

Tot nu toe beschikbare tools nog volop in ontwikkeling

Dataclassificatie nog in kinderschoenen

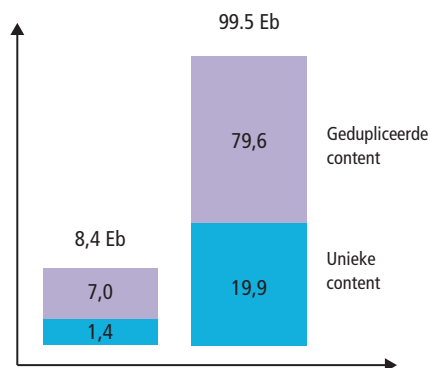
Wereldwijd groeit de hoeveelheid data explosief. Om de beheerkosten van al die gegevens in de hand te houden zijn tools nodig die data efficiënt en op het geschiktste medium kunnen opslaan, afhankelijk van betekenis en type data.

Bram Dons brengt het aanbod van dergelijke classificatietools in kaart en schetst de stand van de techniek op dit gebied.

Bram Dons

De afgelopen jaren heeft er een explosieve datagroei plaatsgevonden (zie figuur 1), een probleem waarmee bijna iedere organisatie de afgelopen tijd wel mee te maken heeft gehad. Het antwoord van de opslagindustrie op dit probleem was eenvoudig: koop meer diskopslagcapaciteit, want dat is het goedkoopste en snelste!

Tot voor kort leek dit ook een afdoende oplossing waarmee de meeste ondernemingen vrede hadden. Naar nu blijkt echter ten onrechte: alhoewel de hardwarekosten per opgeslagen databyte nog steeds dalen, zijn de kosten voor het beheer daarvan explosief toegenomen.



Figuur 1 Wereldwijde digitale content in exabytes (De X-as verbeeldt de periode tussen de introductie van de pc en heden)

Ongestructureerde data

Om de beheerkosten in het petabyte-tijdperk in de hand te houden (het gaat dan nog niet eens om een daling) is een oplossing nodig die data efficiënt en op het juiste medium kan opslaan. Welk medium dat is, is afhankelijk van de betekenis van en het type data en de

gewenste periode van opslag. De vraag is dus welk type data op welk type opslagmedium moet worden opgeslagen en voor hoe lang.

Voor een effectief databeheer is het dan ook van groot belang om informatie te verkrijgen over de data die beheerd moeten worden. In de meeste organisaties worden ongestructureerde data echter nog steeds op één en dezelfde manier behandeld, waarbij geen enkele relatie met de zakelijke waarde bestaat. Spreadsheets met zakelijke data worden bij wijze van spreken met dezelfde Service Level Agreements behandeld als vakantiekiekjes of downloads van iTunes!

Het probleem hierbij is dat de enige persoon die de inhoud en betekenis van een bepaald bestand kent, degene is die het gecreëerd en opgeslagen heeft. Maar een opgeslagen bestand wordt meestal gecirculeerd om het bij andere gebruikers onder de aandacht te brengen, die het vervolgens zelf weer opslaan in een ander mapje, en soms onder een andere naam. Dit zichzelf herhalende proces heeft onbedoeld een epidemische vorm aangenomen, waardoor beheerders helemaal niet weten wat, hoe, en waar wordt opgeslagen. Het wordt nog erger wanneer we ook het back-upproces erbij betrekken, want elk gekopieerd bestand wordt ook nog eens in de back-up meegenomen.

Data met en zonder structuur

Ongestructureerde data: tekst- tot en met videobestanden, spreadsheets, mp3's, et cetera.

Halfgestructureerde data: e-mail.

Gestructureerde data: databasebestanden.

Information lifecycle management

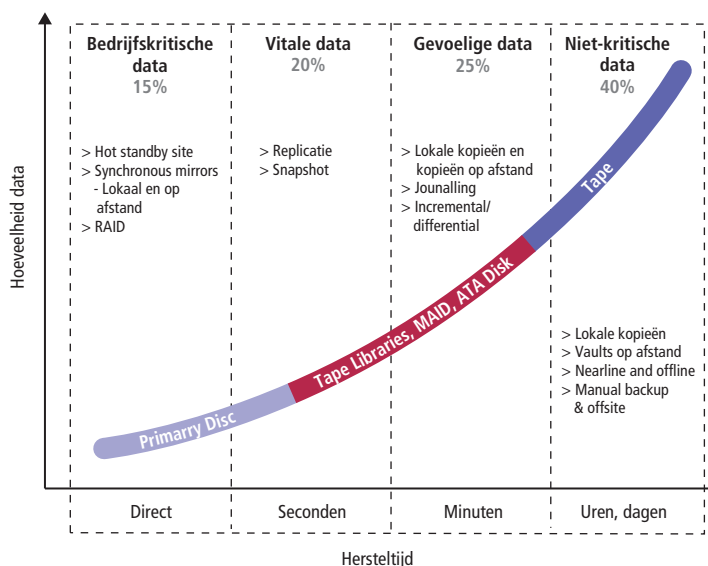
De opslagindustrie dacht met de invoering van information lifecycle management (ILM)

