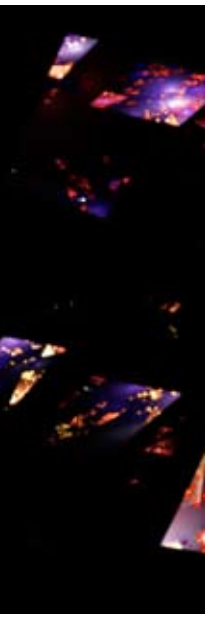


Vier experts brainstormen over het datawarehouse

Data niet kopiëren maar organiseren

Robert de Ruiter

Te gast bij het Centraal Boekhuis praten we met vier deskundigen over het datawarehouse. De conversatie springt van de hak op de tak en er passeren veel onderwerpen. Een voorproefje: de best practices van vandaag zijn gebaseerd op de beperkingen van gisteren, dus genoeg over het verleden.

De inrichting van het datawarehouse staat de laatste jaren flink ter discussie. Kimball heeft heel lang het performance-argument gebruikt om het datawarehouse te rechtvaardigen. Vanwege de ontwikkelingen in dataopslaghardware en flashgeheugen is performance echter een afnemend probleem. Vaak wordt ook het historie-argument gebruikt (transactionele systemen houden geen historie vast). Met de groei van hardware- en software-capaciteit is dit misschien nu nog wel waar, maar het is geen conceptueel argument. Op basis van deze argumenten is het datawarehouse een lapmiddel, zo concluderen de vier experts in een brainstormsessie.
De vraag is of er echte conceptuele redenen zijn om met een datawarehouse te werken. En ja, die zijn er:

- De toenemende heterogeniteit van informatiebronnen. Binnen bedrijven en met name tussen bedrijven vindt gegevensintegratie en datauitwisseling plaats. Vooral de toenemende gegevensstroom tussen bedrijven moet georganiseerd worden;
- Je moet onafhankelijk van de werkelijkheid kunnen modelleren, om te kunnen forecasten, nieuwe strategieën te testen, what-if analyses te kunnen doen (scenario, simulatie).

Aan deze brainstormsessie deden mee: Emiel van Bockel, gastheer en manager information services & Business Intelligence architect bij Centraal Boekhuis; Frank Buytendijk, vice president en fellow voor enterprise performance management bij Oracle Corporation; Rick van der Lans, onafhankelijk adviseur, docent, auteur en spreker over datawarehousing, BI, applicatie-integratie en databasetechnologie; Harm van der Lek, managing consultant bij Van der Lek Advies BV.

Emiel van Bockel: "Het datawarehouse zou moeten worden omgebouwd tot een omgeving waar je wel goed kunt modelleren zonder afhankelijk te zijn van pakketintegraties, consolidaties, proceswijzigingen en andere technische keuzes die leiden tot dataopslagdefragmentatie en redundantie."

"Sommigen vinden het kopiëren van data ten behoeve van BI (wat je doet als je een datawarehouse maakt) een zwaktebod, een lapmiddel, dat je soms wel moet gebruiken, omdat je in wezen het datamanagement binnen je organisatie niet op orde hebt. Is dat laatste wel het geval, dan heb je immers het ideaal bereikt, dat alle informatie maar op één plek redundantievrij is opgeslagen en netjes gemodelleerd. Het integratie argument voor een datawarehouse vervalt hierdoor. Als je dan in deze *operationele* systemen ook nog netjes bij alle wijzigingen de oude waarden bewaart, dan vervalt ook het laatste (belangrijke) argument voor een datawarehouse – het opbouwen van historie van data", zegt Harm van der Lek.

Beste keuze

Tegenover dit 'lapmiddel' standpunt wil Van der Lek een veel principiëler standpunt innemen, namelijk dat de vereisten voor een BI-systeem in de meeste (dus niet alle) gevallen zo immens verschillen van de vereisten aan de operationele systemen, dat je niet eens moet willen om ze in één systeem te vangen: "Ik heb het dan overigens over veel meer dan query performance alleen. Met andere woorden: hoeveel problemen het bouwen en onderhouden van een datawarehouse ook met zich meebrengt, het is toch vaak principieel de beste architectuurkeuze, omdat je nu beide omgevingen (operationeel en BI) los van elkaar voor de eigen vereisten kunt optimaliseren."

