Centennium ontwikkelt methodiek voor generatie van DWH

Toekomstvast datawarehouse

Erik Fransen en Antoine Stelma

Sinds de jaren negentig van de vorige eeuw is het datawarehouse niet meer weg te denken uit de dagelijkse praktijk van managementinformatie-voorziening. Het boek 'Building the Datawarehouse' van Bill Inmon uit 1991 gaf het startschot voor de wereldwijde adaptatie van datawarehousing, gevolgd door Ralph Kimball met de klassieker 'The Datawarehouse Toolkit' uit 1996. Rapportages en analyses worden vanaf dat moment steeds vaker ontwikkeld op basis van een datawarehouse.

Het datawarehouse moest daarbij zorgen voor centrale historische opslag en, waar nodig, integratie van bedrijfsgegevens. De argumenten voor inzet van een datawarehouse waren destijds sterk technisch gedreven: 'the query that dims the light' moest hoe dan ook voorkomen worden, vandaar een strikte technische scheiding tussen brondata enerzijds en data in het datawarehouse anderzijds. Inmon richtte zich met zijn visie sterk op de 'achterkant' van het datawarehouse: een centraal genormaliseerd datamodel voor historische opslag van data en bijbehorende ETL-processen om brondata te ontsluiten en integreren. Kimball verlegde de aandacht echter naar de 'voorkant', voor velen een zwak punt van het Inmon gedachtegoed: via dimensionele modellen, opgebouwd uit feiten en dimensies, werden de brondata omgevormd naar een structuur waar de eindgebruiker zich direct in herkende. Kimball's aanpak, ook wel bottom-up datawarehousing genoemd, had echter als nadeel dat data-integratie en data-opslagvraagstukken pas laat in het ontwikkeltraject zichtbaar werden. Inmon's aanpak, ook wel top-down datawarehousing genoemd, kenmerkte zich door de grote aandacht voor het ontsluiten van brongegevens en het gebrek aan aandacht voor de informatiebehoefte van de eindgebruikers.

In het nieuwe millennium werd de noodzaak voor een gebruikersorganisatie om snel en eenvoudig brongegevens samen te voegen en te aggregeren steeds groter, bijvoorbeeld omdat een integraal klantbeeld noodzakelijk werd geacht. Daarnaast werd eenvoudige en eenduidige toegankelijkheid van historische gegevens omwille van wetgeving en data-analyse belangrijker. Als gevolg groeide het datawarehouse steeds harder qua omvang. Niet alleen in de diepte door behoefte aan detaildata, maar vooral in de breedte; nieuwe databronnen gekoppeld aan en geïntegreerd met het datawarehouse. Met andere woorden: de 'achterkant' en de 'voorkant' van het datawarehouse werden beide cruciaal voor het succes van BI. Anno 2010 spreken we over DW 2.0, onder aanvoering van wederom Bill Inmon. Hierin hebben de genoemde kenmerken van het Inmon datawarehouse, maar ook de kenmerken van het Kimball datawarehouse een plek gekregen. DW 2.0 wordt derhalve ook wel een hybride datawarehouse-architectuur genoemd.

Lessons learned

Centennium heeft zich door de oorspronkelijke concepten van Inmon, Kimball en DW 2.0 laten inspireren. Deze inspiratie is gecombineerd met de eigen ervaringen die zijn opgedaan in het verleden met implementaties van datawarehouses. Uiteindelijk is dit vertaald naar vijf 'lessons learned' die de basis vormen voor het toekomstvaste datawarehouse anno 2010:

1. De datawarehouse-architectuur faciliteert de integratie en opslag van (historische) brondata en geeft de BI-gebruiker zonder drempels toegang tot deze data. Een hybride architectuur sluit hierbij aan;
2. Data worden volledig, correct en waar nodig geïntegreerd opgeslagen en zijn langdurig toegankelijk voor wetgever en eindgebruikers. Het datawarehouse waarborgt dat aan deze eisen wordt voldaan;
3. Data zijn traceerbaar naar de originele databron voor controle. Het datawarehouse waarborgt dat aan deze eis wordt voldaan;
4. Het datawarehouse is adaptief: zowel aan de achterkant (aansluiting van nieuwe bronnen) en voorkant (nieuwe dimensionele modellen) kan snel en eenvoudig gereageerd worden op veranderingen;
5. De eindgebruikerorganisatie ontwikkelt, indien gewenst, grotendeels zelfstandig het datawarehouse. Dit is mogelijk omdat de hedendaagse structuur van datawarehouses een aantal standaardpatronen kent, waardoor datawarehouses grotendeels geautomatiseerd worden gegenereerd.

CDM

Deze vijf 'lessons learned' zijn aanleiding geweest om een methodiek te ontwikkelen, waarmee organisaties in staat zijn om in korte tijd een bedrijfsbreed datawarehouse te realiseren en vervolgens te onderhouden. Met de komst van Data Vault, als de facto standaard voor het intelligent en eenduidig opslaan van data, heeft Centennium het missende stukje van de puzzel kunnen invullen. De Centennium Datawarehouse Methodiek richt zich niet op het uiteindelijk gebruik van informatie uit het datawarehouse, maar is daarentegen wel onderdeel van het overkoepelend proces van het organiseren van Business Intelligence. Organiseren van BI valt buiten de scope van dit artikel.

De vijf 'lessen' hebben geleid tot de Centennium Datawarehouse Methodiek (CDM), met de volgende kenmerken:

- Volledig automatische generatie van het datawarehouse en dimensionele datamarts op basis van beschrijvende metadata;
- Volledig modulaire opzet waardoor aanpassingen en uitbreidingen snel en eenvoudig realiseerbaar zijn, zonder gebruik te maken van complexe datawarehouse- en BI-tools;
- Eenduidige registratie van alle bedrijfsgegevens in het datawarehouse, waardoor volledige transparantie en herleidbaarheid van gegevens is gewaarborgd;
- CDM is geen tool, maar een methodiek en dus toolonafhankelijk;
- De eindgebruikerorganisatie is in staat snel en zelfstandig het datawarehouse te ontwikkelen en beheren.

CDM is opgebouwd uit drie pijlers die tezamen het fundament van de methodiek vormen. De drie pijlers kennen ieder hun eigen voordelen, maar juist de combinatie maakt de methodiek uniek en van grote toegevoegde waarde:

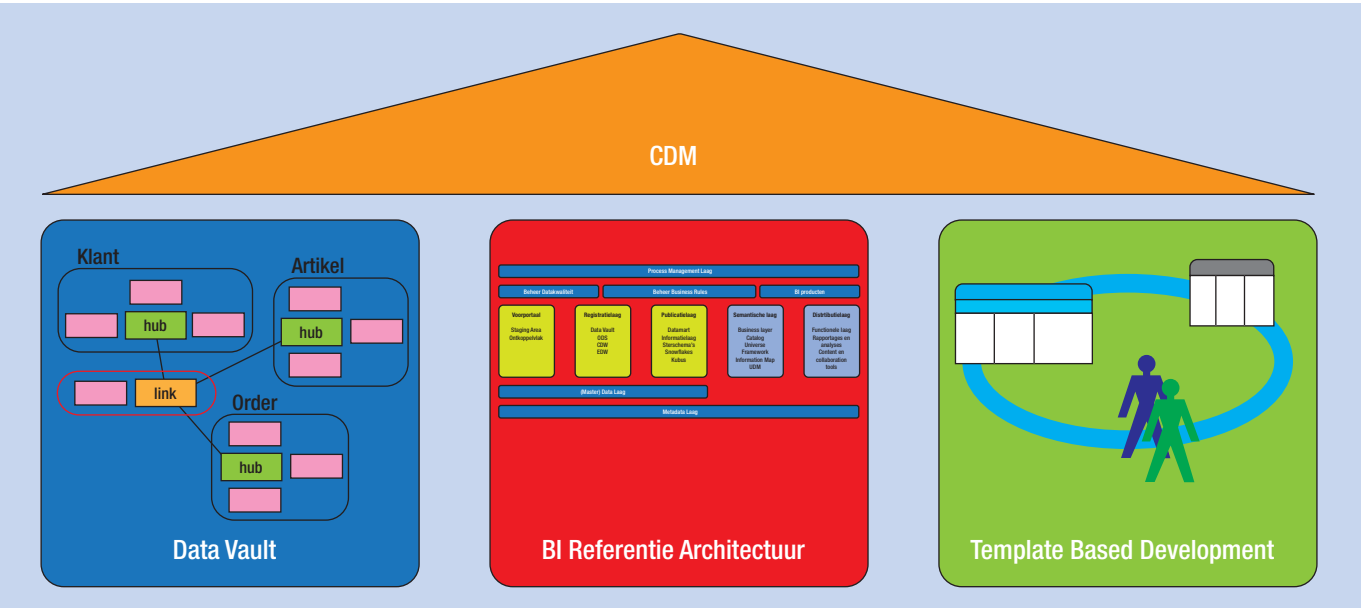
Structureren met BI-referentiearchitectuur. Deze gedegen basis voor moderne datawarehouses combineert praktijkervaring met bewezen methoden uit het vakgebied. De BI-referentiearchitectuur kenmerkt zich door gelaagdheid, waarbij elke individuele laag een duidelijk afgebakende toegevoegde waarde biedt. Deze gelaagdheid draagt tevens zorg voor ontkoppeling van het registreren van gegevens en het gebruik ervan. Hiermee voldoet de BI-referentiearchitectuur aan het paradigma 'loosely coupling & high coherence' en past daarmee naadloos in alle moderne IT-architecturen;

Modelleren met Data Vault. De datamodelleringstechniek Data Vault, bedacht en ontwikkeld door Dan Linstedt, is ontworpen vanuit de gedachte dat het moderne datawarehouse flexibel, eenvoudig uitbreidbaar, volledig traceerbaar en auditeerbaar moet zijn. Data Vault wordt reeds jaren met veel succes toegepast en ontwikkelt zich langzaam maar zeker tot wereldwijde industriestandaard;

Genereren via Template Based Development. Door slim gebruik te maken van de onderliggende database worden, met Template Based Development (TBD), het datawarehouse en de datamarts volledig transparant gegenereerd. Hierdoor wordt het ontwikkeltraject drastisch verkort, dalen ontwikkel- en beheerkosten en is de foutgevoeligheid klein. TBD kenmerkt zich door een incrementele ontwerp- en ontwikkelaanpak en maakt technische realisatie en beheer mogelijk in hoog tempo, foutloos en in korte, snel opeenvolgende cycli.

