Verbleekte ster schijnt met nieuw elan

Inmon 3.0 komt met DW2.0

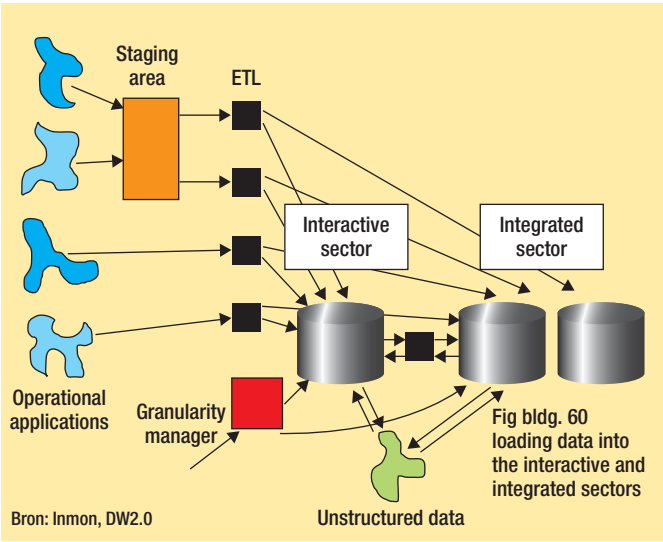
Paul van der Linden

De naam van Bill Inmon is onlosmakelijk verbonden met datawarehousing. De Amerikaan publiceerde in 1992 'Building the Data Warehouse' en wordt sindsdien gezien als de vader van het datawarehouse.

Het belang van dit boek is moeilijk te overschatten. Het maakte in een keer duidelijk dat er naast een operationele of transactionele omgeving ook bestaansrecht is voor een aparte datawarehouse-omgeving. The rest is history. Zo beschouwd is ook Ralph Kimball slechts een dwerg die voortbouwt op de schouders van Inmon. Toch kan gesteld worden dat wie vraagt naar de grondleggers van datawarehousing vaker Kimball dan Inmon hoort. Maar Inmon is terug. En wel met DW2.0: een hele nieuwe kijk op datawarehousing.

Inmon 1.0 en 2.0

Wie Inmon door de jaren heen heeft gevolgd kan grofweg drie fasen onderscheiden. Inmon 1.0 heeft betrekking op de periode 1992-2000 en start met het eerder genoemde 'Building the Data Warehouse'. Voortbouwend op het succes van dit boek werden in een rap tempo verdere boeken gepubliceerd zoals 'Using the Data Warehouse' en 'Managing the Data Warehouse'[1]. Ook aanverwante onderwerpen zoals de operationele data store (ODS) ontkwamen niet aan Inmons publicatiedrift. In deze zendingsfase was Inmon ook voortdurend onderweg om de blijde boodschap te verkondigen. De opzet hierbij was: Inmon naar binnen kruien, een half uur tot maximaal een uur laten praten en vervolgens weer op een taxi naar het vliegveld zetten op weg naar de volgende bestemming. En altijd waren er nog delen van de wereld aan te wijzen waar datawarehousing een onbekend fenomeen was.



Afbeelding 1: Laden van data in DW2.0.

Bij het verspreiden van de datawarehouse-boodschap en het verder uitwerken van de verschillende componenten daarvan, werd allengs duidelijk dat Inmon niet zoveel op had met inmiddels gangbare ontwikkelingen als internet, clickstream analysis, Customer Relationship Management, snellere responsvereisten, kleinere tot niet-bestaande batch windows vanwege 7x24, en soortgelijke ontwikkelingen. De manier waarop Inmon hiermee omging was door steeds meer databases te definiëren. Zo had je ineens ook een exploration warehouse, een data mining warehouse en een oper-mart[2]. Dat laatste is een soort mutatie van een operational data store en een datamart. Al deze componenten leefden gezellig samen in een Corporate Information Factory of kortweg CIF. Dit is Inmon 2.0: het datawarehouse-verhaal uit Inmon 1.0, dat met een veelvoud aan databases is aangevuld tot de Corporate Information Factory. Het is een fascinerende visie en een poging om vanuit een pure data-optiek de lijnen van Inmon 1.0 verder uit te rekken. Deze periode beslaat de jaren 2000 tot 2006.

