

Zwaartepunt datawarehouse-proces ligt bij de Grote Transformatie

De volgende generatie EDW (2)

Lidwine van As

In het eerste deel van dit drieluik, 'De volgende generatie EDW' gepubliceerd in DB/M 5, presenteren we een EDW-architectuur die tegemoet kan komen aan de veranderde (en veranderende) eisen die aan een moderne EDW-omgeving gesteld worden. Opgemerkt werd dat een methodiek als de Data Vault een prima kader biedt voor dit soort ontwikkelingen.

Methodologisch gezien richt de Data Vault zich echter voornamelijk op het centrale datawarehouse. De architectuur biedt een duidelijke filosofie ten aanzien van datakwaliteit, auditbaarheid, traceerbaarheid, performance en standaardisering, maar biedt voor het maken van datamarts weinig tot geen concrete handvatten of richtlijnen.

En dat terwijl daar, door de aard van Data Vault, het zwaartepunt van het datawarehouse-proces komt te liggen: de 'Big T' of 'Grote Transformatie'. We merken dat de onzekerheid over de juiste aanpak voor velen een obstakel is om voor een volgende generatie architectuur te kiezen. In het eerste deel hebben we de volgende generatie EDW grotendeels als een *black box* neergezet. In dit artikel nemen we een kijkje in die box, en stellen we een aanpak voor de Big T voor.

Afbeelding 1 toont de globale architectuur van de volgende generatie EDW. De staging is een vluchtige opslaglaag die – qua structuur – zoveel mogelijk 1:1 met de bron overeenkomt. Deze laag is reeds door Dr. Kimball [1] uitgebreid beschreven. Het CDW is het *system of fact*: de opslag hier ligt zo dicht mogelijk bij wat uit de bron binnenkomt. In het eerste deel van dit drieluik is hier uitgebreid op ingegaan.

Datalogistiek gezien ligt achter het centrale datawarehouse (CDW) het 'klantorder-ontkoppelpunt': tot en met het CDW wordt *aanbodgedreven* gewerkt, alles daarachter is *vraaggedreven*. Er wordt alleen actie ondernomen na het CDW indien er een concrete klantvraag is, en alle ondernomen acties worden gedicteerd door die klantvraag.

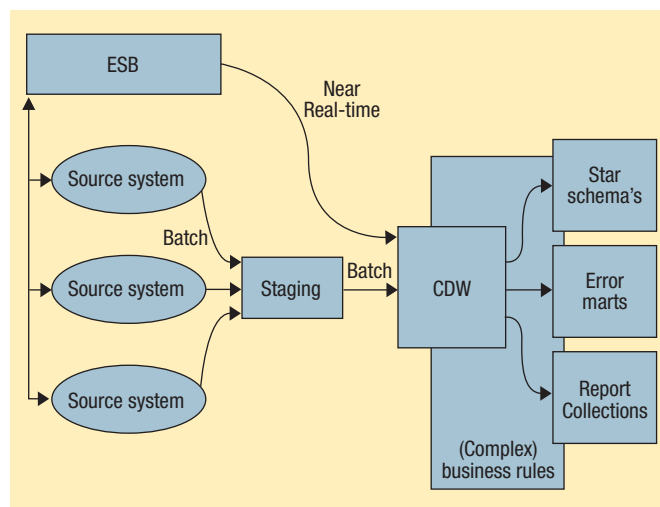
Uncharted territory: the Big T

Zoals gesteld, komt door de system-of-fact aanpak, gecombineerd met het '100% of the data 100% of the time'-principe, het zwaartepunt van de datalogistiek van bron naar klant tussen het

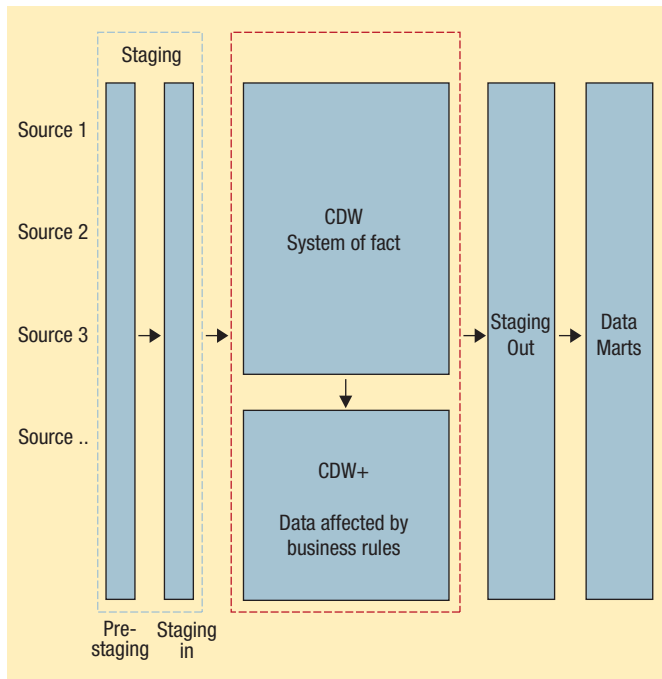
CDW en de datamarts te liggen. Om te bepalen hoe we die component moeten vormgeven, is het zaak om helder te hebben waarom we met het EDW een nieuwe weg zijn ingeslagen.

