhof

Enterprise datawarehouses (afgekort als EDW) hebben een slecht imago. Ze worden gekenmerkt door lange ontwikkeltrajecten, complex onderhoud, ze zijn kostbaar en allesbehalve duurzaam.

De oorzaak van dit slechte imago is tweeledig. Enerzijds heeft het te maken met een datawarehouse-industrie die architecturen neerzet waarvan de verwachtingen niet gehaald worden en anderzijds heeft het te maken met sentiment. Waarbij beide factoren elkaar vanzelfsprekend erg versterken. Dat het EDW niet wordt begrepen moeten de datawarehouse-fabrikanten zich aantrekken. Het is zaak om duidelijk te maken dat het EDW niet synoniem is met duur, log en langdurig, maar aan business case driven, adaptief en duurzaam.

Dit artikel doet een poging het sentiment rond het EDW te doorbreken door vanuit heldere ambities een architectuur neer te zetten die de organisatie in staat stelt haar meest onderschatte productiemiddel duurzaam te gelde te maken: data. Een architectuur die fundamenteel verschillend is van datawarehouses

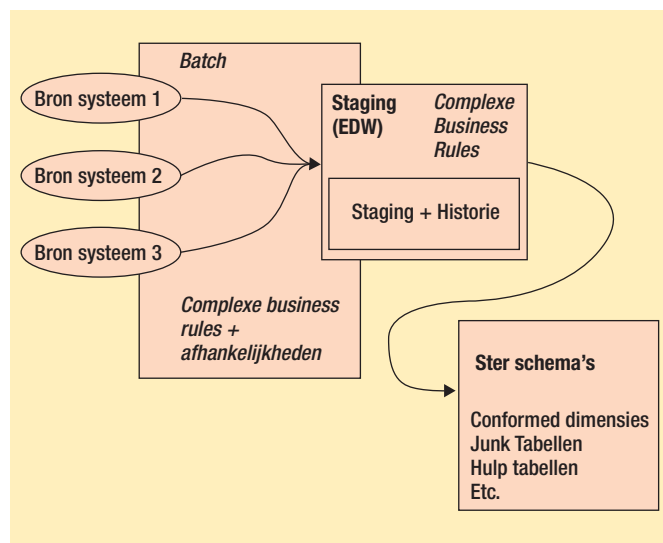
die de laatste decennia zijn gerealiseerd. Het gaat hierbij niet zozeer om de business-rechtvaardiging van een EDW, maar vooral om de ambities van een EDW en de architectuurkeuzes die daarvan het gevolg zijn. Zowel in termen van ambities en architectuur heeft dit artikel de pretentieuze titel 'volgende generatie' gekregen.

Ambities

Het opzetten van een architectuur en infrastructuur van een EDW vraagt allereerst om het duidelijk uitspreken van de ambities. Dit moet niet verzanden in een technisch woud van terminologie. Nee, deze ambities moeten door de organisatie, vooral door het bestuurlijk kader onderschreven – en dus ook begrepen – kunnen worden. Vanzelfsprekend moet achter de ambities een uitgewerkte business case model staan, die het voor de organisatie duidelijk maakt wat de potentiële opbrengsten zijn en wat de terugverdientijd is.

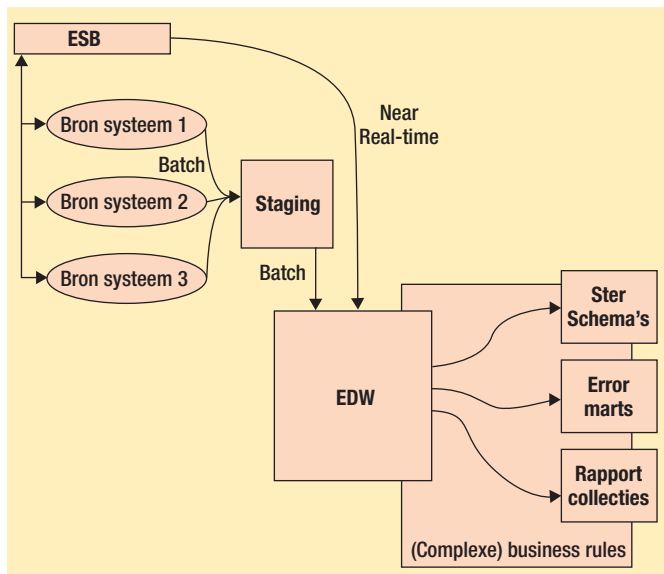
De ambities voor de architectuur en infrastructuur van het nieuwe generatie EDW zijn:

- Volledige traceerbaarheid van gegevens. Gegevens die opgeslagen staan in het EDW en vervolgens gebruikt worden door de organisatie moeten traceerbaar zijn naar de uiteindelijke bron. Ook de data moeten aan de moderne compliance eisen gaan voldoen;
- Betekenisvolle gegevens. Alle gegevens in het EDW dienen vergezeld te zijn van betekenis in termen van definitie, eigenaar, business rule, domeinwaarden etcetera;
- Enige mate van ontkoppeling tussen data en het operationele systeemlandschap. Niet elke verandering moet zwaar resoneren door het EDW heen tot aan de eindgebruiker. Mits goed gemodelleerd kunnen datastructuren in het EDW vele malen stabiel worden opgezet, in vergelijking met de constant



Afbeelding 1: Traditionele architectuur, Business rules voor het EDW.

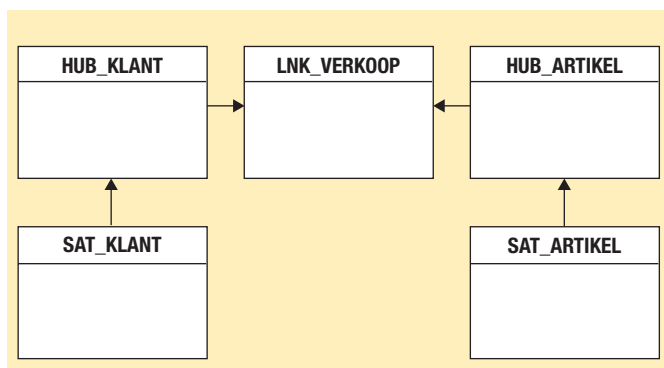
(Bron: Dan Linstedt – The business of Data Vault modeling).



Afbeelding 2: Fundamentele architectuur, Business rules na het EDW. (Bron: Dan Linstedt – The business of Data Vault modeling).

veranderende organisatieprocessen die uiting krijgen in het operationele systeemlandschap;

- Datakwaliteit en business rules zijn de verantwoordelijkheid van de business, niet het domein van het EDW team en vanzelfsprekend ook niet het domein van de ICT. Het EDW moet echter wel voorzien in het duidelijk kunnen scheiden van deze verantwoordelijkheden;
- Schaalbaarheid in al haar vormen. Per definitie is het EDW een incrementeel groeiende architectuur. Meer gegevens, meer gebruikers, meer gebruik. De architectuur moet de infrastructuur in staat stellen om hierin mee te groeien. Met schaalbaarheid wordt hier een schaalbaarheid in resourcing bedoeld, maar ook technische schaalbaarheid. Een voorbeeld hiervan is het streven om de performance van de datalogistieke processen zoveel mogelijk een beslissing van het management te laten zijn. Een EDW-architect moet kunnen zeggen in zijn organisatie dat als hij meer technische resources krijgt, de data sneller geladen kunnen worden. En zo niet, dan gaan de processen langzamer. Zero-update strategieën in combinatie parallelisatie zijn hierin essentieel;



Afbeelding 3.

- Het EDW moet de business in staat stellen om data tegen elkaar te kunnen confronteren ofwel met elkaar te integreren. De nieuwe generatie EDW vereist dat dit 'just-in-time' gebeurt, dus niet op voorhand data proberen te integreren. Met 'just-in-time' integratie wordt bedoeld dat we alleen integreren als de business daarom vraagt. Twee principes zijn hierbij essentieel; business rules worden downstream (naar de eindgebruiker) geïmplementeerd en er moet – tot op het fysieke niveau – een onderscheid worden gemaakt in 'Facts en Truth'. Er bestaat geen 'Single version of the truth';
- Het EDW moet in staat zijn doorlooptijden van gevraagde informatieproducten in toenemende mate te verkorten;
- Gegevens uit datawarehouses worden in toenemende mate operationeel van karakter. Gegevens vanuit het operationele systeemlandschap moeten steeds sneller worden geleverd aan de eindgebruiker (data latency gaat omlaag). Het nieuwe generatie EDW moet voorbereid zijn op een enorm gemixte workload;
- Ook het datawarehouse moet voldoen aan de hedendaagse eisen die gesteld worden aan software engineering. Een belangrijke en zeer bruikbare graadmeter in dit kader zijn de vijf niveaus van het Capability Maturity Model (CMM) opgesteld door het Carnegie Mellon instituut. Een minimale ambitie van een datawarehouse moet level 4 zijn;
- Een ambitie die vaak bij de grotere data-intensieve organisaties een grote rol speelt is het robuust maken van het operationele systeemlandschap. Ongetwijfeld doet deze zin de wenkbrauwen fronsen: wat kan het EDW daarin betekenen? Systemen hebben – op vaak ongecoördineerde wijze – er een rol bijgekregen. Niet alleen moeten ze het primaire proces ondersteunen, vaak worden ze ook belast met vele ad hoc vragen en vele reguliere datasets voor verschillende (externe) gebruikers. Afgezien van de problemen omtrent betekenis, performance, schaalbaarheid en beheer, is dit geen wenselijke situatie. Het nieuwe generatie EDW stelt de organisatie in staat om bronnen eenmaal (goed en volledig) te ontsluiten en veelvoudig te distribueren op een wijze die voldoet aan alle bovenstaande ambities.

