Integrale database voor gestructureerde en ongestructureerde informatie

Meer rendement op informatie

Anja van der Lans en Peter van Til

Het samenbrengen van gestructureerde en ongestructureerde informatie krijgt steeds meer aandacht omdat bedrijven zich bewust worden van de potentiële meerwaarde van de combinatie. Het wordt mogelijk om veel meer aspecten van een bedrijfsprobleem te belichten en bijeen te brengen. De informatiewerker heeft een completer beeld van de problematiek en kan een beter afgewogen besluit nemen.

De vraag is hoe de architectuur moet worden ingericht om beide soorten data efficiënt bij elkaar te kunnen brengen. De probleemstelling is de vraag welke omgeving leidend wordt voor de opslag van informatie in de breedste zin van het woord. Wordt dat Business Intelligence (zoals Bill Inmon stelde op het BI Event van mei 2010), de content management omgeving of is er behoefte aan een geheel nieuwe invulling van de omgeving?
In dit artikel worden de voor- en nadelen van de verschillende mogelijkheden beschreven. Maar voordat deze alternatieven verder worden uitgewerkt is het goed om een beeld te schetsen van het begrip 'integrale database'.
De eerste die een aanzet heeft gegeven tot een integrale database is Bill Inmon. In zijn DW2.0 benadering werd voor het eerst het belang ingezien van het vastleggen van ongestructureerde gegevens naast gestructureerde informatie. In DW2.0 gebeurt het vastleggen van deze gegevens op basis van de resultaten van tekstanalyses, entity extraction, thesauri enzovoort. De resultaten van deze inspanningen moeten dan leiden tot gestructureerde 'ongestructureerde' gegevens die volgens de principes van de gestructureerde gegevens in databases konden worden vastgelegd. Met andere woorden; de ongestructureerde gegevens worden zodanig bewerkt dat ze passen in de databasestructuren van de gestructureerde data. Bill Inmon stelde zich dat in zijn DW2.0 omgeving voor zoals is aangegeven in afbeelding 1.

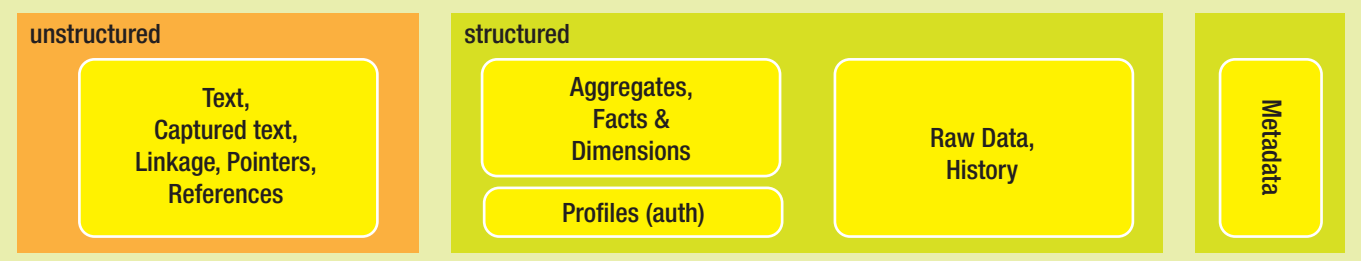
De vraag is of deze voorstelling van een integraal datawarehouse voldoet aan de eisen die hedendaagse kenniswerkers stellen aan het opslaan van gegevens. Waarschijnlijk is dit niet het geval en de eerste tekenen hiervan worden zichtbaar in het veranderende denken over deze benadering.
De ultieme informatieomgeving bevat alle informatie die ook in de benadering van Bill Inmon wordt vastgelegd, maar legt niet de beperkingen op die horen bij DW2.0. Alle soorten gegevens/informatie worden daarbij gecombineerd tot een centrale index (en dus niet langer in twee afzonderlijke indexen) die het mogelijk maakt om alle gegevens op een geuniformeerde wijze te benaderen, om die reden ook vaak 'unified index' genoemd.

