Referentie architectuur voor Metadata- & Masterdata-beheer

Semantic MediaWiki (1)

Henk Scholten

Het concept en de software rondom Semantic MediaWiki voor het beheer van Metadata- & Masterdata is het meest aandachttrekkende deel van de BI-Team Referentie Architectuur voor Business Intelligence/Performance Management (BI/PM). Lichtgewicht en direct productief kennisbeheer door de gebruikers wordt ingezet voor een solide Metadata- & Masterdata-beheer. Het Datawarehouse en de Data Marts worden op basis van die kennis gegenereerd.

Aangezien de Wiki-toepassing een onderdeel is van een complete visie op BI/PM kan het niet zonder die context besproken worden. In dit eerste artikel zal dan ook eerst op de gehele BI/PM referentie architectuur worden ingegaan. In het tweede artikel wordt het concept voor Metadata- & Masterdata-beheer gepresenteerd. Gevolgd door deel drie met een beschrijving van de concrete functionaliteit.

Referentie architectuur

De BI-Team Referentie Architectuur voor BI/PM, zie afbeelding 1, is gebaseerd op de volgende kenmerken van BI/PM (projecten):

1. Snelle realisatie is de belangrijkste succesfactor;
2. Behoeftenanalyse is de belangrijkste probleemfactor;
3. Verkrijgen van de goede data kost de meeste tijd en inspanning bij implementatie;
4. Gebruik is productiever, naarmate (super)gebruikers meer zelf kunnen doen.

Snelheid en wendbaarheid (agility) is waar het om gaat bij BI/PM. Het is onmogelijk om de behoeften op voorhand in detail vast te leggen. De toestand van organisaties verandert voortdurend, dus de behoefte aan informatie ook. Goede antwoorden roepen nieuwe vragen op, een compleet inzicht ontstaat niet in één keer. En als de realisatie van een informatiestroom lang duurt, zijn de behoeften al weer veranderd vóór het antwoord er is; het komt het nooit af. De behoeften moeten dus niet worden vastgelegd, maar moeten kunnen worden ontdekt en gevolgd, dat is een uitdaging die met deze architectuur wordt beantwoord. Uiteraard is inzicht in de behoeften wel van belang, maar dat kan vaak snel worden verkregen. In deze opzet wordt de inspanning van de gebruikers (en consultants/IT) niet gericht op behoeftenanalyse maar op kennisbeheer. Dat is een positieve en intrinsiek zinvolle activiteit in tegenstelling tot behoeftenanalyse. Daar komt bij dat de mate waarin de kennis over de data beschikbaar is veel belangrijker is voor de snelheid van een project dan een behoeftenanalyse ooit kan zijn.

De componenten

Deze architectuur is geformuleerd op basis van ervaring en onderzoek naar wat gebruikers willen en een visie op wat software moet bieden voor optimale ondersteuning van BI/PM-processen in organisaties. En ook al is de invulling van het Metadata- & Masterdata-beheer met een Semantic MediaWiki nieuw, de architectuur is gebaseerd op bewezen en beproefde technologieën. De nadruk op het gebruik van Excel en de koppeling met Statistiek, Document- en Kennisbeheer is weliswaar nog niet algemeen voorkomend in BI/PM-omgevingen, maar de technologieën en toepassingen zijn volwassen en zoals in het geval van Excel, juist zeer veel voorkomend.

> De data kunnen direct na laden vanuit elk mogelijk perspectief worden bekeken en getest

Ook de MediaWiki software is met toepassingen zoals onder andere Wikipedia en Intellipedia goed getest. Het feit dat het gebruik van deze software breed bekend is, is ook een belangrijk aspect. Deze architectuur kan goed met componenten van diverse leveranciers worden gerealiseerd, dat sluit aan bij de praktijk waarin toch vrijwel altijd heterogene omgevingen voorkomen. Zelfs analist Gartner ziet na jaren van aanbevelen van één uniforme omgeving dat de werkelijkheid niet snel zal veranderen. Maar er is natuurlijk niets op tegen om de ambitie te hebben het geheel ooit met één suite te realiseren. Deze architectuur biedt een patroon voor het samensmeden van typische BI/PM-onderdelen tot een complete omgeving.