Naast het bovenstaande moet een EDW vanzelfsprekend voorzien in optimale performance en gebruiksvriendelijkheid en dienen BI services (rapportage, analyse, mining enzovoort) geformuleerd te worden, waarmee data waardevol ingezet kunnen

HUB_KLANT			
SURR_ID	LOAD_DTS	KLANTNR	RECSRC
1	01-01-2007	ABC123	CRM
2	01-01-2007	XYZ456	CRM
3	25-06-2008	MN0789	CRM

Afbeelding 4.

IMP_VERKOOP				
ARTIKELNR	KLANTNR	AANTAL	PRIJS	TRXDTS
...
1904X5	DNZ999	5	5,95	15-07-2008
...
...

Afbeelding 5.

worden. Uiteengezet wordt op welke wijze deze ambities gehaald worden door anders om te gaan met data en data-logistiek. De principes van Dan Linstedt met zijn Data Vault architectuur spelen hierin een zeer voorname rol.

De data: fact en truth

Datawarehouses hadden tot dusver, impliciet of expliciet, het doel om een *single version of the truth* te creëren voor hun gebruikers. Bedrijven die worstelden met inconsistente informatie afkomstig uit onderling gescheiden datasilo's zagen in het datawarehouse dé oplossing om alle neuzen dezelfde kant op te krijgen.

Gaandeweg ontdekte men echter dat hiermee getracht werd een IT-oplossing te creëren voor wat in essentie een business-probleem is: als de organisatie geen bedrijfsbreed gedragen eenduidige kijk op de gegevens heeft, kan het datawarehouse die niet opleggen. Dan zullen er altijd gebruikers zijn die (vrijwillig of noodgedwongen) een afwijkende kijk op gegevens hebben. Door in het datawarehouse één versie van de waarheid af te dwingen, blijven die gebruikers in de kou staan.

Daarnaast zijn ook andere overwegingen een rol gaan spelen. Sinds het begin van deze eeuw zien bedrijven zich geconfronteerd met de eis tot compliance met regelgeving als Sarbanes-Oxley en Basel-II. Deze eis staat op gespannen voet met het opslaan van geïnterpreteerde, voorbewerkte gegevens. Als het oorspronkelijke gegeven niet meer te traceren valt (en meestal is het datawarehouse de enige plek waar historie bewaard wordt, dus de bron valt af), raken we het 'spoor' kwijt: de gegevens zijn niet meer traceerbaar en het datawarehouse is niet compliant.

In de nieuwe generatie datawarehouses wordt het idee van een

single version of the truth losgelaten. "Truth is in the eye of the beholder", oftewel het is maar hoe je het bekijkt.

In plaats daarvan streven we naar een *system of fact*, een 'single version of the facts' dus, waarop iedereen zijn eigen interpretatie mag loslaten. De single version of the facts biedt ruimte voor multiple versions of the truth. Welke interpretaties de gebruikers ook bedenken, ze zijn altijd terug te voeren op de zuivere feiten – en daarmee voldoet het datawarehouse aan de compliance eis.

