Datamanagement: beter omgaan met bezit (1)

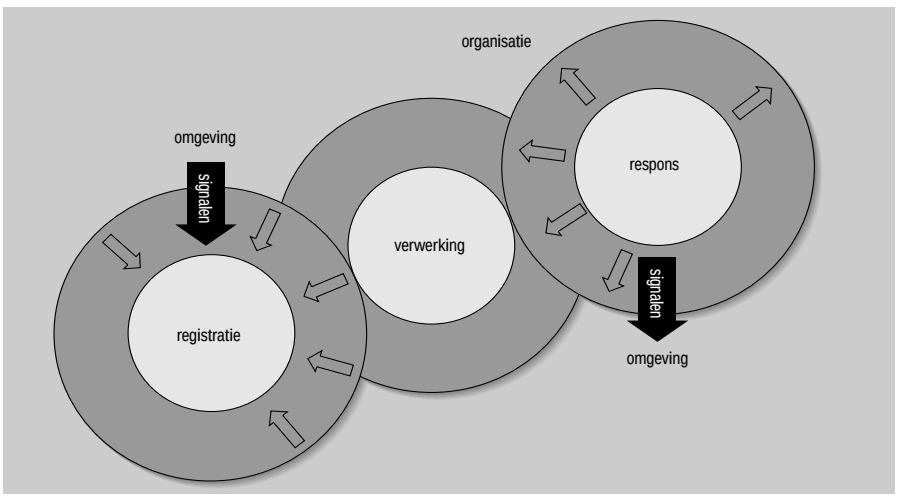
Van nullen en enen tot besluitvorming

Daan van Beek

Gegevens zijn veelzijdig en kennen vele gezichten. Data zijn overal aanwezig: van de betekenisloze binaire vormen in het hart van de computer tot betekenisvolle in een managementinformatiesysteem. Gegevens lopen door de computer, het bedrijfsnetwerk, Internet, middleware en applicaties. Data zijn het basis-ingrediënt van zeer veel computertoepassingen. Het is nog fundamenteler: het ontstaan van onze wereld leverde gegevens op. De allereerste straling, het eerste sterrenlicht en de vroegste geluiden - het is allemaal te duiden als signalen en data.

De allereerste organische cellen bleken nauwelijks in staat die signalen, zoals licht of geluid, te verwerken. Er was geen zichtbare interactie. Later onstonden eencelligen die reageerden op licht. De eerste dataverwerkers waren geboren. Zij illustreren onmiskenbare principes voor datamanagement, die nog steeds opgeld doen: een juiste registratie van de signalen, een correcte verwerking daarvan en het genereren van een adequate respons.

Organisaties doen er goed aan het beheer van gegevensstromen te professionaliseren. Juist in deze moderne tijd, waarin informatie-overbelasting hoogtij viert en de hoeveelheid data ieder jaar verdubbelt, neemt het belang van datamanagement sterk toe. De toegevoegde waarde van een organisatie ligt bovendien steeds meer in de aanvullende informatie die wordt verschaft aan de klant. Juiste, volledige, betekenisvolle en integere data zijn daarbij onontbeerlijk. De betekenis van data in letterlijke zin is een evenzo belangrijk punt. Bezit van gegevens volstaat niet om te blijven voortbestaan. Data moeten renderen en gebruikt worden. De gebruiksvriendelijkheid ervan moet daarom groot zijn.



Figuur 1: De generieke basisprocessen in een organisatie.

Een 'procesbenadering', zoals weergegeven in figuur 1 en tabel 1, werpt onder meer een ander licht op de taken en verantwoordelijkheden van de ICT-organisatie als geheel (zie kader *Basisprocessen voor datamanagement*). Datamanagement fungeert als fundament en middelpunt van de moderne, professionele ICT-organisatie; zonder data geen toepassingen.

DIMENSIES VAN DATA

Professioneel beheer van datastromen binnen een wat grotere organisatie is geen sinecure. Diverse, vaak tegenstrijdige belangen (bijvoorbeeld redundantie versus responstijd) en aspecten spelen een rol. Enige structuur, diverse werkmodellen, definities en een duidelijke verzameling richtlijnen zijn onontbeerlijk om inhoud te geven aan datamanagement.

