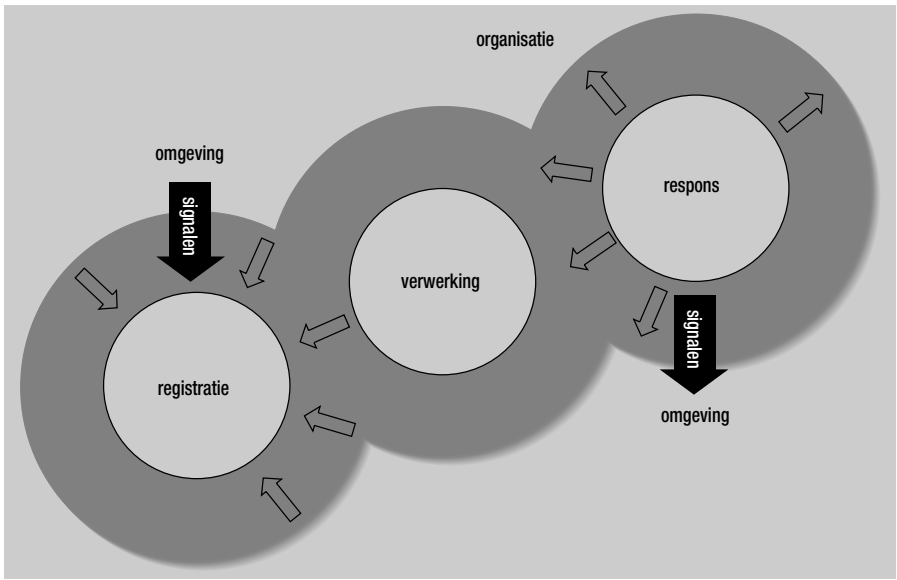
Het verwerkingsproces: de tijd zal het leren