Schemaloos

Een poging om een stap in deze richting te zetten is zichtbaar in de manier waarop de laatste tijd Google en andere zoekmachines indexen opbouwen, namelijk met behulp van schemaloze databases. Dit soort databases is eenvoudig in gebruik, geschikt voor de opslag van minder gestructureerde data en biedt ook bij de opslag van Terabytes aan informatie nog een goede performance.
Een uiteenzetting van deze opslagmethode werd door Bas Peters in DB/M 4, juni 2010 gegeven ('Google datastore: mogelijkheden en beperkingen'). Schemaloze databases werken op basis van sleutel/waardeparen waarin alle informatie (gestructureerd en

Zoeken

De problemen waar Inmon het in zijn textual datawarehouse voorstel over heeft als 'issue of terminology' of 'issue of date' zijn in enterprise search al een aantal jaren terug 'getackeld'. Door middel van entity extraction kunnen op een eenvoudige manier woorden in een bepaalde context aan elkaar worden gerelateerd en als zodanig onderscheiden van andere woorden, om vervolgens op het concept of de entiteit in al haar hoedanigheden te kunnen zoeken. Dan zijn twintig varianten om het begrip 'broken bone' te omschrijven geen probleem meer. De zoekmachine herkent het begrip, legt de relatie met gelijksoortige termen onder een 'common format' als fracturen en interpreteert deze als gelijke waarde voor de resultatenset.



Afbeelding 1: Integraal datawarehouse in DW2.0.

ongestructureerd) kan worden vastgelegd. De Google datastore lijkt een eerste aanzet tot een compleet nieuwe manier van het opslaan van informatie waarmee een combinatie van gestructureerde en ongestructureerde informatie mogelijk wordt. Het concept zal zich echter moeten bewijzen in de combinatie van Business Intelligence en content management.

We keren terug naar de probleemstelling: wordt Business Intelligence leidend, of is het beter om uit te gaan van de content management omgeving, of hebben we iets totaal nieuws nodig (zodat we 'verlost' zijn van ontwerpbeslissingen van het verleden)?

Een gelijksoortige keuzemogelijkheid heeft zich enige tijd geleden ook voorgedaan binnen het Business Intelligence vakgebied en wel bij het modelleren van een enterprise datawarehouse. Daarin waren twee stromingen te identificeren; de relationele en de dimensionele. Dan Linstedt legde van beide stromingen de sterke en zwakke punten vast en kwam met een geheel nieuwe benadering, de Data Vault. Deze werkwijze had geen last van ontwerpbeslissingen uit het verleden en zorgde voor een revolutionaire oplossing van het datamodelleringsprobleem. Deze methode wordt intussen breed ingezet in datawarehouse oplossingen.

Een zelfde soort probleem doet zich nu dus voor bij het creëren van een enterprise information environment, de data opslagcomponent voor de combinatie van gestructureerde en ongestructureerde gegevens.

In afbeelding 3 is het totale speelveld van gestructureerde en ongestructureerde data weergegeven: een omgeving die alle processen binnen de organisatie kan ondersteunen met de benodigde en gewenste informatie; een omgeving die zorgt voor Business Intelligence, content intelligence en search intelligence. Business Intelligence is daarbij het onderdeel dat de intelligence vooral baseert op de gestructureerde data, content intelligence gaat uit van de ongestructureerde data. Search intelligence moet beide onderdelen kunnen combineren. Business Intelligence integreert de gegevens uit de bronsystemen die nodig zijn voor het samenstellen van beslissingondersteunende informatie.