Gegevens hebben een viertal dimensies (zie figuur 2). De inhoud van data verwijst naar concrete zaken in het dagelijks leven, zoals een specifieke klant of een bestelling. Validatie en beveiliging zijn hier de meest in het oog springende aspecten. De vorm waarin de gegevens zich manifesteren raakt de aspecten redundantie en gebruiksvriendelijkheid. De dimensie tijd beschouwt data in relatie tot de aspecten historie, responstijd en actualiteit. Ten slotte is de locatie, de plaats waar men gegevens opslaat, van belang. Actualiteit en redundantie van data zijn aan deze zijde relevante vraagstukken. Het aspect beschikbaarheid heeft eigenlijk betrekking op alle vier de dimensies. Gegevens zijn

niet op tijd, niet in de juiste vorm of niet op de juiste locatie beschikbaar, of zijn afwezig omdat registratie wordt verzuimd.

DEFINITIE EN PROCES

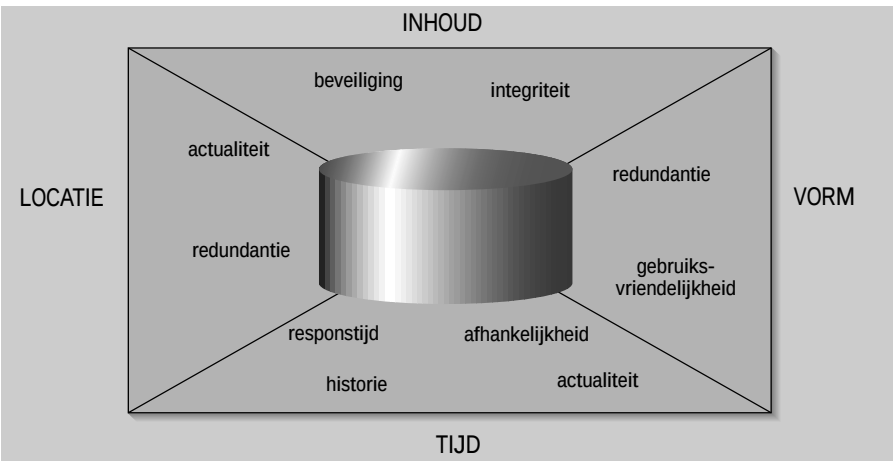
Datamanagement is het proces dat of de functie die voorziet in het verschaffen van toegang tot de gegevens, het uitvoeren en monitoren van de opslag ervan en het beheersen van de daarbij behorende in- en uitvoeroperaties[1]. Datamanagement verschaft ten eerste toegang tot data, ten tweede slaat het gegevens op en ten derde transporteert het die.

De hamvraag is hoe dit te organiseren, zodat data rendabel kunnen worden voor de organisatie. De basisoperaties binnen datamanagement geven hierbij richting. Figuur 3 toont die operaties. De bovenste drie operaties hebben een meer ontwerpachtig karakter. Een draaitijdomgeving, zoals weergegeven binnen de rechthoek, verduidelijkt de overige operaties. In de ontwerpomgeving wordt het datamodel vormgegeven. De entiteiten krijgen attributen en worden voorzien van onderlinge relaties. De levenscyclus van een data-element is begonnen. De database-ontwerper implementeert het model in het informatiesysteem en beveiligt de gegevens door autorisaties. De publicatie van het datamodel is voor een efficiënte ICT-organisatie van groot belang. Ontwikkelaars van applicaties, managementinformatiesystemen en CRM-systemen zijn maar wat graag van het datamodel op de hoogte. Ook gebruikers weten graag welke gegevens worden geregistreerd en wat de betekenis ervan is. Metadata zijn een vereiste voor toegang tot data en het opvragen van informatie.