Datakwaliteit

De wens om een system of fact bij te houden leidt ook tot een andere kijk op de kwaliteit van de data. Uitgangspunt is dat alle data altijd geladen worden, hoe laag de kwaliteit ook is: *100% of the data 100% of the time*. Redenering hierbij is dat 'lage kwaliteit' eveneens een kwestie van interpretatie is: wat voor sommige gebruikers onbruikbaar is, kan voor anderen meer dan voldoende zijn. Daar komt bij dat het data betreft die daadwerkelijk zijn opgeslagen in een operationeel systeem waarop beslissingen gebaseerd zouden kunnen zijn. Het is dus aan de afzonderlijke gebruikersgroepen om te bepalen of gegevens van voldoende kwaliteit zijn om ze te gebruiken – en om te specificeren of en hoe deze kwaliteit in hun version of the truth verbeterd moet worden.

Integratie en schoning vinden pas na het datawarehouse plaats

De T in ETL staat voor Transformatie en de vraag is waar de bulk van de transformatie gepositioneerd moet worden in het EDW. In de traditionele datawarehouses wordt deze gepositioneerd tussen de staging en het EDW (zie afbeelding 1). Een vaak gehoord argument is dat er schone data in het datawarehouse moeten zitten, de term 'een versie van de waarheid' slaat vaak op dit deel van het datawarehouse.

Op termijn gaat dit echter problemen opleveren:

- de data worden moeilijker/onmogelijk om te traceren;
- de ETL wordt steeds moeilijker te schalen, batch windows moeilijker te halen;

HUB_KLANT			
SURR_ID	LOAD_DTS	KLANTNUMMER	RECSRC
1	01-01-2007	ABC123	CRM
2	01-01-2007	XYZ456	CRM
3	25-06-2008	MN0789	CRM
4	15-07-2008	DNZ999	VERKOOP

Afbeelding 6.

SAT_KLANT			
SURR_ID	LOAD_DTS	KLANTNAAM	RECSRC
1	01-01-2007	ABC VOF	CRM
1	01-06-2008	ABC BV	CRM
2	01-01-2007	Xyz corp	CRM
3	25-06-2008	Acme Inc	CRM

Afbeelding 7.

- de projecten gaan in toenemende mate langer duren, worden complexer en vereisen steeds duurder resources;
- de flexibiliteit van de datawarehouse-architectuur, om zich aan te passen aan de steeds veranderende eisen van de business, stort in;
- heel impliciet zijn er aannames gedaan over het gebruik van historie, deze aannames blijken niet te kloppen;
- testen wordt in toenemende mate moeilijker, vooral regressie-testen.

Het resultaat is een datawarehouse dat relatief snel verouderd is en niet meer in staat is om de business te ondersteunen met kwalitatief hoogwaardige en tijdige data.

Het positioneren van Transformatie voor het EDW is fout. In de volgend generatie EDW moet Transformatie altijd gepositioneerd worden na het EDW (zie afbeelding 2). Bij voorkeur zo dicht mogelijk tegen de eindgebruiker en zijn behoeften aan. Ideaal gesproken zou de Transformatie moeten plaatsvinden op het moment dat de gebruiker daarom vraagt. Dit is echter vaak niet mogelijk door vooral (technische) performance redenen.

Een belangrijk architectureel concept van de volgende generatie EDW is; data worden niet geïntegreerd als de business er niet om vraagt. Als de business erom vraagt worden de data geïntegreerd tussen het EDW en de datamarts. Het EDW zelf wordt gevuld met een-op-een data uit de bronsystemen die parallel, asynchroon en in hoge mate gestandaardiseerd geladen kunnen worden in het EDW.

Parallel laden EDW: het EDW moet gemodelleerd worden volgens de strategie van 'zero-updates'.

Asynchroon laden: alles wat wordt aangeleverd in de staging (werkenheden) moet direct verwerkt kunnen worden.

Gestandaardiseerd: het EDW wordt gemodelleerd volgens een standaard methodiek (Data Vault) met een beperkt aantal verschillende typen entiteiten (Data Vault heeft er drie). Elke type entiteit heeft steeds hetzelfde laadgedrag. Deze mate van standaardisatie zet de deur open voor generatie van ETL. Het is belangrijk dat ETL-leveranciers deze ontwikkeling gaan inzien en ondersteunen met ofwel standaard ETL-templates/mappingen of een vorm van model-driven datawarehousing.

Het laden van data vanuit de bron naar het EDW wordt hiermee 'een lopende band proces' met fabrieksmatige efficiency en effectiviteit. Dit fabrieksmatige proces is in staat om met de gemaakte data-halfabrikaten nagenoeg aan elke vraag uit de business te kunnen voldoen.