Structureren

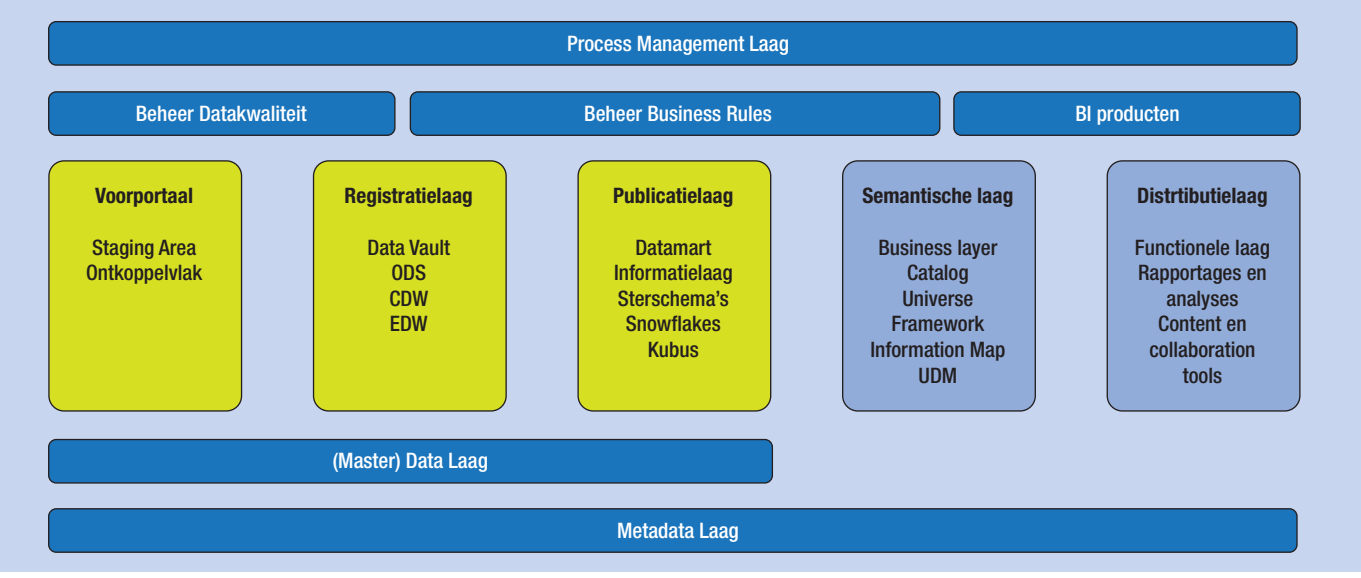
Een flexibele en schaalbare datawarehouse-architectuur (binnen CDM noemen we dit de BI-referentiearchitectuur) maakt bij voorkeur onderdeel uit van de bedrijfsarchitectuur. Een datawarehouse wordt daarbij vaak gepositioneerd aan het einde van de keten, maar dient op haar beurt steeds vaker ook als gegevensleverancier. Hierdoor ontstaat een zogenaamd dataverkeersplein:



Afbeelding 1: Drie pijlers van CDM.

het datawarehouse verwerkt op een consistente en eenduidige wijze grote hoeveelheden data. Het is dus van absoluut belang dat er goed wordt nagedacht over de gewenste architectuur die moet worden ingezet. De architectuur bepaalt voor een groot deel de structuur waarin de gegevens worden opgevangen, verwerkt en gepresenteerd.
Uiteraard is een parallel te trekken met de toekomstige onderhoudskosten van het systeem. Hoe eenduidiger de verwerking wordt opgezet hoe minder kosten er worden gemaakt bij de exploitatie van het systeem. Er dient uiteraard wel een balans te worden gevonden tussen eenduidigheid, uitbreidbaarheid en flexibiliteit. Nadat de in- en uitgangspunten van een architectuur zijn bepaald, worden de koppelvlakken gedefinieerd. De koppelvlakken beschrijven de interactie van het datawarehouse met de 'buitenwereld'. Wij onderscheiden hierbij drie soorten koppelvlakken:
Functioneel: de BI-omgeving houdt zich aan afspraken met IT, gebruikers en aan SLA's;
Applicatief: de interactie tussen de bronapplicaties en het datawarehouse en de interactie tussen datawarehouse en de BI-tools;
Infrastructureel: de interactie met het besturingssysteem, services, connectivity, interoperabiliteit, standaarden (XML, SOAP enzovoort).