Principes voor datamanagement zijn al oud: juiste registratie, correcte verwerking en het genereren van adequate respons

De geschakelde processen aan de onderkant van de figuur geven zicht op de datastroom en de daarbij behorende datamanagementprocessen wanneer het informatiesysteem in werking is. Het eerste proces, valideer, zorgt ervoor dat niet-integere data niet worden opgeslagen. De daaropvolgende processen, transformeer en dupliceer, draaien meestal 's nachts in batchverwerkingsmode. Het betreft gegevens die overdag in de bedrijfsdatabase zijn veranderd. Het proces 'publiceer en consumeer' in deze schakel genereert direct de respons (bijvoorbeeld in de vorm van een factuur) of ondersteunt de organisatie bij het genereren van een adequate respons (bijvoorbeeld door de frequentie van de klantbezoeken op te schroeven, met als doel de omzet per klant te verhogen).

Deze vier processen zijn tijdens het draaien van het informatiesysteem het beste zichtbaar. Maar deze processen en de daarbij behorende vereiste logica en kennis ontstaan niet vanzelf. De databaseontwerper dient deze logica zorgvuldig te modelleren, in de vorm van validatie-, transformatie- en duplicatieregels. Ook voor de doorgewinterde ontwerper is dit een hele uitdaging.



FIGUUR 2: DE VIER DIMENSIES VAN DATA EN DE BESCHIKBAARHEID ERVAN.

REGISTRATIE

Nauwkeurige registratie en gestructureerde opslag van gegevens vormen een belangrijke schakel in de levenscyclus van data. Al te vaak komt het voor dat men zaken dubbel registreert of onvolledig of onjuist invoert. Een goed ontwerp, met

Basisprocessen voor datamanagement

De drie basisprocessen registratie, verwerking en respons nopen tot een uitgekiende methode voor datamanagement. Ieder basisproces heeft namelijk zo zijn eigen behoeften voor de invulling van datamanagement-aspecten. Tabel 1 geeft deze weer.
In nieuwe serie artikelen neemt Daan van Beek deze aspecten voor ieder basisproces afzonderlijk onder de loep. Zij beïnvloeden elkaar onderling en zijn net een verzameling 'communicerende vaten': neerwaartse druk op de vloeistof in een kolom leidt tot opwaartse druk in een of meer van de andere kolommen. Meer redundatie voor een betere reponstijd, minder flexibiliteit bij het invoeren van gegevens voor een betere integriteit - ieder basisproces kent een optimum.
De processen verwerking en respons vallen uiteen in een operationele en tactische variant; het datamanagement ter ondersteuning van deze processen verschilt aanzienlijk.
Twee afsluitende artikelen leggen de nadruk op het beheer van gegevens vanuit organisatorisch en technologisch perspectief. Deze laatste invalshoek toont innovatieve concepten en gereedschappen voor het inzetten van metadata als basis voor datamanagement.

Aspecten	Invulling		
Integriteit	integer	consistent	synchroon
Beveiliging	muteren	stabiliteit	kijken
Actualiteit	seconden	—	uren, dagen
Historie	overschrijven	bewaren	opbouwen
Redundantie	kopie	tabel	database
Responstijd	seconden	uren	minuten
Beschikbaarheid	minuut	uur	uren,dagen
Gebruiksvriendelijkheid	laag	—	hoog
Afhankelijkheid	volledig parallel	parallel/serieel	serieel
Transacties p/sec.	zeer veel	weinig	beperkt
Type transacties	noteer	dupliceer	transformeer
Volume per transactie	laag	fors	reusachtig
Basisprocessen	**Registratie**	**Verwerking**	**Respons**

TABEL 1: INVULLING DATAMANAGEMENT-ASPECTEN VOOR IEDER BASISPROCES.

name van de gebruikersinterface en de integriteitsregels in de database, is daarbij van essentieel belang. Hoge datakwaliteit is het doel. Fouten die er in dit eerste proces insluipen hebben verregaande consequenties in de daaropvolgende processen, verwerking en respons. Het registratieproces vraagt niet alleen een relatief snelle responstijd en bewaking van de data-integriteit, maar dient ook afgestemd te zijn op diverse communicatiekanalen (telefoon, brief, Internet) waarvan de omgeving gebruik maakt.

