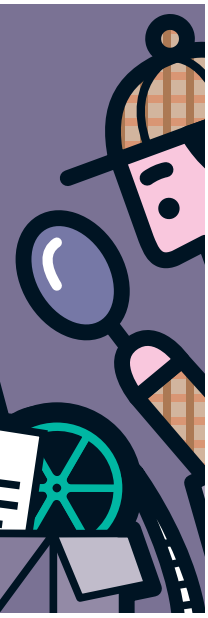


Zoeken en vinden in (on)gestructureerde gegevens

# Uitdaging met twee senario's

Erik Fransen

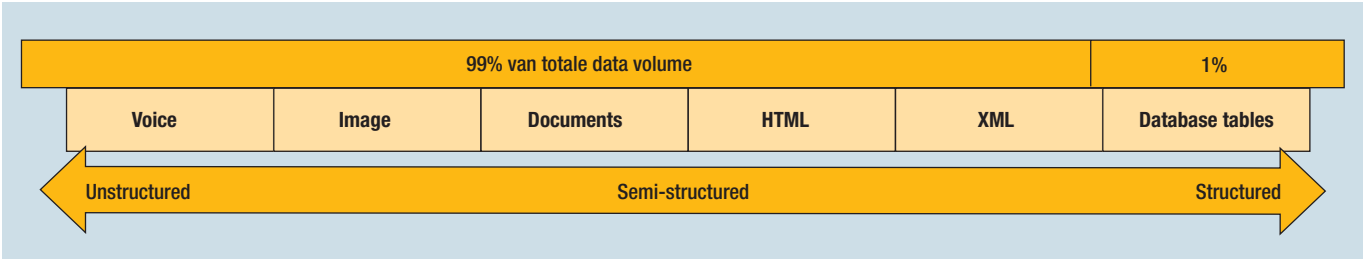
**Iedereen die Informatica heeft gestudeerd of enig andere wetenschap, heeft geleerd om complexe vraagstukken op te delen in kleinere voor de mens beter behapbare vraagstukken, zodat het probleemgebied te overzien is.**

Het vraagstuk wordt geanalyseerd, wat leidt tot een model (systeem dat de werkelijkheid nabootst of tracht te beschrijven) van het probleemgebied. Dit model is een abstractie van het eigenlijke probleemgebied. Daarnaast wordt het model vaak opgesteld met beschikbare oplossingen in het achterhoofd. Modellen worden derhalve vaak op een bepaalde wijze gemaakt, omdat onze oplossingen invulling kunnen geven aan de modellen. Een bekend voorbeeld is de database. Omdat de database, met zijn gerelateerde tabellen en gestructureerde query-taal zo alom vertegenwoordigd is, worden informatievraagstukken meestal geanalyseerd op een zodanige wijze dat het resulterende model (lees: het datamodel) past bij de oplossing, namelijk de database. En aangezien een database in essentie niets meer is dan een vat vol (unieke) rijen met data opgedeeld in (kenmerkende) kolommen, wordt het (data)model ook op deze wijze ontworpen.

Dit betekent dat de oplossingsruimte voor veel informatievraagstukken (en dus ook BI-vraagstukken) beperkt is: de structuur van het te gebruiken model staat vast (entiteiten en relaties) omdat de database vorm moet geven aan de oplossing. In de BI-wereld wordt dit type gegevens overigens gestructureerd genoemd, wat eigenlijk weinig waarde toevoegt en niet BI-specifiek is (gestructureerd: een bepaalde samenhangende structuur hebbend). Elk (goed) model dat wordt gemaakt is gestructureerd (een samenhangend geheel), of het nu wel of niet in een database opgeslagen kan worden. Gestructureerde gegevens zijn dus in de BI-context gegevens die conform het (dimensioneel) datamodel in een database zijn opgeslagen. Voor deze gegevens geldt bovendien dat zowel de metadata (de semantiek) als de data zelf (de content) in de database (en het datamodel) zijn opgeslagen en direct opvraagbaar zijn.