dé oplossing te hebben gevonden voor het opslagprobleem, met een indeling van opslagcapaciteit in een gelaagd opslagmodel. Het is echter de vraag of bij dit soort toepassingen werkelijk sprake is van een ILM-implementatie of dat het meer gaat om een variant op het traditionele model voor hiërarchische opslag. De gecreëerde gelaagde architectuur blijft beperkt tot het verplaatsen van informatie naar een andere (meestal goedkopere) opslaglaag naarmate de informatie ouder wordt. Al dit soort 'ILM-toepassingen' geeft geen inzicht in de betekenis van de opgeslagen data, op basis waarvan een beslissing kan worden genomen over het bewaren, analyseren of verwijderen van data.

Voor een geslaagde toepassing en invoering van ILM is het nodig in eerste instantie een idee te krijgen van de manier waarop ongestructureerde data beheerd moeten gaan worden. Dit kan antwoorden op een aantal cruciale vragen opleveren, zoals:

- Heeft het zin om een ILM-strategie in te voeren?
- Hoeveel niet-zakelijke data worden er beheerd?
- Hoeveel van de opslagcapaciteit gaat verloren aan geduplicateerde data?
- Welke data worden niet langer gebruikt en kunnen naar een minder duur opslagstelsel worden overgeheveld?
- Welke wet- en regelgeving is van toepassing?

Men krijgt alleen een beter inzicht in de levenscyclus van een dataobject als het systeem dat het object beheert zicht heeft op de data vanaf het moment dat die gecreëerd worden. En dus ook op alle daaropvolgende activiteiten. Te veel ILM-methodologieën baseren zich op *events*



Figuur 2 De waarde van data

bij het in kaart brengen van de bedrijfswaarde van een object. Het gaat echter veel meer om vragen als: wie heeft het bestand gecreëerd, wat is waarde daarvan voor de onderneming, wat gebeurt er wanneer iemand de informatie verwijderd, hoe afhankelijk zijn andere personen van deze informatie?

Naast de invoering van een efficiënt opslagstelsel voor het verkrijgen van inzicht in de opgeslagen data is het ook belangrijk dat er informatie wordt gedistilleerd uit alle al opgeslagen 'ruwe' data. Die informatie kan om allerlei redenen van groot belang zijn voor de onderneming. Helaas bestaat er nog geen technologie om enige betekenis toe te kennen aan ruwe data. Een groot deel van de in data opgeslagen zakelijke waarde blijft dan ook verborgen, omdat de tools, tijd en kennis om data te analyseren nog niet beschikbaar zijn. Veel leveranciers van ILM-tools hebben de mythe gecreëerd dat het aanbrengen van een opslagstructuur 'kennis' over de opgeslagen data oplevert, maar niets is minder waar.

Dataclassificatieproducten

Het is duidelijk dat wie een ILM-methodologie wil invoeren, allereerst de opgeslagen data in kaart moet brengen. Daarvoor er zijn de laatste jaren speciale applicaties ontwikkeld, de zogenaamde dataclassificatietools (DC-tools). Het