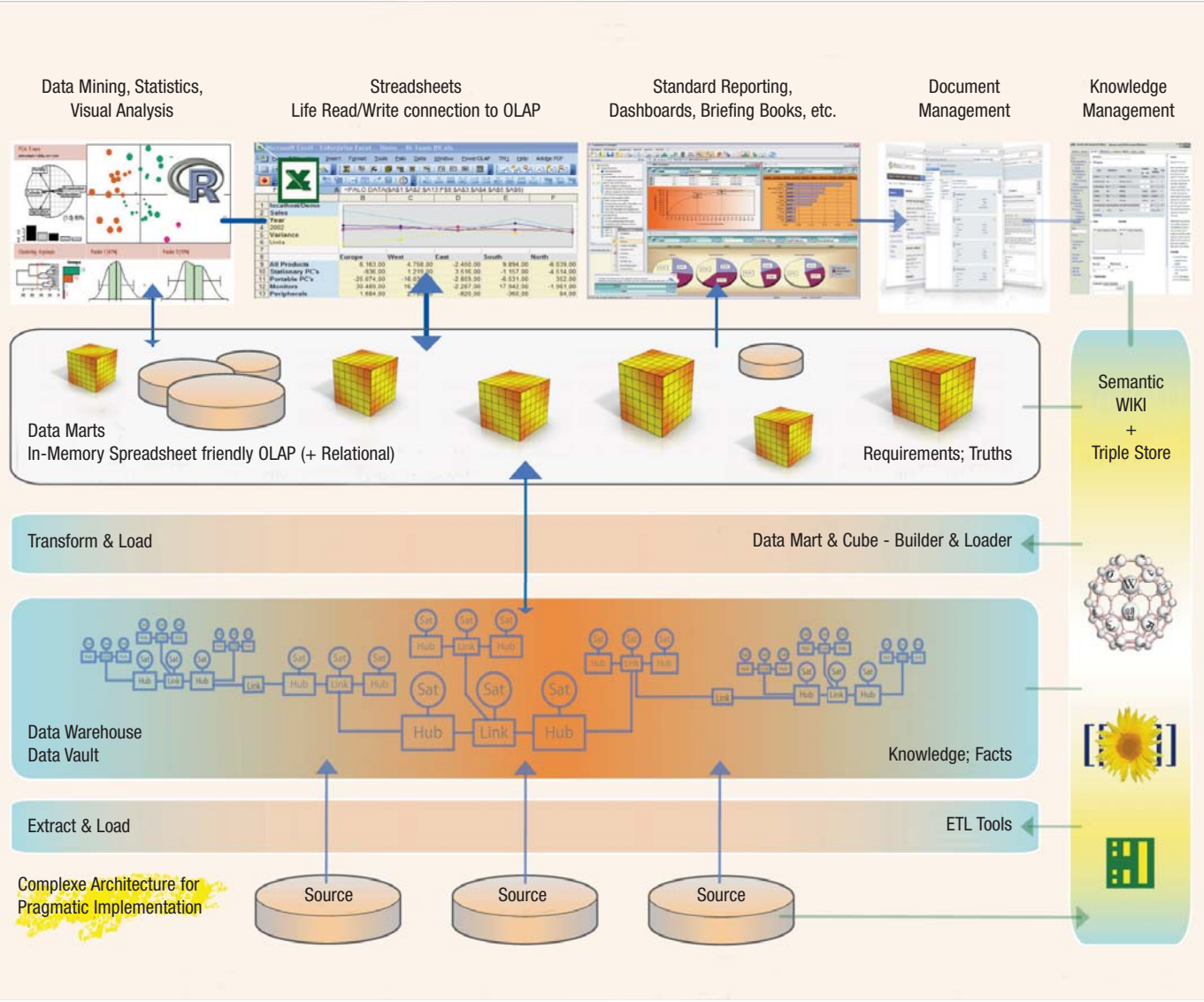
Integratie

De goede integratie van alle denkbare 'evidence based management' ondersteunende toepassingen is een wezenlijk aspect van