> De tooling moet in staat zijn om de juiste entiteiten te herkennen in de onderliggende content

Is dit dan een probleem? Nee en Ja. Nee omdat databases en de gestructureerde query-taal SQL hun waarde uiteraard bewezen hebben en in een bepaalde informatiebehoefte voorzien. Ja omdat databases voorschrijven dat gegevens in tabellen (via rijen en kolommen en relaties tussen tabellen) opgeslagen moeten worden. Zolang we te maken hebben met gegevens die van oudsher op deze manier gemodelleerd worden is het nog steeds



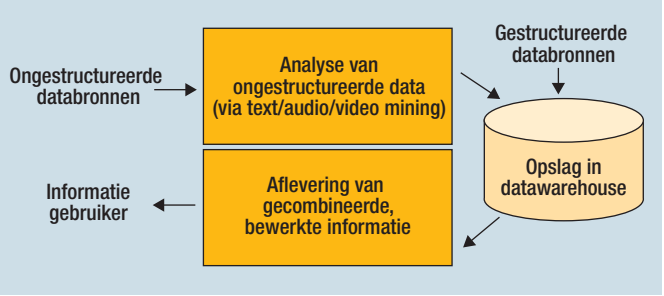

**Afbeelding 1:** Data explosie – 99 procent van de data valt buiten de database.

geen probleem (denk aan financiële administraties, inkoopgegevens, klant- en verkoopgegevens, in feite gewoon kaartenbakken), aangezien de gebruikers lijstjes met (geaggregeerde) rij- en kolomsgewijze (numerieke) informatie nodig hebben.

## Zoekmachine

Echter, indien de gegevens van een heel andere aard zijn, bijvoorbeeld tekstuele documenten, gesproken teksten, videobeelden of geografische gegevens zoals landkaarten, dan is het genoemde datamodel met entiteiten en relaties eigenlijk heel vreemd: waarom zou je documenten die bestaan uit vele pagina's tekst willen proberen te vatten in een paar tabelletjes met relaties ertussen?

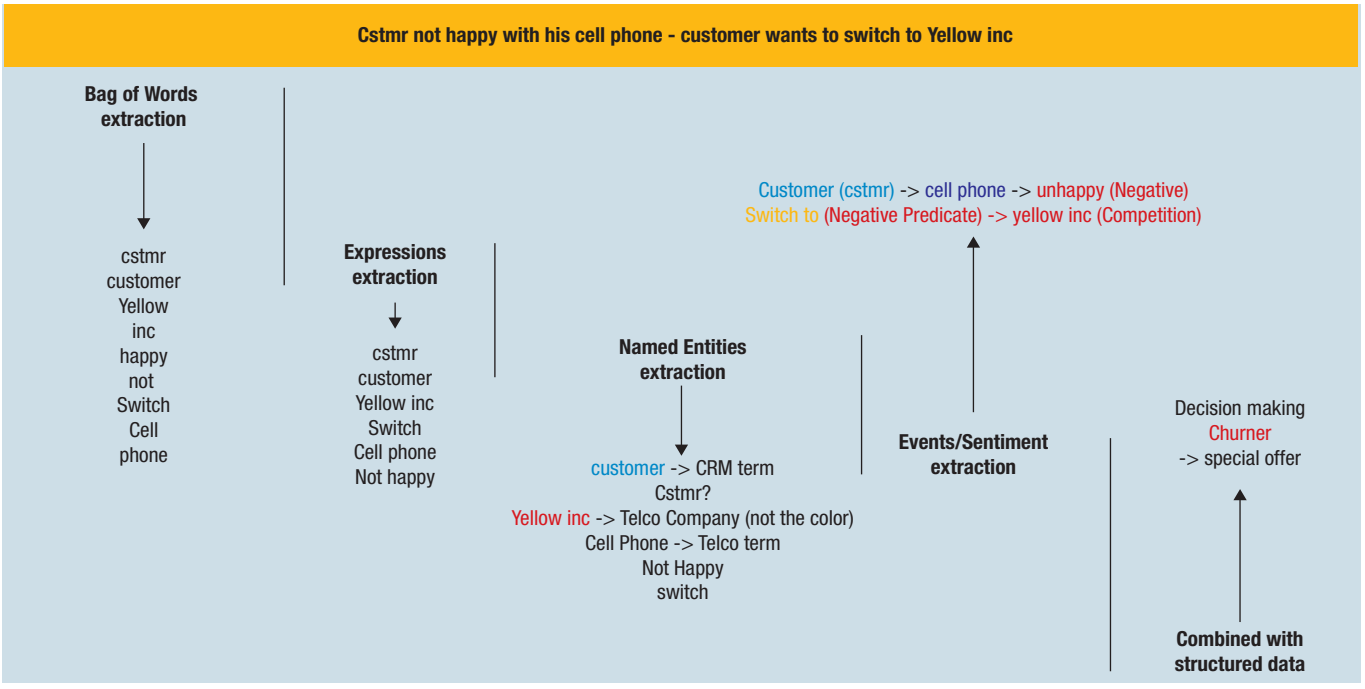
Het antwoord is eigenlijk heel simpel: dat moet je ook niet willen, omdat tekstuele data gewoon in een boek of een Word-document vastgelegd behoren te zijn. Enterprise Content Management-systemen zijn juist ontwikkeld om dit type gegevens op te slaan en te ontsluiten, en dat ontsluiten gebeurt niet met SQL maar met een zoekmachine. In BI-land worden deze gegevens ongestructureerd genoemd. Dit zou impliceren dat een literair boek of een webpagina met bijvoorbeeld reisinformatie niet gestructureerd zou zijn. Dat is natuurlijk onzin, omdat zowel een boek als een reispagina op internet wel degelijk een samenhangend geheel kent (al was het maar een titel, inleiding, conclusie, auteursnaam, de uitgeverij etcetera, in het geval van een boek). Wat we eigenlijk bedoelen met de term ongestructureerde gegevens is dat de gegevens niet zonder meer passen in het datamodel dat hoort bij de database. Een hoofdstuk van een boek (vele pagina's tekst) zou je overigens wel kunnen opslaan in een database (per rij een cel gevuld met de tekst, de kolommen geven het hoofdstuknummer en de hoofdstuktitel aan),