Ieder communicatiekanaal stelt zo zijn eigen eisen aan de organisatie van data. Een telefonische bestelling dient direct gepaard te gaan met een informele respons of het artikel leverbaar is. Een goed informatiesysteem én een juist ontworpen database-architectuur ondersteunen dit principe. Een schriftelijke bestelling stelt deze eis in mindere mate, hoewel frustatie van de gebruikers op de loer ligt. Bij een bestelling via Internet is een snelle responstijd nog essentiëler. Er is immers geen medewerker in de buurt die de klant "aan de praat" houdt. Terwijl veelal slecht ontworpen mid-office-databases voor een snelle responstijd zorgen, zijn actualiteit, integriteit en redundantie van data vaak bedroevend slecht.

VERWERKING

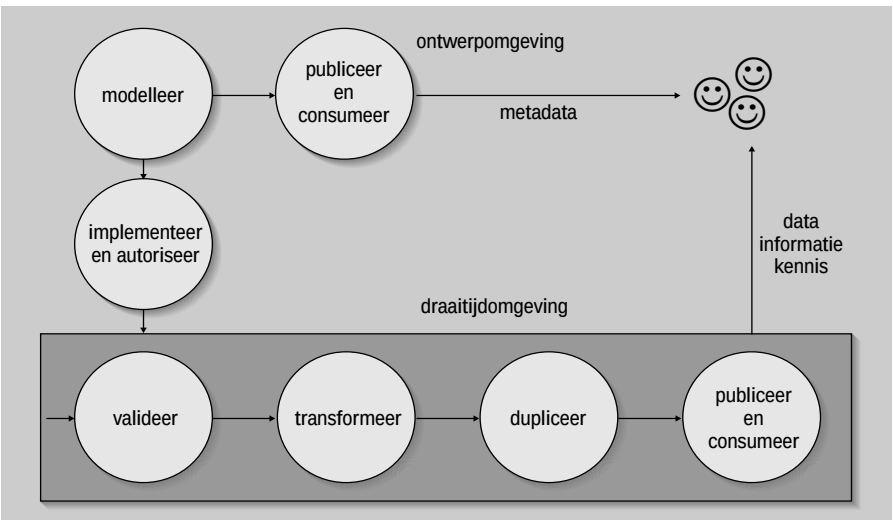
Het verwerkingsproces verschuift de gegevens van het korte- naar het lange-termijngeheugen van een organisatie. Veelal is batchgewijze verwerking daarbij aan de orde. Voordat men de data opslaat in het lange-termijngeheugen zijn controles op integriteit en consistentie noodzakelijk. Immers, geen enkele organisatie is zo professioneel, dat alle integriteitsregels ook daadwerkelijk afgedwongen worden. Daarbij komt dat de huidige dbms'en het valideren van bedrijfsdata nog niet goed ondersteunen[2]. Bij complexere, meer bedrijfsgeoriënteerde integriteitsregels laten ze het afweten.

Belangrijk aspect bij de gegevensverwerking is opbouw van operationele historie. In de eerste plaats omdat men nieuwe attributen en entiteiten in het leven roept, waarvan de inhoud een afgeleide is van andere. Denk bijvoorbeeld aan het BTW-bedrag, dat wordt afgeleid van de omzet en een BTW-code (hoog/laag). Het ontwerp van menige database heeft geen rekening gehouden met mogelijke mutaties op het BTW-percentage. Ten tweede is het opslaan van afgeleide attributen vaak noodzakelijk om een korte doorlooptijd te bereiken voor het genereren van een snelle respons (afdrukken van de factuur). In de derde plaats vereisen bepaalde overheidsinstellingen, zoals de belastingdienst in het geval van het BTW-bedrag, bepaalde gegevens expliciteit te administreren[3] of ervoor te zorgen dat met terugwerkende kracht (tot zeven jaar) de gegevens eenduidig en eenvoudig zijn te herleiden.

Doel van elke vorm van datamanagement is optimale ondersteuning van een organisatie

Daarnaast voorziet dit proces in extractie van mutaties in de productiedatabase en verplaatst het deze ter analyse naar een managementinformatie-omgeving. Die gegevensstroom heeft als doel de tactische en strategische respons te ondersteunen.