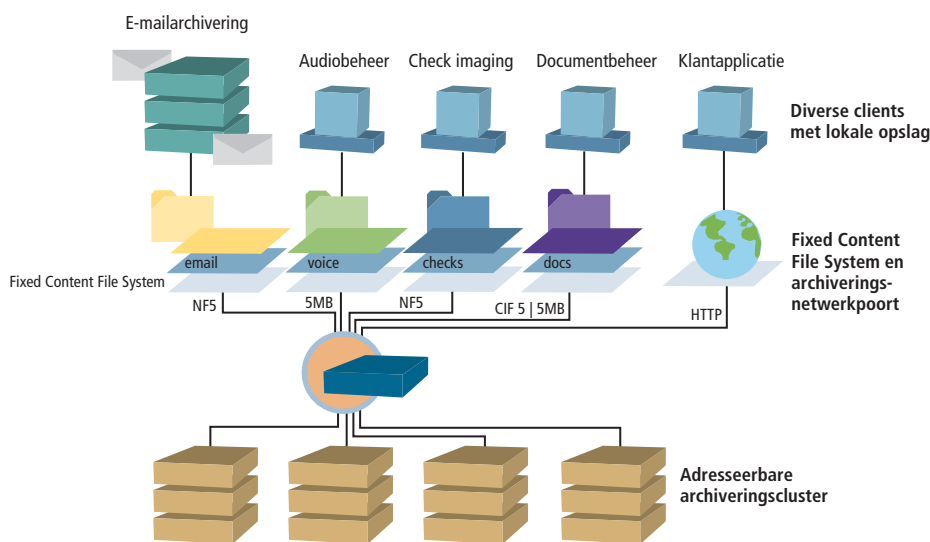
aanbod van DC-producten is echter nog gefragmenteerd: de ene leverancier richt zich op gestructureerde en de andere op halfgestructureerde of ongestructureerde informatie (zie ook kader 'Data met en zonder structuur').

Er zijn enkele ILM-achtige producten van grote leveranciers beschikbaar en er is een aantal *start-up*firmas actief in de markt. EMC heeft als een van de eerste bedrijven EmailXtender op de markt gebracht, maar dat product is alleen geschikt voor halfgestructureerde data. DatabaseXtender van EMC richt zich wel meer op gestructureerde data. Reference Information Storage platform (RIS) van HP richt zich ook in hoofdzaak op halfgestructureerde informatie, en Princeton Softech is meer gespecialiseerd in de classificatie van gestructureerde data. Hier passeren verschillende nieuwkomers de revue, die zich grotendeels richten op de ongestructureerde-DC-productmarkt: StoredIQ, Njini, Kazeon, Arkivio, Archivas en Index Engine.

Archivas

Het eerste product is ArC van de firma Archivas (zie figuur 3), een start-up die archiveringsystemen ontwikkelt voor zeer grote hoeveelheden vastgelegde content. Archivas ziet zijn doelmarkt ergens tussen de high-performance computingomgeving en de langetermijnopslagssystemen. Het product richt zich ook

dossier asset management



Figuur 3 Archiveringsclustersysteem ArC van Archivas

op de markt voor PACS-systemen (Picture Archive Communication System).

Het toepassingsgebied voor ArC begint bij dertig Tb opslag, maar opslagfaciliteit boven de 100 Tb is een aannemelijker waarde. Een van de voordelen van dit systeem is dat de data altijd toegankelijk blijven, ongeacht het type gebruikte applicatie. Het Fixed Content File System (FCFS) van ArC emuleert een bestandsysteem dat applicaties direct toegang geeft tot de archive content alsof het 'gewone' bestanden zijn. ArC biedt WORM-functionaliteit (*write once, read many times*) conform de SEC 17a-4-regelgeving. Het voorziet elk opgeslagen bestand van een digitale handtekening. Het product kent nog geen HIPAA-module, maar Archivas stelt dat ArC voor elke vereiste regelgeving te configureren is.

Arkivio

De firma Arkivio biedt naar eigen zeggen als een van de eerste bedrijven een complete ILM-omgeving, auto-stor, met daarin opgenomen de drie daarvoor essentiële ILM-componenten: *agentless data collection*, *data classification* en *databaseer*. De software verzamelt en analyseert volgens de leverancier gedetailleerde informatie over het gebruik van de opgeslagen data en de zakelijke waarde ervan gedurende hun bestaan. Dit komt in feite neer op het definiëren van enkele eenvoudige gegevens – zoals

bepaalde trefwoorden, teksten, het type bestanden en dergelijke – aan de hand waarvan de tool de waarde van de gegevens vaststelt. Als bedrijfseconoom kun je hier beslist geen bedrijfskundige waarde aan ontlennen.

Aan de hand van de opgegeven criteria worden de data geclassificeerd en van een prioriteit voorzien. Beheerders kunnen policy's creëren voor automatische beheerdoeleinden (waaronder migratie, kopiëren, bewaartijd en verwijdering) voor de consolidatie en optimalisatie van opslagbronnen binnen een heterogene opslagomgeving.