deze architectuur. Het gaat daarbij vooral om de integratie van het 'werken *met* de data' en 'werken *op basis van* de data'. De echte datawerkers binnen de organisatie, de planners, controllers en CFO's, willen *met* de data werken. De werkvloer en het top management werken vooral *op basis van* de data. De integratie daarvan is belangrijk. Bijvoorbeeld de integratie van Sales & Operations Planning met de rapportage, dezelfde data worden ten koste van veel inspanning ontgonnen en geëxploiteerd. Binnen deze architectuur bestaat ook aandacht voor wat er met rapporten gebeurt als die eenmaal zijn geproduceerd. Vooral in het kader van Compliance en Auditability wordt voorgesteld om een Document Management systeem in te zetten voor de distributie. Maar het heeft nog andere voordelen, zoals het afleveren van meerdere documentformaten vanuit één bron, het kunnen opzetten van een workflow voor accordatie en distributie, en het kunnen zoeken in alle documenten en rapporten. Traditioneel gezien bestaat Business Intelligence vooral uit het werken op basis van de data. Het werken met de data is in opkomst, wat zijn uitdrukking vindt in het groeiende gebruik van de term Performance Management.

Werken met de data

Een belangrijke rol speelt de inzet van Excel als client van een 'In-Memory Spreadsheet friendly OLAP' server; een specifieke technologiekeuze. Deze client/server-opzet maakt van Excel niet alleen de meest gebruikte BI/PM tool, maar ook de meest bruik-



Afbeelding 1: De BI-Team Referentie Architectuur voor BI/PM.

Deze architectuur voorziet in alle denkbare functionaliteit ten behoeve van Business Intelligence en Performance Management (BI/PM). Het biedt optimaal support voor activiteiten op het gebied van 'evidence based management' zoals Rapportage, Dashboards, Planning, Forecasting, Kwaliteitsbeheer, Capacity Based Costing, Scenario & Risico Analyse, Balanced Scorecard enzovoort. Deze referentie-architectuur bestaat uit: aanwijzingen; de keuze voor specifieke technieken; en best practices.

bare BI/PM tool. Voorbeelden van 'In-Memory spreadsheet friendly OLAP servers' zijn TM1 en Express Xcelerator van IBM Cognos en het 'commerciële open source' product PALO van Jedox. Cognos (IBM) is het belang van deze technologie gaan zien en heeft in 2007 het sinds 1984 bestaande Applix-TM1 gekocht. Daarmee is deze technologie opeens 'mainstream' geworden. Er bestaan nog diverse andere leveranciers met producten zoals PowerOLAP van Paris Technologies, PM10 van Infor, enzovoort. Het is dus geen single vendor verhaal.

Excel geeft iedere gebruiker die de toegangsrechten heeft direct toegang tot de data. Zonder dat er eerst een rapport of model moet worden gemaakt, kunnen de data direct na laden vanuit elk mogelijk perspectief worden bekeken en getest. Dat is een belangrijk voordeel van deze opzet. De eenvoudige toegang tot de Data Marts met Excel is één van de factoren die data modelleren *on the fly* mogelijk maakt. De inzet van client/server Excel is essentieel voor het kunnen ontdekken en volgen van de behoeften.