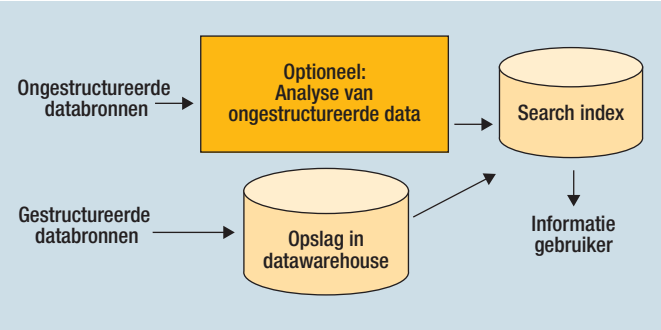

**Afbeelding 2:** Eerste scenario – van ongestructureerd naar gestructureerd.

maar gevoelsmatig past dat niet: een cel in een database is bedoeld om een elementair stukje data op te slaan. De crux zit hier bij de scheiding tussen metadata en de content: de metadata van een boek passen prima in het datamodel van de database, de content zelf hoort er niet in thuis.

Waarom is er dan nu zoveel aandacht voor het opslaan en ontsluiten van ongestructureerde gegevens binnen het vakgebied Business Intelligence? Het antwoord is eenvoudig: er is een gigantische hoeveelheid ongestructureerde data aanwezig die nog niet of nauwelijks door de gemiddelde BI-gebruiker ontsloten wordt, althans niet met de bestaande BI-hulpmiddelen. De meeste schattingen gaan overigens uit van de 80/20-regel: zowel binnen als buiten de bedrijfsgrenzen is 80 procent van de gegevens ongestructureerd van aard, volgens de definitie zoals hiervoor gegeven. Ik vermoed echter dat de 99/1-regel voor de gemiddelde informatiewerker eerder van toepassing is, zie afbeelding 1. De noodzaak voor het ontsluiten van deze gegevens is groot, aangezien het voor mij ondenkbaar is dat alle relevant BI- en stuurinformatie besloten ligt in die ene procent.