Het tweestromenland begrepen

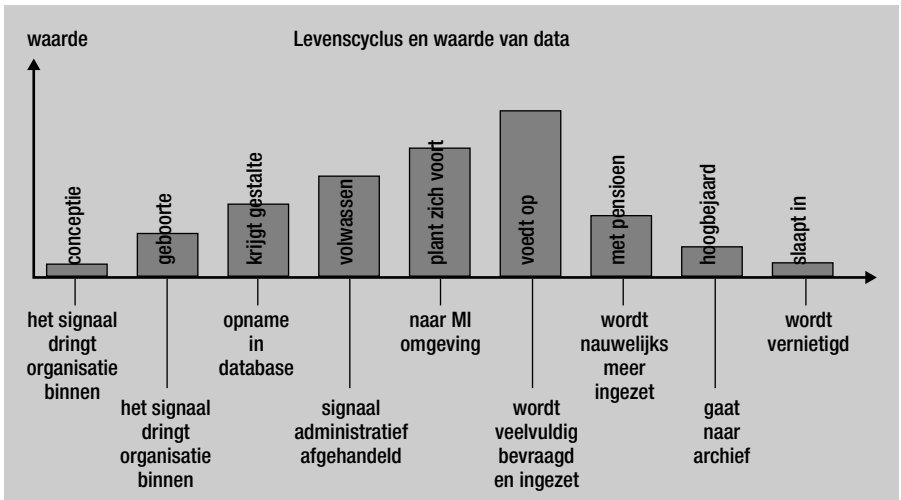
Daan van Beek

Verwerken kost tijd. Dat geldt niet alleen voor mensen die een belangrijke gebeurtenis een plek moeten geven. Ook informatiesystemen verwerken gebeurtenissen. Zondere tijdige en snelle verwerking loopt het vol en uiteindelijk vast. Het verwerkingsproces, het tweede proces in de reeks van drie, geeft de als transactie neergeslagen gebeurtenissen een plek waar ze goed toegankelijk opgeborgen liggen om uiteindelijk een langzame dood te sterven. De levenscyclus van een data-element start in het registratieproces en gaat zijn weg via het verwerkingsproces naar het responsproces zoals weergegeven in figuur 1. Het verwerkingsproces is een belangrijke schakel in het geheel: het verbindt de registratie met de respons en hevelt de signalen van het korte termijn geheugen over naar het lange termijn geheugen. Het geeft data kans om zich te ontplooien en te renderen. Echter, slapende gegevens, zowel in productie- als in managementinformatie-omgevingen benodigen veel kostbare capaciteit van de infrastructuur, de applicaties en de organisatie zelf. Selectief opslaan en verwerken is daarom van groot belang hoewel ieder gegeven dat een organisatie opslaat iets kan bijdragen aan de winst. De hoogte van het rendement op data verhoudt zich tot de pensioengerechtigde leeftijd ervan zoals figuur 2 weergeeft. Daarna neemt de waarde sterk af.

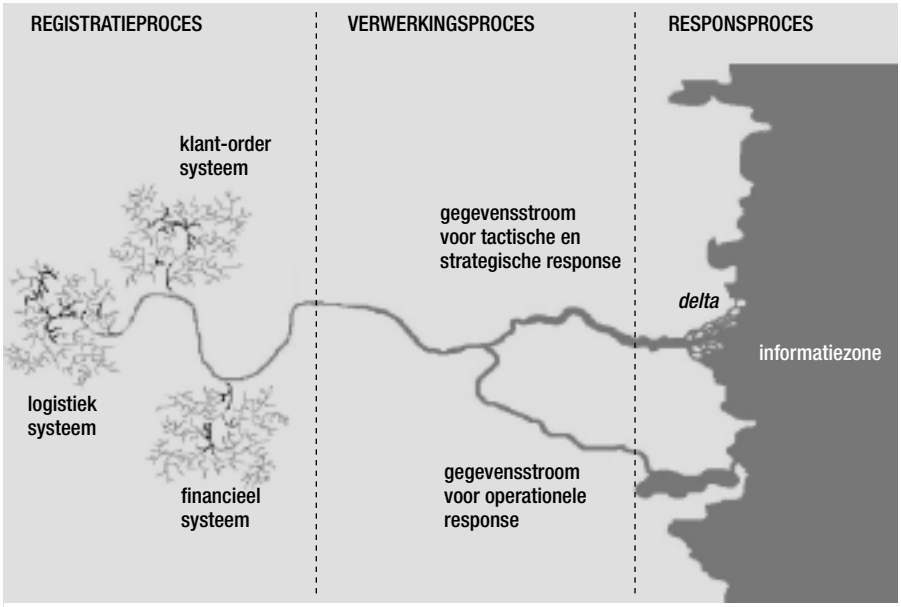


Figuur 1: de generieke basisprocessen in een organisatie

Datamanagement in het verwerkingsproces ondersteunt enerzijds het responsproces, door de gegevens toegankelijker op te slaan en zorgt anderzijds dat het registratieproces soepel blijft draaien, door het verplaatsen en verwerken van de geregistreerde gegevens. Immers, analoog aan het psychologische principe bij mensen,



Figuur 2: de levenscyclus en de waarde van gegevens



FIGUUR 3: HET TWEESTROMENLAND EN DE INFORMATIEZEE

het niet (goed) verwerken van gebeurtenissen leidt tot overspannen situaties waarbij productiedatabases steeds meer haperingen gaan vertonen. Uiteindelijk zal de organisatie niet meer in staat blijken om signalen goed te registreren en zal zelfs mogelijk ten onder gaan. In het verwerkingsproces ontstaan twee datastromen die ieder uiteindelijk een eigen weg gaan en verschillende doelen dienen. De metafoor van het tweestromenland, zoals weergegeven in figuur 3, draagt bij aan een beter inzicht in wat zich veelal 's nachts afspeelt in het rekencentrum van de organisatie. Daarnaast plaatst het het verwerkingsproces in de context van de andere generieke basisprocessen van een organisatie. Zie kader hieronder.