Met Excel en een 'In-Memory Spreadsheet friendly OLAP server' kunnen alle bekende BI/PM-activiteiten worden ondersteund, zie afbeelding 2. Iedereen kan de gratis PALO server plus Excel add-in downloaden en dit zonder kosten zelf toepassen. De integratie van Excel met software van het open source R-Project voor statistiek en datavisualisatie biedt in combinatie met een *in-memory* OLAP server als databron, ongekende mogelijkheden voor geïntegreerde numerieke en visuele analyse. Op de RExcel website van Statcon staat onder de naam RAndFriends een programma dat in één keer alle software installeert om met R en Excel te werken. Hiermee vormt Excel een statistische en visuele analyse-omgeving van de eerste orde. En dat voor nul euro investering boven de Excel licentie. PALO en R werken overigens ook met OpenOffice Calc, maar dan wel minder comfortabel. Er bestaat goede documentatie voor het gebruik van R, zie de referenties voor twee aanbevolen titels uit het grote aanbod. Voor het procesmatig statistisch analyseren staan Data Mining tools ter beschikking zoals bijvoorbeeld RapidMiner of Knime. Dit zijn uitstekende Data Mining-omgevingen die niets hoeven te kosten zolang er geen support wordt gevraagd. Uiteraard zijn producten als Tableau, SAS-JMP en IBM-SPSS ook goed inzetbaar en in zakelijke toepassingen zeker op zijn plaats.