De Arkivio auto-storsoftware biedt de benodigde schaalbaarheid voor grote IT-omgevingen. Gemiddeld genomen moet elke RSA-server vijftig miljoen bestanden of twaalf Tb data kunnen beheren. Er is een maximum van 1 miljard bestanden of 240 Tb opslagcapaciteit.

De verzamelde data kunnen in twee logische groepen worden verdeeld: *volume groups* en *file groups*. De beheerder kan een volume group samenstellen aan de hand van monitoring-, rapportage- en policygebaseerde beheerdoeleinden. File groups worden automatisch door Arkivio auto-stor gegroepeerd op basis van algemene bestandstypen – bijvoorbeeld Microsoft Office-documenten en images – of door een selectie te maken op basis van bekende bestandsgegevens als type, extensie, leeftijd, grootte, toegang en directory path. Auto-stor kan een

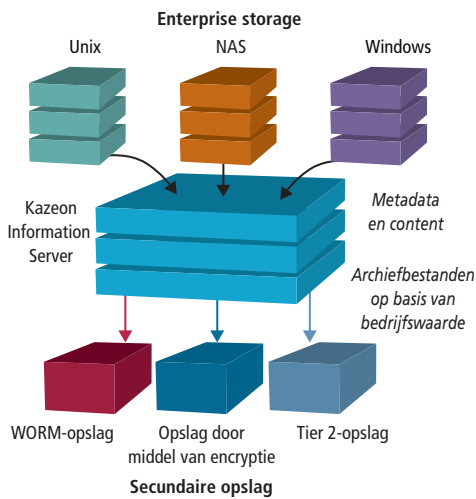
bepaalde classificatie aan een file group toekennen.

Ook behoort het tot de mogelijkheden data onder te brengen in een opslagapparaat dat *retention periods* (de periode vanaf het moment dat een recoveryproces gestart wordt tot het moment dat weer over de data beschikt kan worden) ondersteunt en data te archiveren op een *write-once* apparaat, om te voldoen aan wettelijk gestelde opslageisen. Arkivio auto-stor is sterk geïntegreerd met API's van EMC- en NetApp-apparatuur. De firma pretendeert weliswaar een volledige ILM-oplossing te bieden, maar in feite wordt alleen een selectie gemaakt op basis van metadata over het bestandsysteem en niet op de *inhoud* van bestanden.

Kazeon

Het eerste product van Kazeon, Information Server IS1200 (zie figuur 4), classificeert data op basis van bestandscontent en metadata over het bestandsysteem. Bestandscontent kan worden geclassificeerd op basis van geselecteerde woorden of tekst en bestandseigenschappen (titel, onderwerp, auteur, firmanaam of commentaar) en specifieke velden (project- of afdelingsnamen, telefoon- en accountnummers). Gegevens over het bestandssysteem kunnen worden geselecteerd op basis van de bestandsnaam en het toegangspad, grootte, eigenaar, creatie- en modificatietijd en tijdstip van de laatste toegang tot de data. Door de gebruiker gedefinieerde regels, bestandscontent en opgevraagde metadata kunnen dienen voor de classificatie en policy-toekenning. Daarnaast kunnen via API's en een GUI extra metadata worden aangekoppeld. Verder is een unieke contentgebaseerde 'vingerafdruk' voor ieder bestand te berekenen.

De zoekmachine van de IS1200 kan zoeken aan de hand van de volledige tekst, sleutelwoorden en selecteerbare metadata velden. Er zijn verschillende zoekmethoden: op basis van een booleancon-



Figuur 4 Kazeon-archiveringsarchitectuur Information Server

structie, wildcards en zinsconstructies, of via een veld gespecificeerde metadata en data ranges. De zoekresultaten zijn te filteren.

De Kazeon Information Server is gecertificeerd voor het Centera-opslagsysteem van EMC. Het product is ook te koppelen aan NetApp's SnapLock (past het *write once, read many times*-principe toe op harde schijven in storage-omgevingen), waardoor een complete omgeving met bestandsarchivering wordt ondersteund. Op basis van een gedistribueerde architectuur is de IS1200 schaalbaar tot honderden terabytes. Het systeem is tot zestien systemen te clusteren, waarbij elke clusternode tien miljoen bestanden ondersteunt.