Daarnaast dienen ook de ontkoppelvlakken benoemd te worden. Een BI-architectuur bestaat uit een aantal lagen. Elke laag heeft een eigen functie en dient derhalve bij voorkeur ontkoppeld te zijn van de vorige en/of de volgende laag. Hoewel het aantal lagen in principe eindeloos kan en mag zijn, is er een aantal lagen te onderkennen die als uitgangspunt dienen binnen CDM, zie afbeelding 2:
Staginglaag. De data worden ontkoppeld van de bronsystemen, verzameld en eventueel gecontroleerd op vooraf gestelde kenmerken;
Registratielaag. De data worden opgeslagen voor historisch en actueel gebruik. De de facto standaard is Data Vault, maar alternatieven zijn mogelijk. Bij gebruik van Data Vault worden slechts de feiten uit de bronnen geregistreerd. De data worden dus nog niet verrijkt met business rules. Alle data worden inclusief historie 100 procent auditeerbaar naar de bron toe opgeslagen;
Publicatielaag. De data uit de registratielaag worden verrijkt met business rules en opgeslagen voor raadpleging door eindgebruikers. Deze laag is geoptimaliseerd voor het snel bevragen van data in een structuur die eenvoudig is te begrijpen. Voorbeelden hiervan zijn sterschema's.
Semantische en distributielaag. Deze maakt deel uit van de BI-functie en valt derhalve buiten de scope van dit artikel;
Metadata management. Dit draagt zorg voor de beschrijving van alle metadata in het datawarehouse. Het geeft context aan de data die aan de gebruikers worden gepresenteerd. De metadata worden bijvoorbeeld gebruikt om de Data Vault en data marts (opnieuw) te genereren, maar bieden ook de mogelijkheid om audit-trails uit te voeren (van rapport tot bron);
Masterdata management. Integratie van brondata vindt plaats via masterdata management: in deze laag wordt gedefinieerd hoe integratie van data uit verschillende bronnen moet plaatsvinden: op basis van deze regels wordt het datawarehouse automatisch gevuld met geïntegreerde brondata;
Process management. De process managementlaag draagt zorg voor een ongestoorde afhandeling van de processtappen bij het genereren van het datawarehouse. Indien een stap eventueel niet doorlopen dan wel afgerond kan worden, wordt dit automatisch gemeld aan de procesverantwoordelijke via uitgebreide en gebruikersvriendelijke logging;
Datakwaliteit. Het succes van het datawarehouse staat of valt met de kwaliteit van de gebruikte data. Door de vergaande integratie en samenkomst van data op een centraal punt worden problemen, zowel de bekende en vooruitgeschoven als de voorheen



Afbeelding 2: BI-referentiearchitectuur.

onbekende problemen zichtbaar. In de analyse van de data moet rekening worden gehouden met de bruikbaarheid en waarde van data op korte en lange termijn;
Business rules. Om de feiten uit het datawarehouse in context te plaatsen dienen de data te worden verrijkt met aanvullende berekeningen, filteringen, cumulaties enzovoort. Een business rule kan een beperkte levensduur hebben. Om een consistent beeld te houden met de data is het van belang dat business rules historisch worden opgeslagen. Business rules dienen separaat te worden beheerd.

Modelleren

Data Vault is in 2002 geïntroduceerd door Dan Linstedt. De kern van Data Vault is eigenlijk niet het model zelf maar de manier hoe het met data omgaat. Gedreven door de Amerikaanse wetgeving heeft Linstedt een model ontwikkeld waarbij data op een vaststaande manier worden geregistreerd in een kluis. De kluis is niet toegankelijk voor eindgebruikers, maar alleen voor Data Vault ontwikkelaars. De primaire taak is om de brondata feitelijk te registreren.
De Data Vault wordt gepositioneerd in de registratielaag. Rapportages, kubussen enzovoort worden gegenereerd vanuit de Data Vault. Alle eindgebruikers hebben dus initieel dezelfde data als vertrekpunt. Interpretaties van de data vinden pas plaats na de registratielaag.
De belangrijkste kenmerken van Data Vault zijn:
– een Data Vault bevat data op het laagst mogelijke detailniveau;
– een Data Vault bevat historische data;
– een Data Vault is een uniek gelinkte verzameling van genormaliseerde tabellen;
– een Data Vault is eenvoudig uitbreidbaar.