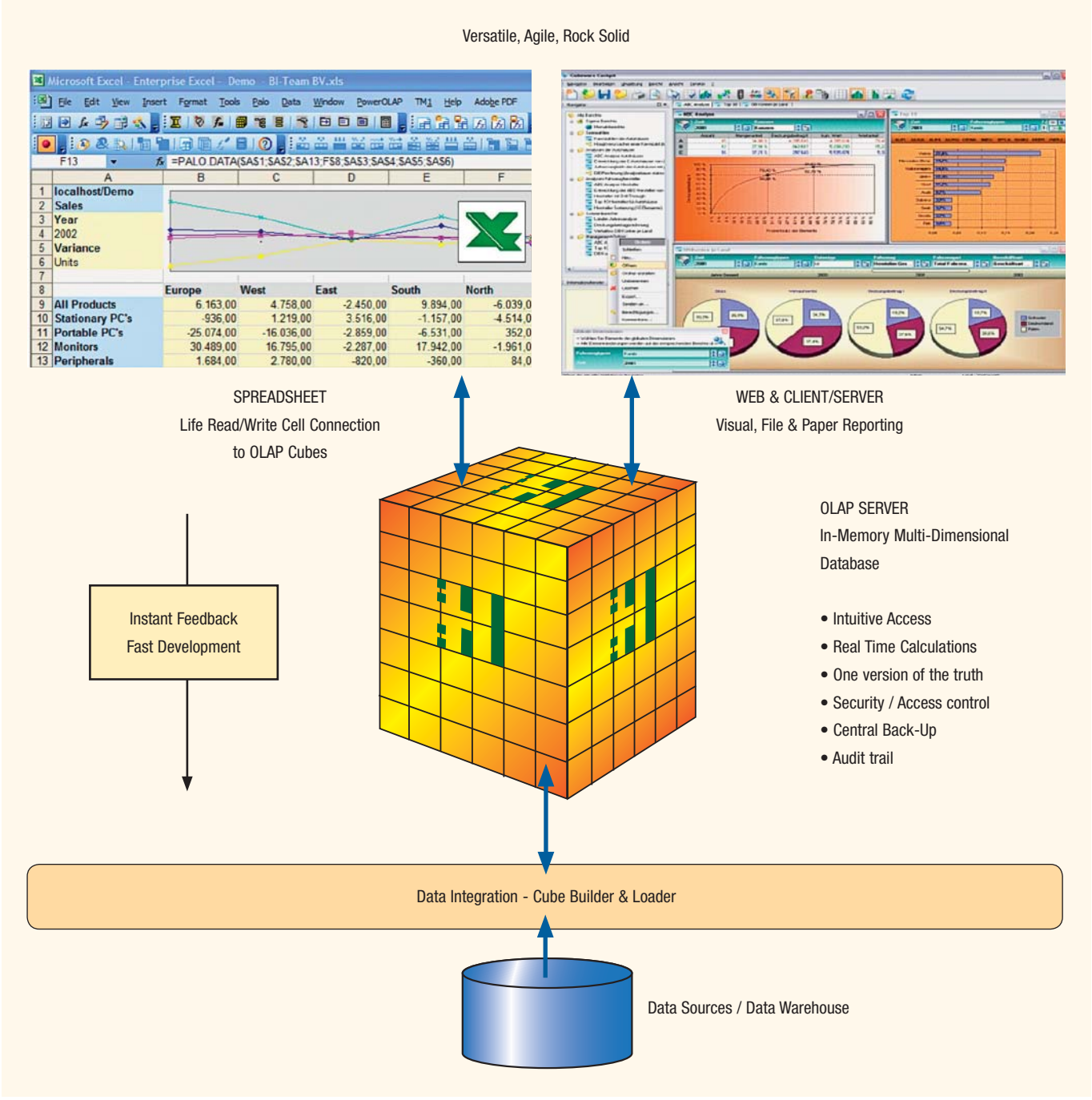
Werken op basis van de data

Op basis van een goede Data Mart kunnen met elk modern rapportage tool zonder enig probleem effectieve rapporten worden gemaakt. Rapportage wordt ten onrechte in de eerste plaats met visualisatie geassocieerd, terwijl geavanceerde visualisatie directer verbonden is met analyse dan met rapportage. Het verstandig gebruik van de grafische middelen speelt wel een grote rol bij rapportage. Vrijwel alle moderne tools bieden vergelijkbare en voldoende grafische mogelijkheden voor rapportage en dit is in feite een behoorlijk onbelangrijke keuzefactor voor een BI-omgeving. Toch is de mogelijkheid om bonte en wilde grafieken te kunnen maken, bij voorkeur *speedometers*, een populair keuzecriterium voor een BI/PM-omgeving. Beschikbaarheid van doordachte en vernieuwende grafische elementen is echter het reële keuzecriterium op dit vlak. In veel BI/PM-producten verschijnen echte vernieuwende weergavemanieren pas jaren nadat ze zijn uitgevonden of nooit, zoals *Sparklines* en *Bullet Graphs*. Nieuwe weergavemanieren zijn niet zelden eerder als add-in in Excel beschikbaar, dan in de meeste BI/PM software. In de BI/PM-praktijk is het vooral belangrijk hoe er van de grafische mogelijkheden gebruik wordt gemaakt. Vrijwel overal waar met rapportage wordt gewerkt kan meer worden bijgedragen aan de kwaliteit van rapporten door iets te leren over visualisatie uit de boeken en websites van Stephen Few en Edward Tufte, dan toolkeuze tot gevolg zal hebben.

> De kwaliteit van rapportage blijkt een goede indicator te zijn voor de levensvatbaarheid van bedrijven

De mogelijkheden op het gebied van distributie en de moeite die het kost om diverse verschijningsvormen (PDF, web, enzovoort) vanuit één bron te realiseren, en vooral de mate waarin (super) gebruikers zelf rapporten kunnen maken, zijn wezenlijker bij de opzet van een rapportage-omgeving dan de grafieken die gemaakt kunnen worden. Uiteraard is de beschikbaarheid van specifieke vereiste weergavemogelijkheden, zoals geografisch, wel belangrijk. In een lokaal netwerk vormt een *in-memory OLAP server* met Excel als client, een prima rapportage-omgeving voor MKB en afdelingen van grotere organisaties. Voor bredere toegang kan de informatie in OLAP-kubussen eenvoudig en snel op het web worden gezet met tools van diverse leveranciers. In een procesmatig werkende rapportage-omgeving met tastbare producten zoals PDF-bestanden, zijn de meeste traditionele BI-tools het best op hun plaats. Bij echte gecontroleerde distributie komen Document Management systemen in zicht. BI/PM-omgevingen bieden niet wat Document Management systemen bieden, daarom zijn dat ook aparte systemen. Vooral als *Compliance* en *Auditability* aan de orde zijn, in grotere organisaties dus, is serieus Document Management een onmisbaar deel van de BI/PM-omgeving.

