

# The Textual Data Warehouse

## Interactieve workshop met Bill Inmon – “The Father of Data Warehousing”

19 mei 2010 – Intres Hoevelaken



### **Dat een tekstgebaseerde datawarehouse veel voordelen biedt, staat buiten kijf. Maar hoe worden tekstgebaseerde datawarehouses eigenlijk gecreëerd?**

In deze ééndaagse workshop bespreekt Bill Inmon wat er allemaal komt kijken bij het bouwen van een tekstgebaseerde datawarehouse en hoe men eventuele obstakels te lijf kan gaan. Natuurlijk gaat Bill verder in op onderwerpen als datumstandaardisatie en de omzetting van tekst naar cijfers. Maar ook de grote issues komen aan bod – denk aan terminologie-kwesties en de logische onderverdeling van bestanden.

Een greep uit het programma:

- veelvoorkomende issues bij het bouwen van een tekstgebaseerde datawarehouse,
- terminologie en logische onderverdeling van tekst,
- invoer van spreadsheets,
- OCR,
- definiëren van scheidingstekens,
- combineren van tekstuele data met klassieke gestructureerde data.

Jarenlang hebben bedrijven beslissingen genomen op basis van gegevens die werden aangetroffen in transactiegebaseerde systemen. Transactiegeoriënteerde gegevens passen goed bij gewone databasemanagement-systemen: deze systemen structureren gegevens immers op een repetitieve manier. Daarbij heeft elke datavindplaats dezelfde structuur als elke andere vindplaats in een tabel. Maar er is nog een andere belangrijke bron van gegevens in het bedrijf: de informatie die wordt aangetroffen in de vorm van tekst. Een gemiddeld bedrijf bezit tekst in alle soorten en maten: e-mails, spreadsheets, contracten, garanties, medische informatie, noem maar op. Maar omdat tekst niet repetitief is, past het niet goed bij gewone databasemanagementsystemen.

Tegenwoordig is er tekstuele ETL en bestaat de mogelijkheid databases en datawarehouses te bouwen die tekstuele informatie bevatten. Als tekstuele gegevens zodanig getransformeerd kunnen worden dat de tekst in een standaard-DBMS past, ontstaan er legio mogelijkheden op het gebied van analyse en besluitvorming.

In deze ééndaagse workshop gaan de deelnemers op zoek naar de ingrediënten die nodig zijn voor het creëren van het tekstuele, ongestructureerde datawarehouse. Het ochtendgedeelte bestaat uit een lezing en het middagedeelte uit een interactieve workshop waarin de databases voor het datawarehouse worden opgebouwd uit tekst. Het seminar wordt volledig in het Engels gepresenteerd.

### **Bestemd voor ú**

Deze dag is bedoeld voor mensen die geïnteresseerd zijn in de manier waarop tekst kan worden ingezet bij het produceren van een analytische database. Data-architecten, projectmanagers, technici en mensen uit het bedrijfsleven zijn van harte welkom om deel te nemen aan de workshop. Reserveer daarom in uw agenda: 19 mei 2010!

## **PROGRAMMA**

**Woensdag 19 mei 2010**

### **Morning Presentation**

#### **An Introduction To Unstructured Data**

This Powerpoint presentation shows what the issues of unstructured data are, what the possibilities, and what opportunities there are. The attendee is introduced to textual ETL and the creation of the unstructured data base. The issues of integration are discussed amply.

#### **Issues of Textual Integration**

There are many issues in the integration of unstructured data and the transformation of unstructured data to an analytical relational data base. Some of the issues are:

- the issue of terminology
- the issue of logical subdivisions of data with a document
- the issues of clustering
- the issues of proximity
- the issues of filtering unnecessary data

These and many other issues are discussed in the context of reading in raw text and creating a viable analytical data base.

### **Afternoon Workshop**

In the afternoon Textual ETL will be run producing a wide variety of data bases/data warehouses using many of the features of Textual ETL. The attendees will observe and participate in the transformation of text into a data base ready for analytic processing.

The workshop begins by examining some textual data. A strategy for capturing and organizing the text is discussed. Then the workshop continues with several types of processing that are done dynamically, under the purview of the attendees. Some of the types of processing that are done include:

- document metadata capture
- document fracturing
- named value indexing
- simple indexing
- semi structured indexing
- merged indexing

Depending on the textual data that has been selected, some or all of these kinds of indexes will be chosen and created. Many of the features of Textual ETL will be used during the workshop. At the end of the workshop the attendees may process some documents that they have brought to the workshop to see what the processing of textual ETL can do.

# The Textual Data Warehouse

## INFORMATIE

De workshop vindt plaats op woensdag 19 mei 2010. Het programma start om 09.30 uur en duurt tot 17.00 uur. Registratie is mogelijk vanaf 08.30 uur.

### PLAATS

Intres  
Koninginneweg 1  
3871 VL Hoevelaken  
Telefoon: 033 - 253 24 85  
Fax: 033 - 253 22 48.

### AANMELDEN

Aanmelden kan op verschillende manieren:

- **online: op [www.arrayseminars.nl](http://www.arrayseminars.nl)**
- per post: stuur de aanmeldingskaart ongefrankeerd naar  
Array Seminars  
Antwoordnummer 10235  
2400 VB ALPHEN AAN DEN RIJN
- per fax: stuur de aanmeldingskaart naar  
0172-424381.