De doelstellingen binnen het verwerkingsproces geven richting aan de werkprocessen van datamanagement, zoals monitoring, foutcontrole, opschonen, back-up en recovery. De datamanagement-aspecten zoals integriteit, beveiliging, actualiteit en redundantie drukken een specifieke stempel op dit generieke basisproces. Immers, deze aspecten verschillen qua invulling binnen ieder basisproces.

Het twee-stromenland verklaard

De bronnen en beekjes duiden op de signalen die een organisatie binnenkomen en worden opgenomen in de hoofdrivier; het registratieproces dat gebeurtenissen als transacties opslaat in de database. De hoofdrivier en de splitsing tonen het verwerkingsproces; de grootste stroom gaat richting de informatiezee en die beeldt de beleidsondersteunende managementinformatie-omgeving af; de kleinere stroom geeft de operationele verwerking weer en mondt uit in het meer. De informatiezee weerspiegelt het responsproces dat de organisatie ondersteunt bij het nemen van beslissingen. In de rivierdelta, met een voedselrijke flora en fauna, is het goed fourageren; dit is verse informatie die sinds de vorige verwerking ter beschikking staat en bevat vaak verrassende trends ten opzichte van het vorige moment. Hoe sneller een organisatie het water uit de bronnen vervoert naar het meer en de informatiezee hoe hoger de bijdrage van die gegevens is aan het resultaat van de organisatie. Hoe sneller het informatiesysteem de geleverde goederen factureert (een operationele respons) hoe eerder de betaling over het algemeen binnen is. Datamanagement beperkt zich dus niet alleen tot vraagstukken over gegevensstructuren en -modellen maar speelt ook een doorslaggevende rol in het proces dat zich 's nacht in het rekencentrum voltrekt.

Het verwerkingsproces bewerkstelligt juist redundantie om de organisatie snel en adequaat te kunnen laten reageren op de oorspronkelijke gebeurtenissen. Nadat de doelstellingen en de werkprocessen aan bod komen, sluit het artikel af met de specifieke invulling van de datamanagement-aspecten binnen het verwerkingsproces.

De nadruk in dit basisproces ligt niet meer exclusief op snelle opslag zoals het geval is in het registratieproces maar ligt hier op een snelle doorvoer naar de managementinformatie-omgeving die het responsproces ondersteunt om de gegevens en informatie snel te ontsluiten. In ieder generiek basisproces is snelheid weliswaar van belang; de richting van de snelheid verschilt echter aanmerkelijk. Figuur 4 geeft dit principe weer.

DE DOELSTELLINGEN

Het datamanagement binnen het verwerkingsproces spitst zich toe op het tijdig vervoeren, dupliceren en transformeren van nieuwe signalen die het registratieproces op heeft geslagen in de productiedatabase. Het heeft een groter procesmatig karakter dan het registratieproces. Aan elkaar geschakelde en van elkaar afhankelijke gegevensstromen transporteren de gegevens naar het lange termijn geheugen van de organisatie. De invulling van het datamanagement verschilt met het registratieproces en kent totaal andere accenten en doelstellingen. Ten eerste kanaliseert het de stroom signalen ter ondersteuning van de operationele en de strategische respons. Ten tweede bouwt datamanagement in het verwerkingsproces essentiële historie op om onder andere te voldoen aan administratieve verplichtingen. Daarnaast stelt de historieopbouw binnen het verwerkingsproces de organisatie in staat om veel betrouwbaarder de exacte

Het niet (goed) verwerken van gebeurtenissen leidt tot overspannen situaties

wijze waarop de bedrijfsprocessen functioneerden te bepalen. Dit belang neemt toe naarmate informatie betrekking heeft op een verder verleden. Tenslotte geeft dit organisaties vervolgens betere mogelijkheden om met terugwerkende kracht het verloop van de bedrijfsprocessen te toetsen aan de genomen maatregelen. Ten derde schoont het de buffers in het registratieproces zodat dit efficiënt blijft functioneren. Ten vierde stelt het datamanagement binnen het verwerkingsproces zich ten doel om data uit verschillende bronnen te integreren zodat een organisatie een compleet en synchroon beeld heeft van wat er zich afspeelt binnen het bedrijf als in de markt waarin zij opereert. Zodoende kan zij beter reageren op haar omgeving.