Inmon 3.0: DW2.0

Eind 2005 presenteerde Inmon zijn nieuwe visie op datawarehousing onder de opwindende naam: DW2.0. Aangezien het hier

Inmon in zijn derde iteratie is, spreken we hier dan ook van Inmon 3.0. Data Warehouse 2.0 is volgens Inmon nodig omdat er sinds midden van de jaren tachtig aanzienlijke vooruitgang is geboekt op architectuurgebied, technologie en op het gebied van informatiesystemen. Waar de eerste generatie datawarehouses volgens Inmon alleen transactiedata integreerden biedt DW2.0 ook integratie van ongestructureerde data, metadata, masterdata, profile data records en heeft het ook online high performance data die te updaten zijn. Nog belangrijker is dat er nu wordt uitgegaan van de levenscyclus van data (of informatie).
DW2.0 staat voor een complete datawarehouse-omgeving. De onderdelen daarvan zijn de interactieve sector, de integratiesector, de near-line sector en de archiveringssector. Data komt meestal via de interactieve sector binnen (soms ook via de integratiesector) en stroomt vervolgens via de near-linesector naar de archiveringssector. Met elke stap neemt de ouderdom van data toe en de beschikbaarheid ervan af. Dat laatste is overigens een keuze gebaseerd op de inschatting of en zo ja hoe vaak de business nog behoefte heeft aan inmiddels verouderde data. Als de verwachting is dat deze data (bijna) nooit meer zullen worden gebruikt, worden ze verplaatst naar de archiveringssector en een minder toegankelijk medium (tape of CD-ROM in plaats van hard disk). Vaak is nog de enige reden om dit soort data te bewaren een wettelijke verplichting.

De interactieve sector is te vergelijken met een operational data store (ODS). Hier kunnen verschillende applicaties gebruik van maken. De data in het ODS zijn actueel en er is maar weinig historie voor handen. De integratiesector is eigenlijk het 'oude' datawarehouse. De wijze waarop data worden gemodelleerd is genormaliseerd. De integratiesector kent weer twee delen: de actuele, geïntegreerde data en de actieve, analytische data.
De actieve, analytische data zijn er ter ondersteuning van de reguliere, bekende statistische analyse. De taak van de actuele, geïntegreerde data is om data te verstrekken ten behoeve van alle andere sectoren van de DW2.0-omgeving. De data in de actuele, geïntegreerde omgeving kunnen gesommeerd, geherstructureerd en geaggregeerd worden. De beide componenten van de integratiesector worden bijna altijd vormgegeven als twee aparte data stores. Dit heeft te maken met het totaal andere gebruik ervan.
De actuele, geïntegreerde sector wordt gebruikt door farmers[3].
De actieve, analytische data zijn bedoeld voor de explorers[4].
De near-line sector en de archiveringssector zijn in feite beide archiveringsomgevingen, alleen is de gradatie anders. Naarmate de inschatting is dat data niet meer nodig zijn, dus minder frequent gebruikt zullen worden, kan naar de near-line sector worden doorgeschoven. Vandaar kan bij verdere veroudering en nog minder of geen voorzienbaar gebruik, worden doorgeschoven naar de archiveringssector.

De data die aanwezig zijn in de interactieve sector, zijn actueel en maximaal een maand oud. Data in de integratiesector kunnen tussen een dag en twee á drie jaar oud zijn. In de near-line sector zijn data tussen zes maanden en tien jaar oud. Data in de archiveringssector zijn tenminste vijf jaar oud. De data in de archiveringssector zijn meestal georganiseerd per tijdsdimensie (jaar) en daarbinnen per subject (onderwerp).
Een belangrijk onderscheid is nog dat in de interactieve sector de data een bijproduct zijn van de applicatie. De groepering van data is hier dus niet conform de verschillende subjecten (klant, order, product etcetera) maar ligt vast in detailtransacties. Voor de overige drie sectoren geldt dat de data wel zijn georganiseerd aan de hand van de subjecten.