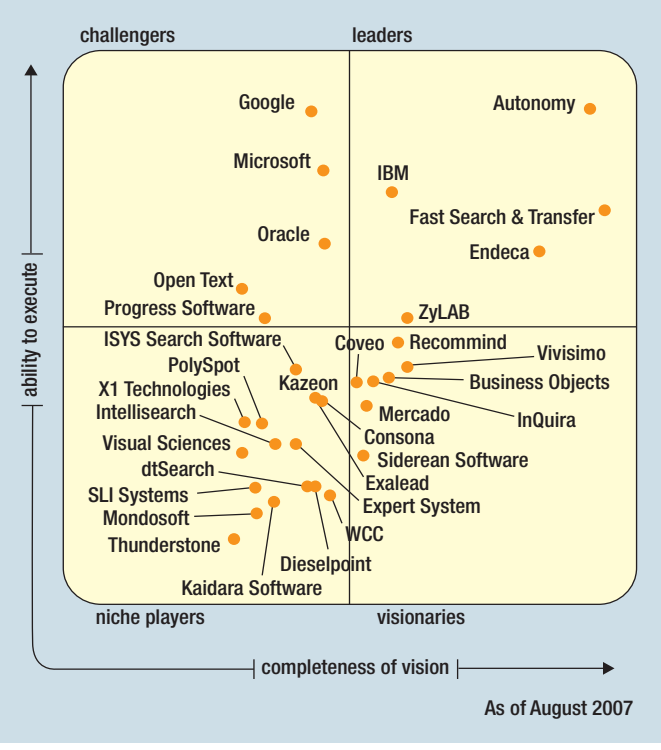

**Afbeelding 3:** SPSS Text Analysis.




**Afbeelding 4:** Tweede scenario – ongestructureerde en gestructureerde data bestaan naast elkaar.

## Data Search: twee scenario's

Uiteraard is de noodzaak voor het ontsluiten van de ongestructureerde gegevens al veel langer onderkend: er zijn vele tientallen volwassen oplossingen voor het opslaan en ontsluiten van ongestructureerde gegevens. Deze oplossingen zitten vaak in de hoek van Enterprise Content Management en enterprise search. Bekende spelers zijn Autonomy, Google, Convera, Fast, Tridion en IBM. Daarnaast zijn er diverse gespecialiseerde bedrijven die opslag en ontsluiting van multimedia-content mogelijk maken. Kern van deze oplossingen is het opslaan van grote hoeveelheden tekstuele of multimediale gegevens, op een zodanige wijze dat deze snel en eenvoudig terug te vinden zijn. Terugvinden geschiedt via zoekmachines die (meestal) op basis van een index content vinden. Daarnaast bieden deze systemen de mogelijkheid om door de content te navigeren: via een thesaurus (woorden/synoniemenlijst) of taxonomie (methodische classificatie/hiërarchische ordening) is content geclassificeerd en vervolgens vindbaar, als dan niet in combinatie met de zoekmachine. Zowel de index als de thesaurus/taxonomie zijn metadata die naast de eigenlijke content worden opgeslagen.




**Afbeelding 5:** Magic Quadrant for Information Access Technology (Bron: Gartner, augustus 2007).

> Textmining geeft structuur aan ongestructureerde gegevens

De uitdaging voor het vakgebied Business Intelligence is om de twee werelden van gestructureerde gegevens en ongestructureerde gegevens bij elkaar te brengen: de BI-eindgebruiker wordt dan in staat gesteld om op een eenduidige wijze beide soorten gegevens te ontsluiten, waarbij enerzijds de bekende technieken uit de databasewereld gebruikt worden (datamodellen en SQL) en anderzijds indexerings-, textmining en zoektechnieken uit de Enterprise Search- en Content Management-wereld worden gehanteerd.

We onderscheiden hiertoe twee data search scenario's.

1. Van ongestructureerde data naar gestructureerd.
a. Via text/video/spraak-mining.
b. Zoeken en Vinden met 'traditionele' BI: SQL, OLAP.
2. Ongestructureerde en gestructureerde data bestaan naast elkaar.
a. Search engine indexeert (ook) gestructureerde data.
b. Text/video/spraak-mining is optioneel.
c. Zoeken en Vinden met search engine en/of SQL/OLAP.