StoredIQ

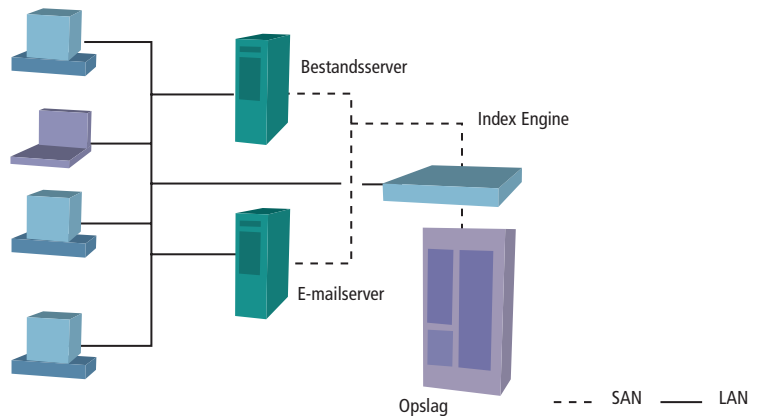
De start-up StoredIQ heeft eind vorig jaar een zogenaamde ICM-toepassing (Information Classification and Management) op de markt gebracht, in de vorm van een apparaat met de weidse naam ICM 5000HC Mini. Hiermee zijn data op bestandsservers te classificeren. Het apparaat is speciaal ontwikkeld voor kleine en middelgrote zorginstellingen en is volgens de firma de enige toepassing op dit gebied die is voorzien van een voorgeconfigureerde module voor de HIPAA/ePHI-regelgeving, die van toepassing is op de gezondheidszorg. Met behulp van deze module kunnen medische specialisten ePHI-content (*Electronic Protected Health Information*) lokaliseren door middel van een combinatie van

persoonlijke en medische content binnen bestanden en e-mail.

Het apparaat is in staat om 175.000 medicijnen en 40.000 medische termen te herkennen, aan de hand waarvan op een intelligente manier kan worden bepaald welk bestand HIPAA-compliant is. StoredIQ kan ook ePHI verhuizen naar een bepaald retention-archief. Het product is door de gebruiker in te stellen voor de meest algemene categorieën van (Amerikaanse) regelgeving.

Njini

Njini, eveneens een start-up, heeft de njiniEngine op de markt gebracht, die alle ongestructureerde bestandsdata analyseert op het moment van creatie. Het systeem bevindt zich 'tussen' de gebruikers en de gedeelde opslagarchitectuur. De njiniEngine houdt automatisch gespecificeerde metadata bij voor alle bestanden. De metadata bevatten het eigenaarschap, de attributen, security en *time stamp* informatie van bestanden. Tevens kunnen met behulp van Njini Helpers uitgebreide metadata van specifieke bestandstypen worden verzameld, bijvoorbeeld van bestanden die verklaringen bevatten als 'confidential', 'copyright', 'merger', 'private', enzovoort. Standaard zijn er Njini Helpers beschikbaar voor Microsoft Office en Adobe pdf, maar met behulp van een *software development kit* zijn ook andere Helpers te ontwikkelen. Toekomstige releases van Njini zullen apparatuur zoals Centera van EMC ondersteunen.



Figuur 5 Index Engine

Index Engine

De firma Index Engine zegt met het gelijknamige apparaat (zie figuur 5) de eerste zoektoepassing voor een opslag-netwerk te bieden die bij het maken van een back-up een index bijhoudt van ongestructureerde data (bestanden en e-mail). Het instapmodel ES-100 kan de indexen opslaan van vier miljoen documenten, de ES-200 van zestien miljoen documenten, waarbij gebruikers snel gegevens kunnen opzoeken binnen een omgeving met de omvang van één terabyte. Het apparaat wordt transparant in het SAN opgenomen en kan in een clusteromgeving tot 64 nodes worden uitgebreid. Het heeft aansluitingen voor zowel SCSI- als fibre channel archiveringstoepassingen. Elk apparaat houdt transparant de indexen bij van tachtig bekende bestandstypen, waarbij snelheden van 3,5 miljoen woorden per seconde worden behaald. Het apparaat decomprimeert en indexeert ook de inhoud van tar-, zip- en gzip-bestanden. Het leest *read-only* bestanden zoals pdf's en scant en indexeert mailboxes (Exchange .pst). Het apparaat ondersteunt de omgevingen van Legato, Veritas en Tivoli Enterprise.

Tools nog in ontwikkeling

Het zal nog wel enkele jaren duren voordat de dataclassificatietools binnen ondernemingen een vaste plaats zullen krijgen. IT-analisten zijn het erover eens dat dataclassificatie nog een lange weg te gaan heeft; in de meeste ondernemingen hebben deze tools nog de status van 'speeltje'. De meer intelligente vormen