Data Vault als enabler

De architectuur van de Data Vault past naadloos bij de geschets-te ambities en architectuur. Zover datawarehouse-professionals/BI-professionals kennism gemaakt hebben met Data Vault, is dat meestal als modelleringsmethodiek (voor een toelichting daarop verwijzen we de lezer naar het uitstekende artikel van Maarten Ketelaars over dit onderwerp in DB/M7, 2005). Het is jammer dat

SAT_KLANT			
SURR_ID	LOAD_DTS	KLANTNAAM	RECSRC
1	01-01-2007	ABC VOF	CRM
1	01-06-2008	ABC BV	CRM
2	01-01-2007	Xyz corp	CRM
3	25-06-2008	Acme Inc	CRM
4	17-07-2008	DeNieuweZaak	CRM

Afbeelding 8.

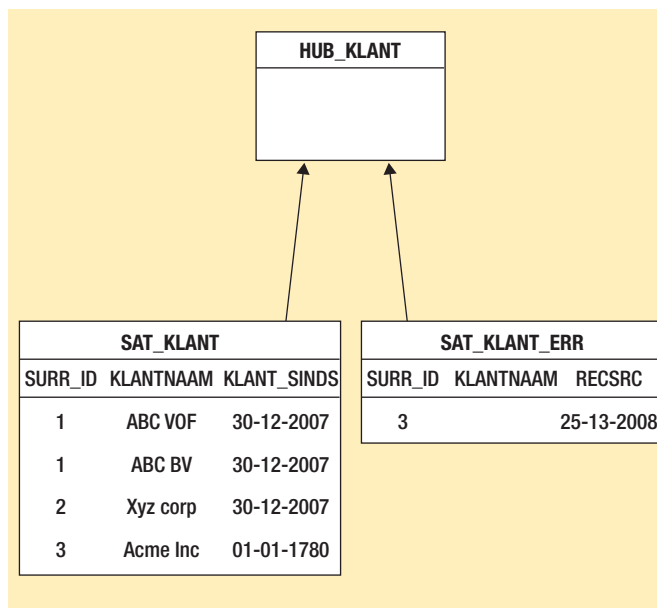
Data Vault vaak verkeerd begrepen, uitgelegd en toegepast wordt. Het unieke van de methodiek is namelijk niet zozeer de gekozen modellering op zich, maar de manier waarop deze de architectuurprincipes ondersteunt, zoals beschreven in dit artikel. Zo is Data Vault bij uitstek geschikt om uitgangspunten als '100% of the data 100% of the time' en 'asynchroon laden' te faciliteren. We lichten dit toe met enkele voorbeelden.

Voorbeeld: problemen met RI.

In het EDW van een denkbeeldige onderneming worden alle verkooptransacties historisch vastgelegd, zie afbeelding 3. Het CRM-systeem van de onderneming is de master-bron voor klantgegevens; maar om het verkoopproces niet in de weg te zitten, kunnen verkopers zelf klantnummers uitdelen, zodat ze verkopen aan nieuwe klanten kunnen registreren in het verkoopsysteem. Deze klantnummers worden achteraf geregistreerd in het CRM-systeem.

De verkooptransacties die aan het EDW worden aangeleverd, kunnen dus een onbekende klantreferentie bevatten.

Op 15-7-2008 ontvangt het laadproces het IMP_VERKOOP bestand, en constateert dat klantnummer DNZ999 nog niet



Afbeelding 9.

voorkomt in HUB_KLANT, zie afbeelding 4 en 5. In een traditionele omgeving zouden we dit transactierecord laten uitvallen; in de Data Vault daarentegen maken we eerst een nieuw HUB_KLANT record aan, zie afbeelding 6. Hierna kunnen de verkopen probleemloos geladen worden in LNK_VERKOOP – we zijn dus in staat om alle aangeleverde data te laden.

Omdat de klant nog niet bekend is in het master CRM-systeem, zijn er nog geen beschrijvende gegevens voorhanden, en kan nog geen *satellite* aangemaakt worden bij de nieuwe klant, zie afbeelding 7. Na enkele dagen is het CRM-systeem bijgewerkt, klant DNZ999 is overgenomen en aangevuld met naam, adres en andere beschrijvende gegevens. Er zijn nu in het EDW geen uitzonderingsroutines of updates vereist, de gegevens van de klant kunnen door het reguliere proces worden toegevoegd in SAT_KLANT, zie afbeelding 8.

Voorbeeld: technisch slechte datakwaliteit.