Bovenstaande processen draaien vaak in de operationele omgeving. Hierbij is het van groot belang dat afhankelijkheden tussen verwerkingsprocessen inzichtelijk zijn. Sommige daarvan mogen pas draaien als



FIGUUR 3: BASISOPERATIES DATAMANAGEMENT.

De 3-lagenarchitectuur

Data worden pas opgeslagen als zij zich toegang verschaffen tot de drie lagen. De eerste laag, de gebruikersinterface, stuurt de gegevens door naar de applicatielaag, die op zijn beurt zorgt voor doorvoer naar de datalaag. Deze gegevenslaag is verantwoordelijk voor de fysieke opslag op schijf (voor dit artikel niet getekend). Ten slotte gaan er signalen terug van beneden naar boven of de transactie succesvol is geweest. De gebruikersinterface kan bestaan uit een command-line-gestuurde interface, voor batchachtige verwerkingen, of een GUI, waarin gebruikers data invoeren met formulierachtige schermen.
Een goed ontwerp van de gebruikersinterface kan sterk bijdragen aan de kwaliteit van de data. Zogenaamde keuzelijsten waaruit een gebruiker kiest beperken de kans op onjuiste gegevensinvoer.
Dit wil niet zeggen dat de gebruikersinterface de meest geëigende plek is voor validatie. Iedere database ontvangt immers doorgaans ook invoer vanuit andere applicaties. Validatieregels horen dan ook thuis in de database. De drie lagen zijn nauw met elkaar verbonden.

een ander verwerkingsproces succesvol is geweest. Handmatige of automatische controle op een juiste verwerking is daarbij essentieel. De professionele ICT-organisatie zal dit in procedures vastleggen.

RESPONS

Juiste toepassing van alle principes van datamanagement in de voorgaande processen zorgt op zijn minst voor een schone responsomgeving. De aandacht kan dan uitgaan naar een optimale rangschikking van de data, zodat een snelle responstijd gegarandeerd wordt. Principes, methodieken en technologieën op het gebied van datawarehousing helpen daarbij. Een datawarehouse is een hulpmiddel om data uit verschillende systemen te synchroniseren, historie op te bouwen en zorg te dragen voor een snelle responstijd. Gegevens worden daar waar nodig opgepoetst en geschoond, het systeem voegt entiteiten samen, vervangt foreign keys door betekenisloze sleutels en zet voorberekende totalen klaar.

Professioneel kennis- en datamanagement gaan gelijk op

Definities en betekenis van attributen raken hier het wezen van de organisatie en de bedrijfsprocessen. Fouten en onjuiste interpretaties in deze procesgang kunnen missers veroorzaken met grote, nare gevolgen. Immers, bij grote hoeveelheden data neemt het belang van een juiste betekenis en interpretatie kwadratische vormen aan. De beschikbaarheid van juiste en volledige informatie over de data is dan essentieel.