Correcte, volledige en tijdige rapportage is van levensbelang voor een organisatie en vormt de basis voor alle andere BI/PM-activiteiten. Het blijkt dat de kwaliteit van de rapportage een goede indicator is voor de levensvatbaarheid van bedrijven. De winnaars van de in 1954 uitgereikte eerste FD Henri Sijthoff-prijzen voor de beste jaarverslaggeving bestaan nog allemaal. Zonder achteruitkijkspiegel gaat ook niemand de weg op. Maar boeiender is natuurlijk wat er door de voorruit te zien is, wat de weg voor ons brengt en hoe we daar op moeten reageren. Kort



Afbeelding 2: Spreadsheet Friendly OLAP server.

door de bocht gezegd: met rapportage wordt geen geld verdiend, met analyse, forecasting en planning wel. De integratie van de informatie uit 'de achteruitkijkspiegel' en 'de voorruit', is daarbij van groot belang voor de productiviteit, kwaliteit en coherentie van de informatievoorziening.

Data-opslag voor BI/PM

De keuze voor de Data Vault opzet als Datawarehouse concept is ook essentieel voor deze architectuur en wel om de volgende redenen:

1. De strikte opslag van feiten en het uitsluiten van interpretaties in het Datawarehouse. Binnen het Data Vault concept worden data vanaf bron naar Datawarehouse uitsluitend op volledig omkeerbare wijze bewerkt. De bron kan volledig worden herleid. Pas als er een Data Mart van wordt gemaakt, worden de data geïnterpreteerd en voor consumptie geschikt gemaakt. Dat is van wezenlijk belang voor het voortdurend kunnen veranderen en groeien van de behoeften en inzet van informatie;
2. De flexibiliteit van de structuur maakt dat een Data Vault naar wens kan worden uitgebreid en gesplitst, zowel in aandachtsgebied, *scope*, als in detail. Data uit verschillende bronnen kunnen uitstekend worden geïntegreerd. En met het ontslui-

ten van meer bronnen kan eenvoudig meer detail worden toegevoegd en/of een uitbreiding van aandachtsgebied worden gerealiseerd. Het is van groot belang dat een Data Vault incrementeel kan worden gerealiseerd. Daarmee past het binnen de eerder geformuleerde uitgangspunten van het kunnen ontdekken en volgen van de behoeften en het snel kunnen realiseren van informatiestromen;

3. Het bouwen en laden van de Data Vault en de Data Marts kan goed worden geautomatiseerd. Dat is ook van groot belang in het licht van de genoemde uitgangspunten.

Het eBook van Informatica met de titel 'The Math, Myth & Magic: An Introduction to Identity Data Search-and-Match' gaat niet over Datawarehouse-architectuur, maar levert wel alle argumenten voor de opslag van feiten. Dit boek maakt duidelijk wat de problemen zijn met opslag van interpretaties. Het lijkt wel een aanbevelingsbrief voor het Data Vault concept en is interessant voor iedereen die met data werkt.