Als je al tot de conclusie komt dat er een datawarehouse moet komen vanwege integratie van systemen en historie komt de vraag boven: wat moet je dan kiezen? En als je het al hebt over een enterprise centraal datawarehouse en datamarts, hoe bouw

Foto: Arjen van den Berg.

Van links naar rechts Rick van der Lans, Harm van der Lek, gastheer Emiel van Bockel en Frank Buytendijk: "De kracht van het DDP zit in de technologische flexibiliteit, niet in de menselijke flexibiliteit".

je dan het centrale datawarehouse? Kimball ziet dat als sterschema's, Inmon in feite hetzelfde in andere terminologie. Ze praten beiden over een boven- en onderlaag. In de bovenlaag (dicht bij de gebruiker) wordt gedenormaliseerd en in de onderlaag moet je het netjes houden.

"Maar eigenlijk is het een ondergeschikt probleem. Het is niet echt belangrijk. Veel moeilijker is het om te achterhalen wat de gebruiker wil, hoe je het beslissingsproces kunt ondersteunen, hoe je de data goed krijgt zodat je deze kunt vertrouwen. Of je dit doet met een sterschema of een sneeuwvlokschema doet helemaal niet ter zake", stelt Rick Van der Lans vast.

> Je hoeft alleen maar te regelen wie welke klanten kan zien

Frank Buytendijk onderstreept dat. "Stel, je bent projectleider van de implementatie van een ERP-systeem. Zou je daaraan durven beginnen zonder dat je kennis hebt van het proces dat je wilt ondersteunen? Het juiste antwoord moet dan zijn: nee. Als je het vakgebied niet begrijpt ga je zeker nat. Nu gaan we naar het vakgebied BI waar het heel normaal is om niets te begrijpen van het besluitvormingsproces en de structuren. Daar gaan we wat rapportjes en dashboards bouwen op basis waarvan mensen 'betere' beslissingen gaan nemen. Met deze business case komen we al 25 jaar weg."

Verdiepen in gebruiker

De vraag is ook of de IT-afdeling deze vragen moet oplossen, want eigenlijk is deze daar niet voor opgeleid. "De informaticus is toch degene die de informatie in kaart moet brengen", vraagt Van Bockel zich af. "Net zoals een bedrijfskundige de processen. De meest foute vraag die je aan de business kunt stellen is 'wat wil je weten?' Deze vraag leidt gegarandeerd tot falen van het project. Veel beter is te achterhalen wat iemand doet. Welke handelingen voert diegene uit? Je moet je dus verdiepen in de gebruiker en zijn processen."

Dat dit niet eenvoudig is blijkt onder meer uit de 'requirements-industrie', die in de IT-branche is ontstaan. Immers, hoe bepaal je als IT'er wat de business wil als je het de business niet kunt vragen, omdat die het niet zouden 'begrijpen'. Buytendijk: "Dat heeft te maken met het collectieve minderwaardigheidscomplex van IT'ers. Die willen allemaal 'van de business' zijn. Kijk maar naar de terminologie; Business Performance Management, Business Intelligence, Business Process Management. Sommigen

noemen zichzelf zelfs business architect, en spreken met een air alsof ze de enigen zijn die 'de business' echt begrijpen. Ik heb moeite met dat soort mensen. Waanzin."

Van Bockel: "Op dit moment is de situatie moeilijk in het modelleren van informatie. We hebben het als informatici voor elkaar gekregen om de vragen te kunnen stellen aan en ons te bemoeien met managementprocessen, waar een bedrijfskundige meer van toepassing zou zijn. Waar de informaticus het vaak nalaat om goed met informatie om te gaan, laten ook procesmensen na om goed met processen en activiteiten om te gaan. Vraag zo iemand maar eens wat het proces is en vraag het zijn buurman. Dan heb je een hele avond discussie. Je kunt dus zeggen dat zowel de informatiewereld als proceswereld niet vakkundig met de professie omgaan, laat staan dat ze op elkaar zijn afgestemd." Tot zover over hoe het de afgelopen 20 jaar was.