Deze omgeving is nodig omdat de bronsystemen vaak niet alle historie vast blijven houden die de basis vormen voor de diverse modellen om voorspellend te kunnen zijn. De content omgeving houdt alle documenten vast die in een organisatie worden beheerd. De vraag is nu of de content kan worden geïntegreerd in de BI-omgeving of dat de BI-omgeving moet worden geïntegreerd met de content of dat de Business Intelligence en de content gezamenlijk moeten worden geoptimaliseerd om search intelligence mogelijk te maken. Alle drie de alternatieven hebben tot doel om meer rendement uit informatie te creëren. Wie de informatie op een betere manier kan inzetten heeft namelijk een concurrentieel voordeel. Hieronder volgen de voor- en nadelen van ieder alternatief.

BI inclusief content intelligence

Het integreren van ongestructureerde en semigestructureerde gegevens in de gestructureerde omgeving is niet zomaar gedaan. Vanuit de ongestructureerde omgeving moet met behulp van text mining technieken, thesauri en sentimentanalyses structuur worden gebracht in de ongestructureerde teksten. Via allerlei ingewikkelde stappen belandt de informatie ten slotte in de gestructureerde omgeving waar op basis van overeenkomstige sleutelvelden zaken kunnen worden geïntegreerd. Indien de informatie uitsluitend in de ongestructureerde omgeving aanwezig blijkt, dan zal het wel geladen kunnen worden in de gestructureerde omgeving, maar zal het niet tot extra informatie leiden. Er wordt immers vaak geredeneerd vanuit de gestructureerde informatie, waarbij de ongestructureerde gegevens als verrijking dienen.

Bill Inmon geeft aan dat hij dit type textual ETL of textual datawarehouse voor de toekomst als mogelijke oplossing ziet voor het

Afbeelding 2: Ultieme opslag voor gestructureerde en ongestructureerde gegevens.

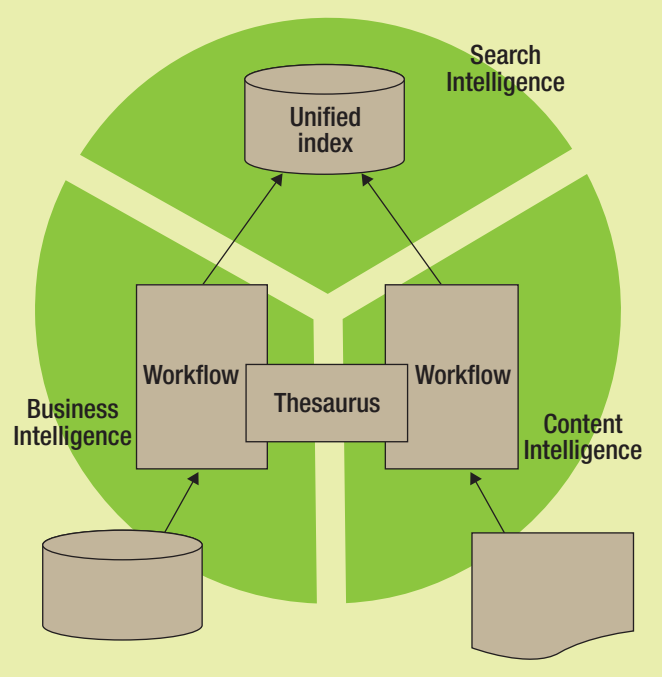
combineren van beide typen informatie. Toch zijn daarbij twee afzonderlijke containers voor opslag noodzakelijk met alle problemen van dien. De door hem genoemde problemen rondom het gebruik van terminologie en standaardisatie van entiteiten als datum of de schrijfwijze van namen van personen en instellingen zorgen voor veel onvolkomenheden.