De metadata worden verdeeld in lokale metadata en enterprise metadata

In alle sectoren is er sprake van verschillende soorten data. Naast de transactiedata en reference/masterdata is er ook sprake van tekstuele onderwerpen, captured text, gelinkte tekst, snapshots van data en profile data. Captured text is afkomstig uit de ongestructureerde omgeving en bestaat in de vorm van e-mails, documenten, transcripties van telefoongesprekken en andere tekstuele informatie. Profile data zijn geaggregeerde of samengevoegde data vanuit verschillende bronnen. Een klantprofiel bevat bijvoorbeeld informatie over aankopen, betalingen, webgedrag, demografische gegevens en wat dies meer zij. Deze informatie is afkomstig uit verschillende bronsystemen. Gelinkte tekst bestaat uit een verwijzing naar een tekst. In de meeste gevallen gaat het hierbij om een bidirectionele verwijzing.
Bij continue snapshot data gaat het om serie van gerelateerde records. Snapshots overlappen nooit, maar het kan voorkomen dat snapshots niet aansluiten (snapshots ontbreken). De tekstuele onderwerpen zijn categorieën van informatie die de ongestructureerde, tekstuele data ordenen. Denk hierbij bijvoorbeeld aan een ontologie. Voor omvangrijke ongestructureerde data die een lage prioriteit hebben en niet in de DW2.0-omgeving thuishoren wordt gebruik gemaakt van eenvoudige tekstuele verwijzingen (pointers).

Metadata

Inmon ziet metadata als het zenuwstelsel van DW2.0. Het belang van metadata is duidelijk: zonder metadata is niet te bepalen waar data vandaan komen, wat ze precies betekenen, welke bewerkingen erop hebben plaatsgevonden etcetera. Praktisch gezien zijn data zonder metadata daarmee waardeloos geworden.
De metadata worden verdeeld in lokale metadata en enterprise metadata. Lokale metadata hebben betrekking op een specifieke component van de omgeving. Het probleem van lokale metadata is dat ze zich niet bewust zijn van de rest van de wereld (het grotere geheel). Vandaar dat de lokale metadata worden doorgestuurd naar de enterprise metadata repository. Invoeren en

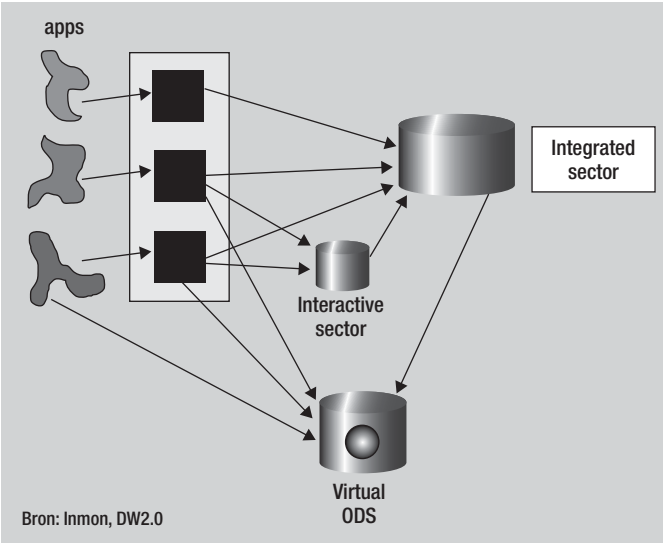
updaten van metadata gebeurt hierbij nog steeds lokaal en niet in de enterprise metadata repository. De lokale metadata vallen uiteen in de business metadata en de technische metadata. Dit onderscheid heeft te maken met de groepen gebruikers waarvoor de metadata van belang zijn.

Ongestructureerde data

Terecht wordt gesteld dat bestaande datawarehouses voornamelijk of uitsluitend gestructureerde data bevatten. Bedoeld wordt dan data die vanaf het moment van de elektronische vastlegging een gestructureerd formaat hebben gekregen, waarbij gestructureerd gelijk wordt gesteld met vastlegging in de rijen en kolommen van een database. Dit soort data is echter maar 20 procent van de totale hoeveelheid data. De overige 80 procent van de data wordt hierbij dus buiten ogenschouw gehouden. We hebben het dan bijvoorbeeld over de informatie die zich bevindt in emails, in presentaties of in Word-documenten. Ongestructureerde data kunnen in DW2.0 worden opgenomen door ze eerst om te zetten in gestructureerde data. In dit geval zal de omzetting buiten de DW2.0-omgeving plaats vinden door specifieke software. Ook kan er voor worden gekozen om een verwijzing (link, pointer) in DW2.0 op te nemen naar de betreffende ongestructureerde data.

The global datawarehouse

Grote organisaties hebben hun vestigingen wereldwijd en beschikken ergens op de wereld over een hoofdkantoor. In deze situatie is er sprake van meerdere lokale datawarehouses en een globaal datawarehouse (voor het hoofdkantoor). DW2.0 gaat er vanuit dat er eerst lokale datawarehouses zijn en vervolgens een globaal datawarehouse wordt opgezet. Dit globale datamodel is meestal veel kleiner dan een lokaal datamodel en heeft vaak niet meer dan 30 á 40 attributen. In de meeste gevallen gaat het dan om financiële data, soms aangevuld met klantdata. Het proces bestaat eruit dat de attributen van het globale datamodel worden gemapt op de verschillende lokale datamodellen. Vervolgens kunnen fysieke modellen worden gemaakt en kan het ETL-proces worden ingezet om gegevens met de vereiste frequentie door te zetten van de lokale datawarehouses naar het globale datawarehouse.