Het eerste scenario is gericht op automatisch genereren van structuur in de ongestructureerde gegevens, zie afbeelding 2. Concreet betekent dit dat met tooling een datamodel wordt gegenereerd uit de content waarin slechts verwijzingen zijn opgenomen naar de feitelijke content. Het gegenereerde datamodel kent dan bijvoorbeeld dimensies als genre, thema, hoofdstuk, auteur, trefwoord(en) en een of meerdere feitentabellen met daarin verwijzingen naar specifieke content (een hoofdstuk, het hele boek, een samenvatting, tekeningen etcetera). Dit datamodel kan dan worden gekoppeld aan een bestaand BI-datamodel (dimensioneel model met gestructureerde data zoals boekenverkopen, lezersinformatie, profielen etcetera) waardoor bijvoorbeeld bezoekers van een boekenwebsite niet alleen een boek kunnen vinden dat binnen hun profiel past, maar tegelijkertijd ook de inhoud (van delen van het boek) kunnen lezen.
De uitdaging van dit scenario ligt bij het genereren van het datamodel: de tooling moet in staat zijn om de juiste entiteiten (dimensies en feiten) te herkennen in de onderliggende content. Moderne content management en enterprise search tools worden hier steeds sterker in door het toepassen van textmining-technieken (onder andere Autonomy, Cunvera en IBM). Het toekennen

van structuur gebeurt echter vaak nog met de hand: per document (of documentonderdeel, bijvoorbeeld de titel, de inleiding, conclusie, auteur etcetera) worden tags toegekend, die dan eventueel onderdeel zijn van een thesaurus. Op die manier wordt het mogelijk om met synoniemen te zoeken (auteur versus schrijver). Dit handmatig toekennen van tags is arbeidsintensief, vandaar dat binnen dit scenario veelal gebruikt wordt gemaakt van technieken als textmining en videomining.
Textmining geeft structuur aan ongestructureerde gegevens. Bekende technieken zijn Feature Expression Extraction waarmee entiteiten in de gegevens worden herkend, ER-modelling waarmee relaties tussen entiteiten worden herkend, maar ook classificatie (toekennen van inhoudelijke tags aan een stuk tekst waarmee het wordt getypeerd), genereren van samenvattingen en automatisch vertalen. Overigens kennen textmining-systemen die gebaseerd zijn op linguïstische analyse vaak een leercurve; het systeem moet gevoed worden met voorbeelden waarmee het leert dat tags bij een bepaald stuk tekst horen. Hoe meer het systeem wordt gevoed, hoe beter het kan worden. Dit in tegenstelling tot textmining gebaseerd op statistische analyse, waarbij het systeem vooral kijkt naar frequentie en distributie van termen.

Het voorbeeld in afbeelding 3 van SPSS Text Analysis geeft aan welke stappen worden doorlopen om ongestructureerde gegevens (in dit geval de tekst die een call center agent heeft ingevoerd naar aanleiding van een gesprek met een ontevreden klant) structuur te geven. Op basis van de tekst worden allereerst de woorden met de meeste informatie herkend, vervolgens worden woorden die bij elkaar horen herkend (bijvoorbeeld *not happy*, daarna wordt de relatie gelegd met entiteiten zoals die reeds bekend zijn bij de gebruikers (Yellow Inc is een Telco Company). De laatste stap in het textmining-proces bestaat uit het relateren van entiteiten en aanduiden van het belang van de relatie. De output van dit textmining-proces (de entiteiten en relaties en hun waarden) wordt opgeslagen in een dimensioneel datamodel. Deze informatie wordt gecombineerd met profielinformatie in het datawarehouse: het feit dat de klant wil gaan