Bovenstaande doelstellingen met betrekking tot het datamanagement binnen het verwerkingsproces staan dus volledig ten dienste van beide andere generieke basisprocessen. Door te verwerken komt er ruimte in de productiedatabase

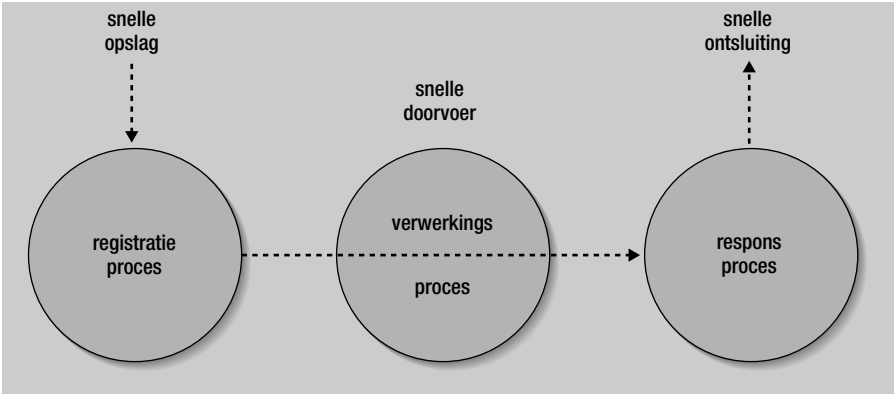
Van elkaar afhankelijke processtromen transofrmeren gegevens tot informatie

waarin het registratieproces de signalen wegschrijft en stelt het de organisatie in staat om snel en adequaat te reageren op de oorspronkelijke signalen die zijn opgevangen.

DE WERKPROCESSEN VAN DATAMANAGEMENT

In het eerste artikel zijn de datamanagementoperaties in grote lijnen aan de orde gekomen. De processen in draaitijdomgeving valideer, transformeer, dupliceer en consumeer hebben niet slechts betrekking op het generieke basisproces verwerking. Valideer (en uiteraard ook het opslaan) is een datamanagementoperatie horend bij het registratieproces, de datamanagementoperatie consumeer hoort bij het responsproces. De processen transformeer en dupliceer vallen onder het verwerkingsproces en deze worden ondersteund door de volgende werkprocessen:

- monitoring
- functionele controle en verslaglegging
- back-up en recovery
- opschonen en archivering
- scheduling.



FIGUUR 4: DE RICHTING VAN SNELHEID IN DE DRIE GENERIEKE BASISPROCESSEN

MONITORING

Monitoring ten aanzien van de beschikbaarheid, schijfruimte en geheugengebruik is niet alleen voorbehouden aan andere generieke basisprocessen. Het ligt voor de hand dat beheerders gewaarschuwd worden wanneer tijdens openingstijden de productiedatabase of -machine down gaat. Iedere minuut uit de lucht heeft vaak directe gevolgen voor de omzet. Het monitoren van een goede werking en verwerking van de gegevensstromen naar het lange termijn geheugen van de organisatie ligt minder voor de hand maar is minstens zo belangrijk. Betrouwbare en bruikbare informatie en de daarvoor benodigde verwerking van data wordt steeds meer een onderscheidend element in de concurrentiestrijd. Daarnaast speelt het steeds krapper wordend tijdsvenster bij calamiteiten in de verwerking nogal eens op. Uitstel van de verwerking of het vastlopen ervan kan ertoe leiden dat de productiedatabase 's morgens later in de lucht komt hetgeen ook tot directe omzetverliezen kan leiden. Een ander argument om de monitoring van het verwerkingsproces goed te organiseren is het gegeven dat vroegtijdige signalen van een onjuiste of incomplete verwerking herstel nog mogelijk maken. Dit voorkomt dat onjuiste informatie het responsproces binnendringt.