Cloud

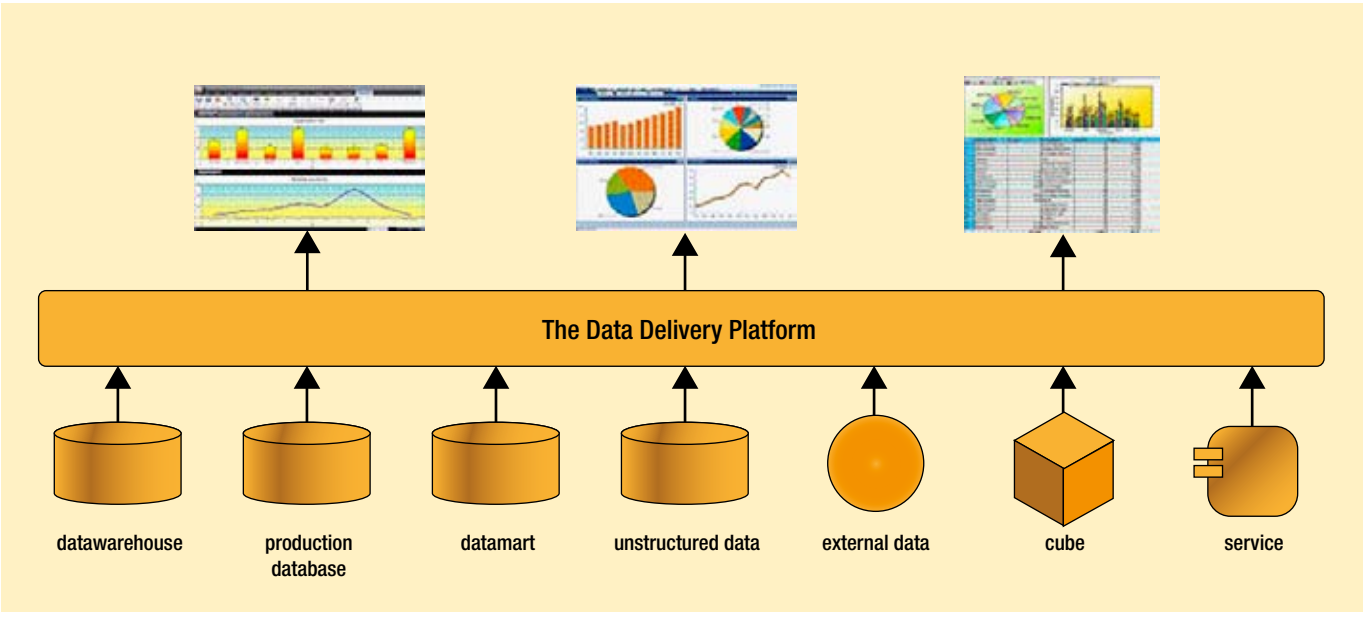
Van Bockel vraagt zich af waarom hij nog zelf zijn klantdefinities moet beheren en zich niet voornamelijk met de transacties kan bezighouden. "Waarom levert Oracle mij niet de structuur voor de klant?" Van der Lans doet daar nog een schepje bovenop; "Je kunt je voorstellen dat er in de Cloud een situatie komt waar de klanten van bijvoorbeeld Shell en Albert Heijn in dezelfde database zitten. Waarschijnlijk zijn dit voor 80 procent dezelfde klanten. Je hoeft alleen maar te regelen wie welke klanten kan zien."