| Tool | Product | Propositie |
|---|---|---|
| IBM | Unstructured Information Modeler | Automatisch genereren en aanpassen van taxonomieën van ongestructureerde gegevens. |
| | IBM Enterprise Content Management Search and Discovery solutions (OmniFind) | IBM OmniFind biedt mogelijkheden voor 'enterprise search': het doorzoeken van alle informatiebronnen van een bedrijf. |
| Microsoft | Office SharePoint Server | Office SharePoint Server biedt standaard zoekfuncties voor het zoeken in bestanden, websites, SharePoint-sites, openbare Exchange-mappen en in Lotus Notes-databases. |
| | Fast | Microsoft Fast's Enterprise Search platform levert een set van tools voor het herkennen van talen en synoniemen, de thesaurus maakt het mogelijk om zelf termen toe te voegen. Add-on's geven de mogelijkheid tot data cleansing en multimedia mining. |
| Oracle | Oracle Secure Enterprice Search | Oracle Secure Enterprise Search (SES) biedt een vergelijkbare gebruikersinterface voor het zoeken op internet, maar zorgt tevens voor een veilige toegang tot alle gegevensbronnen binnen organisatie-websites, file servers, Content Management Systemen, ERP- en CRM-systemen, BI-systemen en databases. |
| BO | Text Analysis | BusinessObjects Text Analysis leest tekst uit 30 talen, onttrekt sleutelinformatie zodat gegevens uit tekstdocumenten gebruikt kunnen worden als bron voor data-integratie of BI, ontdekken van 'verstopte' informatie in CRM-systemen, internet, e-mails en andere tekstuele bronnen. |
| | Business Objects Intelligent Search (Voormalig Inxight) | De mogelijkheid voor het extraheren van entiteiten en feiten uit ongestructureerde gegevens. Zoeken binnen alle bronnen vanuit een single secure search box. Resultaten zijn geordend naar relevantie en geclusterd naar mensen, bedrijven en andere concepten. |
| | Intelligent Search (for Google Desktop) Voorheen Inxight Search Extender for Google Desktop | Intelligent Search clustert de resultaten al tijdens het zoekproces en geeft de mogelijkheid tot filteren van de zoekresultaten op mensen, bedrijven, producten en andere informatie. |
| Cognos | Cognos 8 Go! Search | Zoeken naar rapporten in IBM Cognos 8 BI. |
| SAS | SAS Text Miner | SAS Text Miner bestaat uit een aantal tools om kennis in documenten te ontdekken en onttrekken. Het maakt het mogelijk om documenten te classificeren, categoriseren en relaties te ontdekken tussen deze documenten.<br>Niet als individueel product te gebruiken. SAS Text Miner draait op een SAS implementatie en SAS Enterprise Miner. Zoeken in en analyseren van ongestructureerde gegevens. |

**Afbeelding 6:** Tabel.

overstappen naar Yellow Inc omdat hij/zij ontevreden over zijn/haar mobiele telefoon betekent dat de klant als 'churner' wordt aangeduid, waardoor een gerichte actie kan worden opgestart door de marketingafdeling.

Zoals opgemerkt zijn er diverse tools beschikbaar die in staat zijn om ongestructureerde gegevens op te slaan en te ontsluiten (Enterprise Content Management- en Enterprise Search-tools). In aanvulling daarop zijn de textmining-tools in staat om structuren te genereren uit de teksten, waarna deze gerelateerd kunnen worden aan gestructureerde gegevens in het datawarehouse. De grote uitdaging zal zijn om de twee werelden te verenigen in een alomvattende data- en informatiearchitectuur. Momenteel zijn er diverse architecturen beschikbaar die invulling geven aan deze vereniging. De bekendste zijn het Uniform Information Management Architecture (UIMA) van IBM en DW 2.0 van Bill Inmon. Het UIMA-model is uitvoerig beschreven in DB/M2 2008 en is een concrete verzameling van (open source) functies waarmee de gestructureerde en ongestructureerde wereld met elkaar verbonden worden. DW 2.0 van Inmon (www.inmoncif.com) is vooral een architectuurraamwerk, waarin de ongestructureerde component is geïntegreerd, op de wijze zoals in dit artikel geschetst. Aan de BI-architecten nu de schone taak om binnen de context van UIMA dan wel DW 2.0 invulling te gaan geven aan het verbinden van gestructureerde en ongestructureerde data.