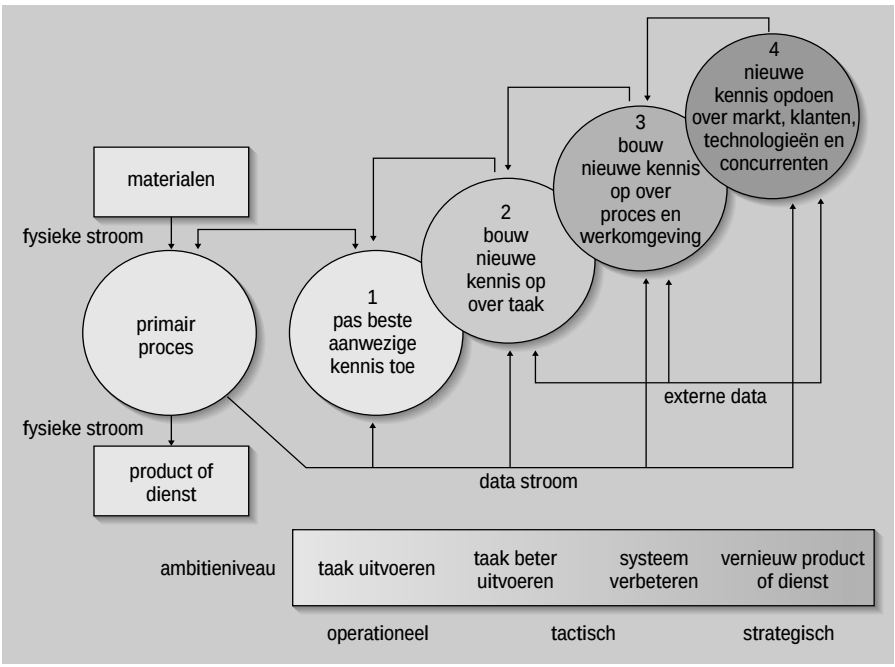
Om succesvol te grasduinen in gegevens is niet alleen de betekenis van belang. Gebruiksvriendelijke rapportage- en analysetools, gecombineerd met een zeer hoge responstijd van het datawarehouse, stellen medewerkers in staat trends, afwijkingen en hot-spots op te duikelen. Door te jongleren met en dankzij de wil te leren van data kan een organisatie hierop adequaat reageren. Concrete maatregelen en strategische beleidsontwikkeling maken de cirkel van actie-reactie rond.

LEREN VAN DATA

Dat mensen nieuwe 'feiten' en de toepassing daarvan zich eigen maken in (de bedrijfsprocessen van) de organisatie is niet vanzelfsprekend. Mensen, en daarmee ook organisaties, zijn niet snel tot verandering geneigd, vaak ook omdat helder zicht op de meerwaarde ervan ontbreekt. Kennismanagement, als proces van voortdurende verbetering en verandering, speelt hierin een grote rol. Hoe beter een organisatie de datastromen beheerst, hoe beter kennismanagement tot ontwikkeling kan komen.

Figuur 4 geeft de diverse ambitieniveaus weer van kennismanagement[4]. Een hoger ambitieniveau vraagt om een meer doordacht en integraal datamanagement. Immers, ambitieniveau 2 vereist reeds integratie van gegevens uit verschillende systemen. Het volgende ambitieniveau verplicht tot het samenvloeien van interne met externe datastromen. Professioneel kennis- en datamanagement gaan gelijk op.

De beslisser anno 2002 komt om in de



FIGUUR 4: AMBITIENIVEAUS VAN KENNISMANAGEMENT.

informatie die hem of haar ter beschikking staat. Selectieve en voorbedachte kengetallen zijn nodig om *information overload* te voorkomen. Wanneer datamanagement in de onderneming juist geïmplementeerd is, vormen de gegevens in de databases een heldere spiegel van de bedrijfsprocessen. Observatie van deze spiegel levert gedetailleerde kennis op over het bedrijfsproces, de markt, de klanten en de organisatie.

Sinds de opkomst van OLAP -eind jaren tachtig, begin jaren negentig- en later ook door datamining kreeg men in de gaten dat er allerlei goudklompjes verbor-

Nauwe samenwerking tussen disciplines is vereist

gen zitten in de bedrijfsdatabases. Correlaties tussen allerlei attributen geven bedrijven beter inzicht in hun bedrijfsprocessen en allerhande waarschijnlijkheden over bijvoorbeeld de winstgevenheid van klanten. Dergelijke klantkennis brengt de onderneming naar een hoger niveau dan een concurrent die nog niet zo ver is dat hij data en kennis kan inzetten in het bedrijfsproces. Leren van data is essentieel voor organisaties die vandaag de dag willen overleven.