Afbeelding 2: Virtual Operational Data Store (VODS).

De bouw van DW2.0

Het zal niet verbazen dat Inmon onveranderd een voorstander is van genormaliseerde datamodellen. Dit terwijl er hele legioenen BI-ontwikkelaars rondlopen die alleen maar in Kimball-sterren denken en voor wie normaliseren even bekend is als de precieze locatie van Atlantis. Niet dat Inmon niet weet wat stermodellen zijn of er geen plaats aan geeft. Datamarts hebben volgens Inmon een stermodel als basis. Hij voegt er overigens wel aan toe dat het stermodel voor elke datamart anders is. Dat kan worden opgevat om vooral niet te denken in conformed dimensions. Gezien de opzet van DW2.0 zou dat ook een nutteloze eis zijn. Alle componenten uit de DW2.0-omgeving kunnen bron zijn voor een datamart. Er kunnen echter ook geen andere bronnen voor een datamart zijn dan een DW2.0-component.
Inmon onderscheidt twee verschillende soorten datamarts: de spontane en de permanente. Een permanente datamart is een datamart die over een lange tijdsperiode Key Performance Indicatoren traceert. Zo'n datamart kent meestal een afdelingsoriëntatie. Voor een spontane datamart geldt dit niet. Deze ontstaat om een specifieke, ad hoc behoefte in te vullen en heeft een minder lange levensduur dan zijn permanente broer.
Een van de bezwaren tegen eerste-generatie datawarehouses zoals door Inmon beschreven, was dat je een ongelooflijk lange tijd nodig had om het corporate datamodel op te stellen. In veel gevallen bleek in de praktijk dat dit alle lucht uit het datawarehouseproject liet lopen en men niet verder kwam dat een corporate datamodel dat permanent in wording was. Inmon stelt in DW2.0 voor om een datamodel te kopen, of te leasen! Vervolgens kan er getuned en geïmplementeerd worden. Hierbij wordt steeds een bepaald deel van het model ingevuld. In termen van datamodellen onderscheidt Inmon drie modellen: het entiteiten-relatiediagram (ERD), de data item set (DIS) en het fysieke model.

VODS: the virtual ODS

Indien behoefte bestaat aan extreme flexibiliteit, extreme snelheid en we het hebben over eenmalige query's, is de Virtual Operational Data Store (VODS) wellicht een uitkomst. Inmon ziet het als een versterking van de interactieve sector. Beide technologieën zijn in zijn ogen complementair. Een groot verschil tussen beide is dat in het geval van de interactieve sector we het hebben over een permanente, complete en te auditen dataset. Bij een VODS is dit niet het geval. Het werkt op een beperkte dataset (hetgene wat op dat moment aanwezig is), is niet compleet en auditen heeft hier dus geen zin. Zoals elk ODS heeft ook de VODS geen of slechts beperkte historische gegevens. VODS kan gezien worden als een specifieke invulling van Enterprise Information Integration (EII). Kalido-oprichter en chief strategist

Andy Hayler heeft in het verleden al zijn licht laten schijnen over EII als alternatief voor een datawarehouse-omgeving. Inmon ziet het dus als aanvullende technologie. VODS kan dan ook het beste worden vergeleken met de functie van een pleister in een verbanddoos. Wellicht eventjes nuttig, maar nooit een volwaardige oplossing.

Conclusies

Hoe nieuw en verrassend is Inmons DW2.0? Wie zijn bekendste boek 'Building the Data Warehouse' heeft gelezen ziet vooral nuanceverschillen. Voor wie dit boek niet kent en datawarehouse-land heeft betreden aan de hand van Ralph Kimball biedt het echter veel nieuws.