"Dat gaat nooit lukken omdat mensen – al zullen ze het nooit toegeven – dan geen mogelijkheid meer hebben om dingen te manipuleren, tot en met fraude plegen aan toe. Dat zit gewoon onderin de pyramide van Maslov: sex, fear en greed. Als je die emoties ontkent krijg je dit heus niet ingevoerd", zegt Buytendijk. "Tenzij wellicht je de mensen hun eigen gegevens laat beheren. Iedereen heeft een burgerservicenummer en je laat mensen de mogelijkheid om daar verschillende profielen aan te hangen. Zodat ze zich bijvoorbeeld in het ene traject bekend kunnen maken als 'Piet van Putten' en als 'Kinky99' in het andere."

Van der Lans ziet het niet zo somber in. Hij vraagt zich af of, indien je een (aantal) centrale databanken verplicht stelt, dit wellicht eenvoudiger wordt geaccepteerd dan één database in de business. "Want dan is het nu eenmaal zo. Dan valt er niet meer over te discussiëren." Toch zoekt ook hij geen oplossingen voor de alom rondwarende data in een centraal geleide database. Blijft dus de vraag hoe je het probleem op een praktische en voor de business hanteerbare manier kunt oplossen. Van Bockel stelt dat hij met ERP-systemen zowel processen als informatiedefinities inkocht. "En ik wil procesdefinities bij de ERP-leverancier kopen en datadefinities bij de database leveranciers. Ik hoef niet per se een centraal systeem te hebben als het maar met elkaar communiceert. En hoe interesseert me niet."

Federation

Van der Lans ziet een softwarelaag tussen de data – ongeacht waar die vandaan komen – en de BI-omgeving als de oplossing. "Het maakt voor de business toch niets uit waar de data vandaan komen of hoe ze die kunnen benaderen. Dat is allemaal irrelevant. Het gaat erom dat ze de juiste informatie krijgt en dat die aan al hun eisen voldoet. Die softwarelaag noem ik het Data Delivery Platform. We kunnen alle data die zij nodig hebben achter het DDP plaatsen en we zorgen dat het DDP alle data vertaalt en filtert. Een DDP wordt tegenwoordig waarschijnlijk gebouwd met 'federation servers'. Het doet er dan niet meer toe wat je achter het DDP plaatst. Dat kunnen databases, datawarehouses, XML bestanden, HTML, datafeeds of gegevens van het internet zijn. Je moet als business wel even door een pijngrens,



Afbeelding 1: Data Delivery Platform.

omdat je die laag moet bouwen, maar daarna krijg je een enorme flexibiliteit."
Ook is de hoeveelheid opgeslagen data en de verwerkingssnelheid heel goed te managen. Op het moment waarop achter het DDP te weinig capaciteit dreigt te ontstaan kun je de databronnen over meer database servers verdelen. Je kunt ook veel eenvoudiger switchen van de ene naar de andere databasetechnologie, en je kunt nieuwe technologie gemakkelijker adopteren.
Van der Lans vergelijkt de laag software tussen data en de rapportagetools met het luikje in een Chinees restaurant. De bezoekers krijgen datgene voorgeschoteld wat zij hebben besteld. Het zal ze daarbij weinig interesseren hoe 'nummer 83' tot stand is gekomen, welke ingrediënten voor het gerecht zijn gebruikt of waar die zijn ingekocht.

Options based strategy

De andere deelnemers aan de brainstormsessie hebben veel detailvragen bij het voorstel van Van der Lans, maar zien uiteindelijk ook het nut van deze aanpak. Buytendijk: "Dit is een options based strategy. Je investeert ermee in de toekomst, omdat het er niet toe doet wat je in de toekomst achter het luikje zou willen zetten." Van Bockel: "Je moet in die laag dan ook de business rules of process rules vastleggen. Daar worden de definities bepaald en je kunt vóór de muur alle rapporten queryen die je nodig hebt."

> De grote kracht van het DDP schuilt in de terugverdientijd

Van der Lans: "Je moet natuurlijk nog wel je data modelleren. De tools die het DDP raadplegen zijn die van BusinessObjects, Cognos of wat dan ook. Die tools willen gewoon tabelstructuren zien, maar weten niet welke database ze precies benaderen. Ze weten dus niet of dat een datamart is, of relationeel of wat dan ook. Je kunt overigens nog een stap verder gaan door je productiesysteem aan het DDP koppelen. Als je dan toch een rapport wilt maken waarin je de oude data naast de nieuwe data wilt leggen, dan betekent dat voor de rapportbouwers niet veel. Wij moeten alleen die productiedatabase koppelen aan het DDP."
Voor alle duidelijkheid: de architectuur van het DDP tussen data en BI is leverancieroverstijgend, maar alle benodigde tools zijn ook nu al gewoon te koop.