Een Data Vault is dus eigenlijk een soort bibliotheek voor data. Alle binnenkomende data worden geregistreerd en kenmerken worden toegekend. De regel is dat niets mag worden verwijderd en hooguit door middel van een start- en einddatum door de tijd heen aangepast. Er wordt geregistreerd door wie en wanneer de data zijn aangeleverd. Hierdoor is het te allen tijde mogelijk om in de tijd terug te herleiden naar een bepaalde status van de data.

Genereren

In de wereld van datawarehouses speelt ETL een grote rol, ooit bedoeld om grafisch de ETL-stromen te modelleren. Door algemene generieke code te generen zijn plotseling veel mensen in staat om ETL toe te passen. De tools hebben hun kracht bewezen maar tegelijkertijd leverde dit een scala aan extra problemen op. De ETL-tool werd een wereld op zich en menige organisatie verslikte zich in de overgang naar andere tools. Ook het 'patchen & upgraden' levert de nodige hoofdbrekens op. Niet zelden raakt goed werkende code beschadigd of werkt het anders na een dergelijke actie. Eigenlijk is het weer tijd om terug te keren naar de basis. Met de komst van modelleringtechnieken als Data Vault en al eerder Kimball's sterschema zijn er eigenlijk enkele vaste patronen te herkennen in het ETL-proces. In Data Vault zijn de regels voor het vullen van de tabellen (hubs, satellieten en links) eenduidig. Doordat Data Vault zich richt op het registreren van feiten kan men de regels strak hanteren. In de wereld van Kimball, die zich richt op presenteren ligt dit iets complexer. Maar het vullen van dimensies, feiten en aggregaten gebeurt weer op een eenduidige manier. Maar hoe lossen we dit op in de ETL? We zouden een aantal voorbeelden kunnen maken en deze telkens kopiëren en aanpassen. Als we teruggaan naar de code op database niveau kunnen we vrij eenvoudig een aantal templates definiëren. Deze templates zijn eerst geoptimaliseerd voor de taak die ze uitvoeren. Daarna kunnen we ze eindeloos aanroepen op basis van metadata. Immers, als we de patronen herkennen is dit repeteerbaar.

Als voorbeeld noemen we hier het laden van een hub-tabel in Data Vault:
Eis: laad alleen niet bestaande aangeleverde business keys;
Voorwaarde: controleer of business key bestaat;
Actie: voldoet aan voorwaarde: doe niets;
Actie: voldoet niet aan voorwaarde: voeg business key toe, geef een uniek id af als primaire sleutel, voeg gegevens over de bron en de laaddatum toe.
Om de templates aan te roepen wordt gebruik gemaakt van metadata. Deze metadata beschrijven eenvoudig de route tussen een aanleverbestand en de hub. De code die vervolgens gegenereerd wordt op basis van de metadata is 100 procent in de programmeertaal van de database, eenvoudig en doordacht gericht op de taak en transparant in de opbouw. Immers, zonder kans op verschillen wordt de code nu eenduidig gegenereerd. CDM is geheel template-gebaseerd, wat inhoudt dat het gehele datawarehouse te genereren is.

Conclusie

Duidelijk is dat de Centennium Datawarehouse Methodiek een weldoordachte methodiek is die voor een toekomstvast datawarehouse het fundament legt. Optimaal gebruik makend van bestaande methoden uit de markt in combinatie met onze eigen ervaring leidt dit tot een pragmatische aanpak voor het realiseren van een datawarehouse. Nu het technische pijnpunt uit een gemiddeld datawarehouse-project kan worden gehaald, is er meer ruimte om na te denken over de succesroute voor de toekomst. Een datawarehouse is vaak verre van statisch en levert optimale toegevoegde waarde als het gecontroleerd kan meegroeien met de eindgebruikersorganisatie. Een belangrijk aspect is dat een organisatie zelfstandig het datawarehouse kan aanpassen, beheren en verder uitbreiden. Samen met een BI-kennispartner kan dan optimaal rendement worden behaald.

Erik Fransen is senior business consultant en **Antoine Stelma** is Lead BI Architect bij Centennium BI Expertisehuis.