Monitoring is een veelomvattend aspect van datamanagement en is jammer genoeg in veel gevallen een sluitpost op de begroting. In niet weinig situaties trekken de gebruikers eerst aan de bel en zijn het niet de beheerders die een probleem zien (aankomen). Het proactief monitoren in het verwerkingsproces richt zich met name op de volgende zaken:

- zijn de verwerkingsprocessen succesvol en goed doorlopen
- de groei van de doorlooptijd van de verwerkingsprocessen ten opzichte van het beschikbare tijdvenster
- de groei van de schijfruimte (ook tijdelijke schijfruimte)
- het gebruik van het intern geheugen tijdens de verwerking
- uitval van data door slechte kwaliteit
- overige relevante bottle-necks (locks, connecties, buffers, etc.)

Het instrument om deze indicatoren in kaart brengen is logfiles. Deze geven op ieder gewenst tijdstip de huidige situatie van de indicatoren van het systeem weer en, mits deze goed gestructureerd en toegankelijk zijn, zijn het voorboden van een op handen zijnde crisis.

FUNCTIONELE CONTROLE EN VERSLAGLEGGING

Ook al zijn de verwerkingsprocessen succesvol afgerond en technisch goed doorlopen wil dat niet zeggen dat het ver-

Datamanagement en de markt

De maatschappij en de markten waarin organisaties opereren veranderen continu. Organisaties veranderen mee om geen klanten te verliezen en om hun toekomstige bestaan veilig te stellen. Markten en wensen ten aanzien van de producten en diensten veranderen in rap tempo en de organisatie moet mee. Ondernemen lijkt tegenwoordig op snelschaken. Er is nauwelijks tijd om de volgende zet te beramen. De voorsprong van vandaag is overmorgen weer verdwenen.
Deze veranderingen in ons huidige economisch stelsel beïnvloeden ook de inrichting en organisatie van datamanagement. Weken wachten op cijfers is er niet meer bij. Dagelijks moet het management gevoed worden met nieuwe feiten om grip te houden op de processen en de concurrentie voor te blijven. Dagelijkse verwerking is dan ook een absolute must; real-time verwerking soms onontbeerlijk. Daarnaast trekt de 24-uurs-economie met zijn ruimere openingstijden een zware wissel op de infrastructuur en architectuur van het verwerkingsproces.
De eisen die de economische en maatschappelijke veranderingen stellen aan het verwerkingsproces zijn aanzienlijk: een hoge doorvoersnelheid van de gegevens om de actualiteit zo hoog mogelijk te laten zijn, een hoge beschikbaarheid, een steeds nauwer wordend nachtvenster waarin de verwerking draait en een schaalbaarheid die gelijk op kan gaan met de groei van de organisatie. Soms zijn flinke investeringen nodig om aan deze eisen te voldoen.

werkingsproces de data goed overbrengt, transformeert en opslaat. Ten eerste kan een lek in de validatiestrategie binnen het registratieproces niet-integere data door hebben gelaten waardoor het verwerkingsproces data verdubbelt of juist weglaat. Het verwerkingsproces knoopt namelijk voor het registratieproces min of meer alleenstaande tabellen aan elkaar wat kan leiden tot het genoemde effect. Het aantal rijen van de brontabellen dient overeen te komen met die van de doeltabellen tenzij in het transformatie-ontwerp hier bewust rekening mee is gehouden. Het ontwerp en de implementatie daarvan kan een tweede oorzaak zijn van functionele fouten tijdens de verwerking. Verschillen treden aan het licht door een mechanisme in te bouwen dat voor en na de verwerking het aantal rijen telt of attributen sommeert naar verschillende invalshoeken. Bij verschillen dient de oorzaak ervan te worden achterhaald en te worden gerapporteerd aan de gebruikers van de gegevens en indien mogelijk te worden hersteld.