De Data Vault biedt ook een geïntegreerde oplossing voor gegevens die zo slecht van kwaliteit zijn dat ze voor de datawarehouse-omgeving niet hanteerbaar zijn. Denk aan te lange waarden, numbers die geen number zijn, afwijkende datumformaten etcetera. In de Data Vault worden deze waarden (met instemming van de gebruiker) omgezet naar default waarden: NULL voor numbers, een defaultwaarde (bijvoorbeeld 1-1-1780) voor datums. De oorspronkelijke waarde gaat echter niet verloren bij de omzetting, deze wordt ondergebracht in een speciale error-satellite, zie afbeelding 9.

De error-satellite bevat dezelfde velden als zijn reguliere 'broertje', maar dan met een character datatype; iedere mogelijke binnengekomen waarde kan er dus in worden opgeslagen. De probleemrecords uit de error-satellite kunnen doorgeladen worden naar een Error Mart, zodat de gebruiker ze kan bekijken, beoordelen, en, als hij dat wenselijk acht, (laten) corrigeren in het bronsysteem.

In de Data Vault is datakwaliteitsborging dus onderdeel geworden van de architectuur en het reguliere datawarehouse-proces.

Conclusie

De volgende generatie EDW heeft qua architectuur twee fundamentele kenmerken: er wordt verschil gemaakt op conceptueel, logisch en technisch niveau tussen feiten en waarheden; de mate van data-integratie wordt bepaald door de eindgebruiker en wordt zoveel als technisch mogelijk uitgevoerd op het moment dat er om gevraagd wordt.

Wat betekent dat voor de architectuur en het datawarehouse-proces? Een radicale omslag. Het centrale datawarehouse is niet langer een geschoonde, geïntegreerde gegevensverzameling; het dient zuiver als opslag voor de feiten. Integratie en schoning vinden pas na het datawarehouse plaats, als de datamarts gemaakt worden. Ingericht naar gebruikersbehoefte, geven zij de kijk van de gebruiker op de facts weer, en daarmee de versie van de waarheid van die gebruiker.

Waar organisaties in de eisen met betrekking tot datamanagement de laatste tien jaar zijn geëvolueerd (vooral wat betreft compliance en schaalbaarheid) is de datawarehouse-industrie in slaap gesukkeld en niet meegegroeid. Met bovenstaande architecturale principes kan deze zich weer aansluiten en kan weer voldaan worden aan de steeds hogere eisen die organisaties stellen aan datawarehouses.

Wie op zoek is naar methodologische ondersteuning, vindt in een methodiek als de Data Vault een uitstekend kader voor de volgende generatie EDW, die daarmee beter aan de verwachtingen kan voldoen en daarmee het negatieve sentiment rondom datawarehousing kan doorbreken.

Ronald Damhof (ronald.damhof@prudenza.nl) is Senior Information management Architect.

Met dank aan Lidwine van As, consultant bij Grey Matter.

Update

InterSystems presenteert embedded BI-oplossing DeepSee

Intersystems heeft bekend gemaakt te werken aan een BI-oplossing die geïntegreerd wordt in Caché en Ensemble en dus niet los verkrijgbaar zal zijn.

DeepSee is voorlopig alleen beschikbaar voor 20 proefprojecten.

Gaandeweg zullen de bestaande Caché-applicaties met de BI-oplossing worden verrijkt, waarvan vooral de vele zorg-applicaties (zoals Philips Medical, Siemens Medical, I-Soft, Epic) profite-

ren, maar wellicht ook de bibliotheeksoftware Vubis Smart van Infor.

QlikTech brengt QlikView 8.5 uit

De belangrijkste vernieuwingen in QlikView 8.5 zijn:

- Geavanceerde 'Set Analysis'-functies vergelijken gerelateerde datasets met elke willekeurige andere dataset – in één view, met een druk op de knop;
- Verbeterde integratiefuncties, zoals open API's voor koppeling met andere voorzieningen, en bulkimplemen-

taties van AJAX- en Java-clients maken het ook voor de grootste organisaties eenvoudiger om QlikView in te zetten en te beheren;

- Verbeterd gebruiksgemak, zoals de rechtstreekse integratie van QlikView in generieke bedrijfsapplicaties, maakt samenwerken makkelijker, terwijl aanvullende visualisaties QlikView duidelijker en plezieriger maken om mee te werken;
- Vereenvoudigde licentieregeling, geen onderscheid meer tussen 32- en 64-bit versies.