Resources komen zolang de wet van Moore stand houdt steeds weer beschikbaar

Aangezien aan een voordeel ook een nadeel kleeft, is het goed om te zien welke prijs betaald wordt. Deze bestaat uit de inzet van meer resources, de opslagruimte neemt toe en query's vragen meer inspanning van de hardware. Toch is dat geen probleem, want de historie van computing bestaat uit steeds betere bruikbaarheid en het op steeds hoger abstractieniveau werken op basis van de inzet van meer resources. En die resources komen zolang de wet van Moore stand houdt ook steeds weer beschikbaar. De Data Vault is geen besparingsmethode, het ligt niet op de weg terug maar op de weg voorwaarts.
De Data Vault is vandaag al productief en economisch inzetbaar, het is geen toekomstmuziek. Belangrijk is ook dat de IT-afdeling behoorlijk zelfstandig een Data Vault kan bouwen, in tegenstelling tot klassieke Datawarehouses worden er immers geen interpretaties opgeslagen en wordt de structuur niet door de behoeften van de gebruikers bepaald. Enige kennis vanuit de organisatie is daarbij uiteraard wel noodzakelijk, de 'business-keys' moeten wel bekend zijn, maar *requirements* komen niet aan de orde. Het gaat er dan om de informatie in de organisatie, database voor database, met een geringe inspanning te ontsluiten. IT kan de organisatie er *sooner or later* blij mee maken.
De gebruikers kunnen dan hun Data Marts direct vullen met perspectieven op de organisatie die tot dan toe vergeten waren, en mogelijk zonder zo'n initiatief niet meer beschikbaar zouden zijn geweest. Ook is het van belang om data, en dan vooral de informatie over de data, centraal te vangen als de personen die weten wat het betekent er nog zijn en het nog weten.
De Data Marts bestaan in deze opzet vooral uit *in-memory* OLAP-kubussen. Maar ook Exploration Marts, dat wil zeggen platte (CSV) bestanden of platte tabellen in een relationeel DBMS, horen bij deze standaard architectuur. Exploration Marts zijn goede bronnen voor Data Mining en Rapportage. De Data Vault vraagt veel computer-inzet bij het bevragen, omdat er dan veel data bij elkaar worden gezocht. Eén van de functies van de Data Marts is het opvangen en beantwoorden van de vragen die het systeem anders te traag zouden maken.

Voor wat betreft de vraag of het DW echt tastbaar of virtueel moet zijn; dat maakt voor de architectuur niet veel uit. Maar kantoor *malheur* maakt het leven niet fijner en een centrale fysiek aanwezige database maakt duidelijk waar de data staan en wie verantwoordelijk is. Ook worden niet plotseling 'onbelangrijke' velden of tabellen veranderd, back-ups lopen centraal, de historie is geborgd. Belangrijk is ook dat alle Data Marts met één technologie op hun bron kunnen worden aangesloten. Dat geeft *super-users* toegang, personen die direct in de primaire processen van de organisatie staan en de BI/PM-omgeving meebeheren en daar zelf dingen mee doen.

Referenties

1. *Informatica eBook: The Math, Myth & Magic, http://vip.informatica.com/?elqPURLPage=553.*
2. *Richard M. Heiberger & Erich Neuwirth, R through Excel, A Spreadsheet Interface for Statistics, Data Analysis, and Graphics, Springer 2009, ISBN 978-1-4419-0051-7.*
3. *Dianne Cook & Deborrah F. Swayne, Interactive and Dynamic Graphics for Data Analysis with R and GGobi, Springer2007, ISBN 978-1-387-71761-6.*
4. *Henk Scholten, BI-Team whitepaper, 2009, BI-Team Referentie Architectuur, Een compleet raamwerk voor Performance & Knowledge Management.*

Websites

5. *Stephen Few, Perceptual Edge: www.perceptualedge.com*
6. *Stephen Few – Graph Design I.Q. Test: www.perceptualedge.com/files/GraphDesignIQ.html*
7. *Edward Tufte: www.edwardtufte.com*
8. *Express Xcelerator, IBM/Cognos: www.ibm.com/software/data/cognos/products/cognos-express/xcelerator*
9. *PALO, Jedox: www.jedox.com/en/home/overview.html*
10. *PowerOLAP, Paris Technologies: www.paristech.com*
11. *RExcel: http://rcom.univie.ac.at*
12. *RExcel introductievideo: http://rcom.univie.ac.at/RExcelDemo*
13. *GGobi: www.ggobi.org*
14. *GGobi demo videos: www.ggobi.org/demos/ Videos: teaching-with-ggobi.html, brushing-simple.html, brushing-biotin.html, tour.html*
15. *RapidMiner: http://rapid-i.com*
16. *Knime, Konstanz Information Miner: www.knime.org*
17. *Sparkmaker: www.bissantz.de/palo/index_de.asp*

Henk Scholten (hscholten@bi-team.com) is Managing Director van BI-Team BV.