Van Bockel: "Waar gemakkelijk overheen wordt gestapt is dat er bij de federation server van wordt uitgegaan dat aan beide zijden de unieke identificaties goed zijn. Als dat zo is en het systeem goed is ingericht, heb je ook eigenlijk geen federation meer nodig. En juist die unieke identificatie levert immense problemen op." Van der Lans: "Dat probleem blijf je houden. Als het voor een rapport nodig is om data uit verschillende systemen bij elkaar te brengen en het koppelpunt zit niet goed in elkaar, dan heb je met elke technologie een probleem. Je hebt een rommel gemaakt en daar ga je voor betalen. Dus vakmanschap blijf je nodig hebben. Maar als het in de federation server eenmaal is opgelost kan elke tool er gebruik van maken. Terwijl je, wanneer je het probleem in BusinessObjects oplost, weer aparte oplossingen moet ontwikkelen voor spreadsheets die ook die data willen combineren."

Stap in de goede richting

Buytendijk: "Ik vind een DDP volstrekt helder en logisch. Iedere architect moet ermee uit de voeten kunnen. Het is vergelijkbaar met SOA met een Enterprise Service Bus." Van der Lans: "Voor een gebruiker is relevant of zijn rapporten snel genoeg komen, voldoende kwaliteit hebben en compleet zijn. Met deze structuur kun je dat bereiken met de goedkoopste oplossingen. En eigenlijk is het helemaal niet zo bijzonder, maar veel mensen die in de BI/datawarehousewereld rondlopen zijn nog niet bekend met deze producten. Ze lezen elk artikel over snowflakes, stars en ontwerpen, maar hebben geen idee van welke nieuwe ontwikkelingen op de markt verschijnen. Ik sprak laatst een datawarehousespecialist van een groot Nederlands overheidsorgaan die niet wist wat een datawarehouse appliance was. Dat is vreemd."

Hoe je een en ander met de gebruiker communiceert is een onderwerp van discussie. Wat kun je wel en wat kun je niet met de gebruiker communiceren? Mag je hem de modellen laten zien, of de sterschema's? Volgens de opleidingen mag dat geen van beide, omdat de gebruiker daar niet mee overweg kan. Van der Lans: "De gebruiker wil weten of hij de gegevens kan extrapoleren en hij wil een scherm zien. Wellicht kun je nog over processen praten, maar verder moet je niet gaan."
Met sommige appliances is dat ook niet nodig. Ze zijn gemakkelijk in gebruik, je hoeft ze niet te tunen, je hebt bij wijze van spreken niet eens een gebruiksaanwijzing nodig. Met andere appliances moet je daarentegen wel tunen en aan allerlei palletjes gaan draaien om de machine te laten doen wat je wilt. Het is overigens ook nog afhankelijk van de gebruiker. Een bepaalde appliance kan voor een deel van de klanten correct zijn ingericht en getuned, terwijl een ander deel er nog veel aan zou willen aanpassen om bijvoorbeeld de performance te vergroten.

De conclusie van een middag discussiëren over nieuwe ideeën voor datawarehouse en BI is volgens de bijna eensgezinde brainstormdeelnemers: het oude doen we omdat we het zo doen, maar de echte redenen zijn achterhaald. Datamarts en snowflakes zijn achtergrondgerommel. Het wordt dus tijd voor nieuwe oplossingen. En het Data Delivery Platform is – zonder te weten wat de toekomst brengt – een stap in de goede richting. De grote kracht van het DDP schuilt in de terugverdientijd, de vele opties en mogelijkheden en de flexibiliteit naar de toekomst toe.

Robert de Ruiter is hoofdredacteur van Optimize.