Na ontvangst van uw aanmelding krijgt u de bevestiging, factuur en routebeschrijving toegestuurd.

### KOSTEN

De kosten voor deelname aan de workshop bedragen € 770,- (excl. BTW) per persoon. Documentatie, lunch en koffie zijn inbegrepen. Abonnees van **Business Process Magazine, Database Magazine en Optimize** komen in aanmerking voor € 75,- korting op de toegangsprijs. Medewerkers van universiteiten en rijkserkende HBO-instellingen ontvangen 20% korting op de deelnemersprijs. Deze korting kan niet

worden gecombineerd met abonneekorting. Werkt u bij een gemeente of provincie? Dan kunt u BTW terugvorderen via het BTW compensatiefonds.

### AANTREKKELIJKE KORTINGEN

Als u zich uiterlijk vijf weken voor het evenement registreert, wordt u beloond met 10% **vroegboekorting**. Meldt u op dezelfde dag **meerdere personen van één bedrijf** aan voor hetzelfde evenement, dan geldt vanaf drie deelnemers een aantrekkelijke korting. De derde en vierde persoon ontvangen ieder 20 % korting, de vijfde en elk volgende deelnemer ontvangen 30% korting (de deelnemers dienen op dezelfde factuur te staan).

### ANNULEREN

Annuleren dient uitsluitend schriftelijk te geschieden. U kunt annuleren tot drie weken voor het evenement plaatsvindt. Er wordt echter wel € 75,- (excl. BTW) administratiekosten in rekening gebracht. Annuleren is niet meer mogelijk vanaf drie weken voordat het evenement plaatsvindt. Vervanging door een ander dan de aangemelde persoon is te allen tijde mogelijk.

### MEER INFORMATIE

Alle informatie over deze workshop en een complete agenda van de evenementen vindt u op [www.arrayseminars.nl](http://www.arrayseminars.nl). U kunt zich hier ook online registreren onder vermelding van de volledige foldercode die onderaan de aanmeldingskaart vermeld staat. Voor meer informatie over het programma kunt u contact opnemen met Array Seminars, telefoonnummer 0172-469030. Voor meer informatie over de registratie kunt u bellen met Quatuor, telefoonnummer 036-5409111.



Bill Inmon is oprichter, voorzitter en de belangrijkste technoloog van Inmon Data Systems. Hij heeft een uitgebreide ondernemersachtergrond. Voorafgaand aan het oprichten van Inmon Data Systems, heeft hij in 1991 Prism Solutions opgericht en naar de beurs gebracht. In 1995 richtte hij 'Pine Cone Systems' op, later hernoemd tot 'Ambeo'. Wereldwijd bekend als 'The Father of Data Warehousing', heeft Bill adviezen verstrekt aan veel 'Fortune 1000'-ondernemingen en vooraanstaande IT-managers op het gebied van Datawarehousing, Business Intelligence en Database Management. Bill kondigde onlangs DW 2.0 aan, de architectuur van de volgende generatie datawarehousing. Bill Inmon is een belangrijke spreker bij diverse grote technologie-instituten en IT-congressen en hij staat bekend om zijn technologie-seminars. Ook heeft hij 46 boeken geschreven, waarvan veel in 9 talen zijn vertaald. Hij heeft ruim 650 artikelen geschreven over verschillende onderwerpen in de technologie. Zijn artikelen zijn gepubliceerd in diverse vooraanstaande, gerenommeerde vakbladen, zoals 'Data Management Review' en 'The Business Intelligence Network', waar hij vaste columnist is.

## AANMELDINGSKAART – The Textual Data Warehouse

Woensdag 19 mei 2010 in Intres Hoevelaken

Bedrijfsnaam: .....

Naam: .....  Dhr.  Mevr.

Functie: .....

Afdeling: .....

E-mailadres: .....

Naam: .....  Dhr.  Mevr.

Functie: .....

Afdeling: .....

E-mailadres: .....

Postadres: .....

Postcode/plaats: .....

Telefoon: ..... Fax: .....

Btw-nummer (indien van toepassing): .....

Uw ordernummer: .....

Indien het factuuradres afwijkt van bovenstaand adres, graag factuuradres vermelden:

Bedrijfsnaam: .....

Afdeling: .....

T.a.v.: .....

Postadres: .....

Postcode/plaats: .....

De kosten voor deelname aan de workshop bedragen € 770,- (excl. 19% BTW) per persoon. Documentatie, lunch en koffie zijn bij de prijs inbegrepen. Inschrijvingsvoorwaarden zoals vermeld in deze brochure.

### Zo komt u in aanmerking voor korting:

- Ik ben abonnee van **Business Process Magazine/Database Magazine/Optimize** en ik profiteer daarmee van € 75,- korting.  
Mijn abonneenummer is: .....
- Ik abonneer mij tot wederopzegging op **Business Process Magazine/Database Magazine/Optimize** (doorhalen wat niet van toepassing is) en profiteer direct van de korting.
- Ik ben medewerker aan een universiteit of rijkserkende HBO-instelling en ik profiteer van 20% korting op de deelnemersprijs.

Datum: ..... Handtekening: .....

Array Seminars: specialist in IT-vakkennis!

FOLDERCODE A1014