> De output van het textmining-proces wordt opgeslagen in een dimensioneel datamodel

Het tweede scenario (zie afbeelding 4) integreert ongestructureerde en gestructureerde data op een alternatieve wijze (dit is overigens ook het meest gebruikte scenario op dit moment): een zoekmachine wordt ingezet, op basis van een index op zowel ongestructureerde als gestructureerde data, waardoor niet alleen teksten worden ontsloten, maar ook database-gegevens zoals metadata en content. Tevens worden BI-rapportages geïndexeerd, waardoor rapporten via zoektermen teruggevonden worden.

Het indexeren van BI-rapportages lijkt het meest zinvol, aangezien daardoor heel gericht gegevens beschikbaar komen voor de gebruiker (bijvoorbeeld een rapport met de omzet voor een specifieke klant via de zoektermen 'JANSSEN' en 'OMZET', waarbij het meest recente rapport als eerste wordt getoond en het systeem om kan gaan met typefouten, synoniemen en vreemde talen). In aanvulling daarop worden persberichten, relevante e-mailuitwisselingen of accountplan(nen) van deze klant(en) getoond. Op het moment dat het klantnummer van Janssen bekend is, kan gericht met dit nummer gezocht worden, direct via de zoekmachine.

Een alternatief is de mogelijkheid om tijdens het opstellen van een BI-rapport de dimensies doorzoekbaar te maken via een index: grote dimensies zoals klant, product of leverancier zijn dan sneller te doorgronden waardoor gerichte rapportages voor selecties uit deze dimensies sneller te genereren zijn.
Optioneel kan dit scenario als een uitbreiding van scenario 1 worden gezien: de onttrokken gestructureerde gegevens uit de ongestructureerde bronnen worden geïndexeerd en daardoor ook terugvindbaar met de zoekmachine. De gegevens (metadata en content) zijn vervolgens via SQL, OLAP en/of via een zoekmachine te ontsluiten.

## Tools voor Data Search

Het Magic Quadrant for Information Access Technology (augustus 2007) van Gartner kent ook voor de gemiddelde BI-professionals een aantal bekende leveranciers: Microsoft (eigenaar van Fast sinds februari 2008), Oracle, IBM, maar ook Business Objects zijn opgenomen omdat zij oplossingen leveren die al invulling geven aan genoemde zoek-, indexerings- en textmining-functionaliteiten. Zij zouden dus bij uitstek in staat moeten zijn om tools te leveren die invulling geven aan zowel scenario 1 als 2. Zowel Oracle (met Secure Enterprise Search), IBM (met OmniFind), Microsoft (met Fast en SharePoint) alsook Business Objects (met Text Analysis en Intelligent Search) bieden hiervoor diverse oplossingen. Zie voor een uitgebreider beschrijving afbeelding 6.
Endeca, Autonomy, OpenText en Google zijn leveranciers die hun wortels in Content Management en/of Enterprise Search hebben: de integratie met gestructureerde gegevens ligt voor de hand en er zijn diverse oplossingen beschikbaar: OpenText levert Livelink ECM/Business Intelligence waarmee databases en OLAP-kubussen bevraagd kunnen worden vanuit de Content Management-omgeving. Cognos 8 Go! Search (Google Enterprise Search) geeft gebruikers toegang tot functionaliteit en onderliggende content in de systemen van Autonomy, Fast, IBM en Google.

## Nog 99 procent te gaan!

We verwachten dat de komende jaren de trend van het verenigen van gestructureerde en ongestructureerde data zal leiden tot producten die niet meer te classificeren zijn als een BI- of een ECM/Search-product, zoals dat nu nog wel het geval is. De integratie zal verder versnellen doordat de ontwikkeling van BI-, ECM- en search-tools zal plaatsvinden binnen de visie en randvoorwaarden van architectuurmodellen zoals DW 2.0 en UIMA. Voor de BI-professional betekent dit dat hij/zij zich ook, veel meer dan nu het geval is, zal moeten gaan richten op het implementeren van de geschetste scenario's, tools en architectuurmodellen. We hebben nog 99 procent van de beschikbare data te ontginnen voor onze eindgebruikers!

**Erik Fransen** is senior business consultant bij Centennium BI Expertisehuis.
Met dank aan Philip du Maine, BI consultant bij Centennium BI Expertisehuis.