OVERIGE WERKPROCESSEN

Het behoeft geen lang betoog dat back-up en recovery belangrijk zijn. Wanneer de productiedatabase plat gaat en er is geen recente back-up (terug te zetten) dan kan het bedrijf zijn deuren wel sluiten. Hoe groter het gat tussen de datum van de back-up en het moment waarop de back-up teruggezet moet worden hoe groter de kans daarop. Daarnaast is de terugzettijd van de back-up eveneens van belang. Als het 4 dagen duurt om een back-up van één dag oud terug te zetten dan biedt dat ook geen soelaas meer. Geteste recovery-procedures zijn onontbeerlijk om verrassingen bij calamiteiten te voorkomen. In de hectiek van de calamiteit is dat een goede houvast.

Het opschonen en archiveren is van een andere orde. Het dient een ander doel dan de backup namelijk het bewaren van historische toestanden in plaats van actuele toestanden. Toch zijn deze twee processen met elkaar verweven en kan interferentie voor nare dingen zorgen. Daarom is het zaak om voor en na het archiveren een back-up te maken zodat bij eventuele calamiteiten de toestand van voor en na het archiveren kan worden hersteld.

Ten slotte zorgt het werkproces scheduling ervoor dat alle verwerkingsprocessen, maar ook de back-up-processen, na of naast elkaar kunnen draaien en dat fouten in de ene stroom niet de andere stroom tot stilstand brengen. Parallelle en van elkaar afhankelijke seriële processen, over meerdere platformen heen, zorgen voor grote complexiteit in de verwerking. Het vastleggen van deze afhankelijkheden is absoluut nodig om bij vastlopers het proces snel weer aan de praat te krijgen zonder dat er een leger ontwikkelaars aan te pas hoeft te komen. Ten slotte vergemakkelijkt dat het in productie nemen van nieuwe verwerkingsprocessen.

DE DATAMANAGEMENT-ASPECTEN

De woorden opslag, transport en ontsluiting geven richting aan de invulling van de datamanagement-aspecten in de drie generieke basisprocessen; zo ook in het verwerkingsproces. Bij een transport met bijvoorbeeld een vrachtauto, waar in plaats van gegevens goederen van A naar B gaan, spelen min of meer dezelfde aspecten een rol. De integriteit wordt bewaakt door op punt A en punt B de vrachtauto te wegen. Een ongelijk gewicht duidt op een hapering, een inconsistentie,

Geteste recovery-procedures zijn onontbeerlijk om verrassingen bij calamiteiten te voorkomen

in de lading. De vrachtauto kan niet te lang onderweg blijven want dan raken de goederen over datum. Het aspect actualiteit. De spoorwegen dienen als metafoor om een ander aspect toe te lichten. Wanneer de ene trein te laat is dan wordt aansluiting met de andere trein gemist. Het aspect afhankelijkheid.

De nadruk ligt in het verwerkingsproces toch wel op redundantie, een datamanagement-aspect dat in het registratieproces uit den boze is maar in het verwerkingsproces is toegestaan. Het voordeel ervan is dat de weg van A naar C niet meer via B hoeft te lopen. Daardoor zijn

minder joins nodig en over het algemeen bevordert dat de responstijd van queries. Redundatie moet echter niet leiden tot het ongebreideld kopiëren van gegevens die her en der verspreid liggen in de database of het ondernemingsnetwerk. Bij sommige slimme DBMS'en leidt redundantie zelfs tot een lagere responstijd doordat ze kleinere tabellen vóór de uitvoer van de query in het interne geheugen laden waardoor joins vermeden worden. Voor iedere situatie dient een afweging te worden gemaakt of de voordelen opwegen tegen de nadelen. De volgende criteria kunnen daarbij helpen:

- blijft de laadprogrammatuur onderhoudbaar
- verhoogt het de responstijd aanzienlijk
- leidt redundantie niet tot een te lange verwerkingstijd; met andere woorden blijft het passen in het tijdvenster
- wordt de gebruikte vorm van redundantie ondersteund door de ontsluitingssoftware.

Tot slot: redundantie manifesteert zich in allerlei vormen[1] en dat maakt een juiste en functionele toepassing ervan binnen het verwerkingsproces tot een uitdaging.