Content intelligence inclusief BI

Het onderbrengen van Business Intelligence in de content omgeving zou kunnen door de rapportages uit de BI-omgeving te bewaren en te onderwerpen aan content technieken zoals text mining, thesauri, en dergelijke. Er wordt op deze manier van gestructureerde data, ongestructureerde tekst gemaakt en op basis van deze nieuwe ongestructureerde gegevens wordt er met behulp van content technieken weer structuur aangebracht. Het enige voordeel is dat de content technieken op beide omgevingen worden toegepast. De wenselijkheid van dit alternatief laat echter nogal te wensen over. De analysemogelijkheden die in de BI-omgeving mogelijk zijn worden door het gebruik van de statische rapporten in een ongestructureerde omgeving volledig teniet gedaan. Het enige wat mogelijk blijft is de combinatie van gestructureerde en ongestructureerde gegevens. Aan de gebruikswaarde van de oplossing wordt volledig voorbij gegaan.

Search intelligence als integrator

In deze opzet worden beide omgevingen, zowel Business Intelligence als content intelligence, gelijkwaardig beschouwd. Het is dus niet zo dat de ene omgeving wordt gebruikt ter verrijking van de andere. Beide omgevingen zijn gelijk. Maar hoe krijgen we deze ogenschijnlijk verschillende omgevingen gelijkwaardig? Gelijkwaardig betekent dat de brongegevens in de oplossing aanwezig moeten zijn. Dit betekent dat niet alleen de resultaten van de content technieken moeten worden opgenomen, maar ook de informatie waarop deze technieken zijn gebaseerd zoals een thesaurus en dergelijke. Dit betekent ook dat we bij de integratie van gestructureerde en ongestructureerde data niet alleen de gegevens ter beschikking moeten hebben maar ook de technieken die deze gegevens kunnen relateren aan de thesauri en andere hulpmiddelen. Het gaat dus om data én functionaliteit. Het moet immers mogelijk zijn om product 101 (damesfiets Batavus) te koppelen aan ongestructureerde informatie over damesfietsen van Batavus zonder dat de ongestructureerde gegevens de term 'product 101' bevatten. Dit kan met een 'unified index'.



Afbeelding 3: Informatieomgeving met thesaurus als verbindende factor.

De 'unified' index heeft een speciale opbouw waardoor de searchfunctie optimaal wordt ondersteund. De gegevens vanuit de BI-omgeving én vanuit de content intelligence-omgeving worden door speciale workflows geschikt gemaakt voor deze index. De thesaurus is in beide workflows actief waardoor alle synoniemen, homoniemen, spellingsvormen en hiërarchieën op de juiste manier kunnen worden vormgegeven en worden vastgelegd in de 'unified' index, zie afbeelding 3. De structuur van de index is zodanig dat in de structuur de databases vanuit de gestructureerde omgeving (BI en OLTP) kunnen worden opgenomen en worden geïntegreerd met de informatie uit de ongestructureerde omgeving (content intelligence en content management systemen – OLCP). Het is dus mogelijk om met één zoekvraag de gegevens uit beide bronomgevingen te selecteren en op basis hiervan de informatie weer te geven in een applicatie.

Een nieuwe structuur voor het opslaan van gegevens maakt het mogelijk om de gegevens geschikt te maken voor nieuwe informatiewensen. Het brengt nieuwe mogelijkheden waardoor onmogelijkheden nieuwe mogelijkheden blijken. Enterprise Information Management is niet langer alleen een begrip, maar kent nu al concrete toepassingen. Dit is weer een stap dichter bij de juiste informatie op de juiste plaats op het juiste moment; weer een stap op weg naar meer 'rendement op informatie'.

Anja van der Lans (anja.vanderlans@vlc.nl) en **Peter van Til** (peter.vantil@vlc.nl) zijn beiden senior EIM-consultant bij VLC.

Volgende stap

Het idee van Inmon wordt verkocht als de volgende fase van Business Intelligence maar zou beter uitgewerkt kunnen worden tot een volgende stap op weg naar Enterprise Information Management. Hierbij gaan we uit van een volledige fusie tussen Business Intelligence, content management en enterprise search. De Business Intelligence wereld zou er goed aan doen om over de schutting van haar eigen vakgebied heen te kijken. Dan blijkt dat er al bloemen opkomen in de EIM-weide.