> Inmon onderscheidt twee verschillende soorten datamarts: de spontane en de permanente

Inmon geeft zelf aan dat de grootste verschillen zitten in het meenemen van metadata, master data, ongestructureerde data, profile data en als centraal uitgangspunt de life cycle van data/informatie. De meeste van deze onderwerpen zaten ook al in 'Building', alleen wordt er in DW2.0 meer aandacht aan besteed. Dit zijn ook de bekende, actuele onderwerpen – dus niet echt nieuw. Wel zijn het onderwerpen waaraan niet elke organisatie voldoende aandacht besteedt. Voor deze organisaties gaat het wel om nieuwe onderwerpen.
Positief is dat Inmon inmiddels aandacht besteedt aan meer recente ontwikkelingen zoals Enterprise Information Integration (zie zijn Virtual Operational Data Store), ongestructureerde data en master/reference data. Hij geeft een plaats aan stermodellen (toepassen in datamarts) en probeert pragmatischer om te gaan met het corporate datamodel (kopen of leasen). De enige echte verandering betreft de datamarts. Waren deze voorheen subsets uit het grote datawarehouse, nu zijn het datasets die kunnen putten uit elke component van de DW2.0-omgeving.
Op een aantal punten is het verhaal te kort door de bocht. Bij de bespreking van het global datawarehouse gaat Inmon uit van een specifiek besturingsmodel. Er zijn echter nog andere besturingsvarianten die zeker leiden tot een andere verhouding tussen lokale en globale datawarehouses. Inmon roept dat je soms een staging area nodig hebt – in mijn ervaring heb je die altijd nodig. Terughalen van data uit de archiveringssector kan volgens Inmon met weinig kosten. Juist de verschillen in gebruikte software en verouderde technologieën zorgen ervoor dat dit soort acties niet alleen veel tijd en geld kosten, maar soms niet eens praktisch uitvoerbaar zijn.
Ondanks deze kanttekeningen en de weinige vernieuwende theorie is DW2.0 toch interessant voor organisaties. Niet zozeer voor Inmon-kenners die wellicht nieuwsgierig zijn of hij ook actuele thema's toelaat (kort antwoord: ja), maar met name voor de datawarehouse- en BI-ontwikkelaars die 'Kimballiaans' aan de weg timmeren. Een schepje Inmon helpt bij het verbreden van de horizon. Van al dat staren naar sterren wordt je ook maar blind.

Noten

1. *In 1997 publiceerde The Data Warehousing Institute (TDWI) onder redacteurschap van Herb Edelstein een boek met de title 'Building, Using and Managing the Data Warehouse'. Het boek bevat bijdragen van verschillende auteurs, maar niet van Bill Inmon.*
2. *Zie boekbespreking 'Het pakhuis is een complete fabriek geworden' (DBM 1, februari 2002) over 'Corporate Information factory' en 'Data Warehousing for E-Business'.*
3. *Farmers: type gebruiker dat behoefte heeft aan van te voren te definiëren informatie (bijvoorbeeld standaardrapportages of geparametriseerde rapporten).*
4. *Explorers: type gebruiker dat op basis van de beschikbare informatie pas de vraag kan formuleren.*

Paul van der Linden (Paul.PFH.vanderLinden@AtosOrigin.com) is senior consultant Data Warehousing/BI bij Atos Origin en geeft leiding aan Data Warehousing Cost & Lifecycle Management (CLM).

Update

Nieuwe versie Cognos 8

Cognos introduceert een nieuwe versie van Cognos 8 Business Intelligence (MR1). Deze versie maakt BI toegankelijk voor meer gebruikers, met nieuwe functionaliteit voor zoeken en rapporteren en maakt het mogelijk investeringen in de operationele infrastructuur te maximaliseren door uitgebreide ondersteuning van bedrijfsapplicaties.
De nieuwe functionaliteit omvat onder meer Cognos Go!, een zoekservice waarmee gebruikers direct relevante strategische bedrijfsinformatie kunnen vinden. Naast de mogelijkheid om in Microsoft Excel rapporten te integreren die steeds geactualiseerd worden, kunnen gebruikers nu via de Cognos Office Connection ook analyses en meetgegevens zien, bewerken en actualiseren binnen PowerPoint en Excel. Tevens bevat de nieuwe versie Report Packs voor mySAP FI/CO (voor SAP R/3-data) en Siebel CRM, en biedt gecertificeerde ondersteuning van SAP NetWeaver. Bovendien heeft Cognos heeft ook de ondersteuning uitgebreid voor HP Integrity-servers met HP-UX 11i, Microsoft SQL Server 2005, BEA WebLogic Server 9, Red Hat Enterprise Linux 4.0, Sybase Adaptive Server Enterprise (ASE) 15, Teradata Warehouse 8.1, ORACLE Database 10g v2 en het Netezza Performance Server-systeem.

Meer informatie op www.cognos.nl