TRANSPORT IS DE SPIL

Datamanagement in het verwerkingsproces kenmerkt zich door een procesmatig karakter. Veelal van elkaar afhankelijke processtromen kanaliseren en transformeren de gegevens tot bruikbare grondstoffen om informatie te genereren. De informatiezee, waar alle gegevens uiteindelijk terechtkomen, zal een organisatie in staat stellen om de informatie in te zetten in het bedrijfsproces om zodoende rendement te behalen op het bezit van de geregistreerde gegevens. Het transport, zoals dat in het verwerkingsproces plaats vindt, is de spil waar het om draait. Voor een zo hoog rendement op data dient de spil snel, gecontroleerd en wellicht altijd te draaien zodat de gegevens tijdig en zonder informatieverlies beschikbaar komen voor het responsproces. ●

1 Horizontaal, verticaal en totaal slaan respectievelijk op denormalisatie, aggregaties en een kopie van een complete database.

Daan van Beek MSc (daanvanbeek@hetnet.nl) is Manager Datamanagement Services bij een groothandel in geneesmiddelen en farmaceutische producten.

Congressen, beurzen e.d.

10/10: Cognos Enterprise 2002
Relatiedag van deze BI-leverancier. Amsterdam, the Factory.
Org./inf.: www.cognos/enterprise2002.

10-11/10: OGh Jubileumcongres
Oracle-gebruikersgroep, derde lustrum. Maastricht, Crowne Plaza.
Org./inf.: www.oracle-usergroup.nl

21-24/10: IDUG 2002
Europese conferentie voor DB2-gebruikers. Lissabon. Org./inf.: www.idug.org

28-29/10: Data Mining Summit 2002
Congres van leverancier SPSS. Parijs, Le Meridien Etoile.
Org./inf.: www.dataminingsummit.com

3-8/11: TDWI World Conference
Congres van The Data Warehouse Institute. Orlando (VS), Gaylord Palms Resort & Convention Center.
Org./inf.: www.dw-institute.com

Cursussen, seminars e.d.

7, 8 en 14/10: Dimensionaal modelleren
Cursus met Harm van der Lek. Amsterdam ZO, Planetarium Gaasperplas.
Org./inf.: VanderLek Advies BV, www.vdlek.nl, (035) 6216928.

15-17/10: Informatie-analyse en logisch database-ontwerp
Seminar met Rick van der Lans. Gent (B), Holiday Inn Gent Expo, 9.00-17.00 uur. Kosten: € 1350. Org./inf.: I.T. Works, www.itworks.be/logdbontwerp.html, (00) 32 9 2415613.

23/10: Enterprise portals
Seminar met Peter Hinssen e.a. Diegem (B), Hotel Sofitel Brussels Airport, 14.00-21.00 uur. Kosten: € 545. Org./inf.: I.T. Works, www.itworks.be/, (00) 32 9 2415613.

6/11: Ervaringen met datawarehousing
Seminar. Diegem (B), Hotel Sofitel Brussels Airport, 14.00-21.00 uur. Kosten: € 545. Org./inf.: I.T. Works, www.itworks.be/, (00) 32 9 2415613.

7/11: Enterprise applicatie-integratie
Seminar met Peter Hinssen e.a. Diegem (B), Hotel Sofitel Brussels Airport, 14.00-21.00 uur. Kosten: € 545. Org./inf.: I.T. Works, www.itworks.be/, (00) 32 9 2415613.

11-12/11: Ontwerpen van de nieuwe generatie datawarehouses
Masterclass met Rick van der Lans. Leiden, Holiday Inn, 9.30-17.00 uur. Kosten: € 1250 (€ 1175 voor DB/M-abonnees).
Org.: Array Publications, info: www.array.nl, (036) 5409111.

13-14/11: Fysiek database-ontwerp
Seminar met Rick van der Lans. Gent (B), Holiday Inn Gent Expo, 9.00-17.00 uur. Kosten: € 980.
Org./inf.: I.T. Works, www.itworks.be/, (00) 32 9 2415613.

Alle vermelde bedragen zijn excl. BTW.