CONCLUSIE

In dit artikel is stilgestaan bij de inhoud en de betekenis van datamanagement voor een organisatie. Enerzijds verschilt de precieze invulling van datamanagement per generiek basisproces; anderzijds kent datamanagement zelf diverse generieke werkprocessen en -operaties. Doel daarvan is altijd optimale ondersteuning van een organisatie voor het generen van een adequate respons. Professionalisering van datamanagement is een voorwaarde om te leren van data en voor het bedrijven van kennismanagement. Gebruik van metadata speelt bij die professionalisering een zeer belangrijke rol. Ten slotte is datamanagement nauw verweven met andere disciplines, die applicaties ontwikkelen en gebruikersinterfaces ontwerpen. Nauwe samenwerking tussen deze disciplines is vereist, zodat de drie lagen optimaal op elkaar zijn afgestemd. ●

Daan van Beek MSc is manager datamanagement services bij een groothandel in geneesmiddelen en farmaceutische producten.

Noten en literatuur

1. McDaniel, George, ed. *IBM Dictionary of Computing*. New York, NY, McGraw-Hill, Inc., 1994
2. Veldwijk, René, Verschuivende grenscontroles in het datadomein, DB/M 3, mei 2001.
3. Zie ook de 'interessante' brochure *Uw geautomatiseerde administratie en de fiscale bewaarplicht*
4. Rob van der Spek en André Spijkervet, *Kennismanagement: Intelligent omgaan met kennis*, Utrecht, Kenniscentrum CIBIT, 1996

Marco, David, *Building and Managing the Meta Data Repository*, New York, John Willey & Sons, Inc., 2000

Vervolg van pagina 21.

In een latere fase maakte de bank ook gebruik van SSD's op zijn Sybase-systemen. Hierbij bleek dat bijvoorbeeld het aanmaken van indexen geen 3,5 uur meer duurde, maar slechts een halfuur in beslag nam - een performancewinst van 700 procent. Een Scandinavisch bedrijf ('s werelds grootste containervervoerder) heeft onlangs SSD ingevoerd voor de logging in MQSeries. MQSeries maakt het mogelijk mainframedata aan te bieden aan open systemen en omgekeerd. Door de zeer kleine logfiles op SSD te plaatsen heeft men een performancewinst van 800% geboekt (figuur 2).

GROTE PERFORMANCEWINST

Door een slimme inzet van een kleine hoeveelheid solid state disks ligt een grote performanceverbetering van de database binnen bereik.

Voordeel van deze toepassing is de snelheid van implementatie. Er zijn geen aparte drivers nodig en in sommige gevallen (fiber) is het zelfs mogelijk 'on flight' aan te koppelen. Een mogelijk nadeel is de moeilijkheid de juiste 'hotspots' te vinden. Maar vaak is dat te omzeilen door de voor de gebruikte database bekende hotspots op SSD te plaatsen. In elk geval moet de organisatie er zeker van zijn dat de performanceproblemen waarmee zij kampt I/O-gerelateerd zijn.

Een ander nadeel is de hoge aanschafprijs. Dit wordt echter gecompenseerd door de langere afschrijvingstermijn; het systeem gaat meerdere servergeneraties mee. Bovendien valt dikwijls te besparen op processoren en op dure cache-centric disksystemen. Leveranciers van deze laatste, die daarmee succesvol zijn in de mainframemarkt, claimen vaak ten onrechte dat hun oplossingen ook doelmatig functioneren in open systemen. In de mainframewereld zorgen pakketten als DFDFS (IBM) voor het handhaven van controle over wat de cache bijhoudt. In open systemen gebeurt dit niet. Daardoor is een veel grotere cache nodig, die vaak minder efficiënt is dan een kleine SSD.

Door de grote belangstelling voor de SSD zal de prijs de komende jaren sneller dalen dan de kosten van geheugenruimte. De hoge prijs wordt nu vooral bepaald door het geheugen en de relatief kleine aantallen waarin de apparatuur wordt geproduceerd. Overigens zijn de kosten van een systeem van ondergeschikt belang als het de gebruiker op een snelle manier van zijn de problemen verlost. ●

Ing. G.R.M. Brouwer
(G.Brouwer@nehgroup.com) is adviseur bij Storage Expertise Holland en mede-oprichter en partner van dat bedrijf. In samenwerking met de Evaluator Group uit Colorado levert hij leverancieronafhankelijk advies binnen de storagemarkt, vooral op het gebied van storage area networks.