In deel 1 in DB/M 5 zijn enkele problemen van eerdere generaties datawarehouses uitgebreid behandeld. Deze lagen met name op het gebied van beheersbaarheid, onderhoudbaarheid, uitbreidbaarheid, traceerbaarheid en schaalbaarheid. Deze drivers zullen ook in hoge mate de inrichting van de Big T bepalen. Daarbij streven we dan ook naar:

- Inzichtelijkheid, duidelijke structuren, iedere component krijgt een duidelijke eigen plaats in het verwerkingsproces;
- Hergebruik (en daarmee beperking van duplicatie) van functionaliteit;
- Eenvoud van verwerking: liever drie eenvoudige procedures dan een complexe; modularisatie van functionaliteit. Hierbij merken we tevens op dat de eenvoud van een object of een procedure mede bepalend is voor zijn herbruikbaarheid – uit de wereld van de object-oriëntatie weten we dat eenvoudige objecten een grotere mate van herbruikbaarheid hebben dan complexe;
- Standaardisatie, van verwerkingsslagen (met andere woorden het inperken van de customisatie van functionaliteit), maar ook van het onderliggende datamodel;
- Traceerbaarheid, ten aanzien van de resultaten van gegevens-



Afbeelding 1: Volgende generatie EDW-architectuur (Bron: Dan Linstedt – The business of Data Vault modeling).



Afbeelding 2: Volgende generatie EDW-architectuur – CDW en CDW+.

afleidingen (business rules) en integratieslagen gelden dezelfde eisen met betrekking tot traceerbaarheid en historie-management als die gesteld worden aan brongegevens;

- Efficiency en performance-optimalisatie, er wordt naar gestreefd om 'zuinig' met de beschikbare middelen om te springen. Het is bijvoorbeeld zinloos om verwerkingstijd te besteden aan gegevens die in een later stadium alsnog uitgefilterd worden. Dit vraagt om een doordachte opzet qua volgorde van het proces. Hoe efficiënter het datalogistische proces is, hoe beter dat is voor de performance ervan;
- Testbaarheid, de dreiging van een steeds grotere vervlechting van datalogistiek is groot en kan resulteren in een degenererend systeem in termen van testbaarheid naarmate het EDW groeit.

Laten we nu eens nader bekijken wat er aan verwerkingen plaatsvindt in die Big T. (Voor een diepgaander inzicht in het ETL-proces verwijzen we naar [1]). Enerzijds worden er datagerichte bewerkingen uitgevoerd: bewerkingen die puur betrekking hebben op de data zelf, en impact hebben op hetzij de gegevensinhoud, hetzij de samenstelling (en omvang) van de gegevensverzameling (Kimball refereert hieraan als 'Cleansing & Conforming'). Denk aan: controle op datakwaliteits-issues, en afhandeling daarvan; uifilteren/selecteren van gegevens voor verdere verwerking; integratie en consolidatie van ruwe gegevens; gegevensafleidingen. De uit te voeren bewerkingen stellen geen eisen aan de modellering van de gegevens, we kunnen deze dus zelf kiezen. Daarnaast vindt, dichter naar de datamart toe, een serie bewerkingen plaats die erop gericht is om de eerder bewerkte data adequaat aan de gebruiker te presenteren (door Kimball aangeduid als 'Data Delivery'). Deze bewerkingen

worden in hoge mate gestuurd door de eisen die door de end-user tooling opgelegd worden; de wijze van gegevensmodellering is er daar één van.

Denk aan: hermodellering naar dimensies en feittabellen (inclusief opbouw van aggregaten); de opbouw van kubussen; het platslaan in een flat file bijvoorbeeld voor datamining-toepassingen en wat verder maar denkbaar is aan gewenste user formats.

CDW+

Om de Big T inzichtelijk te houden, ligt het voor de hand om deze in te delen op een wijze die de tweedeling weerspiegelt, zodat iedere functie en gegevenselement op een voorspelbare plaats terug te vinden is. Daarom maken we een onderverdeling in een gegevensgerichte CDW+-omgeving, en een presentatie-georiënteerde Staging Out area (STO).

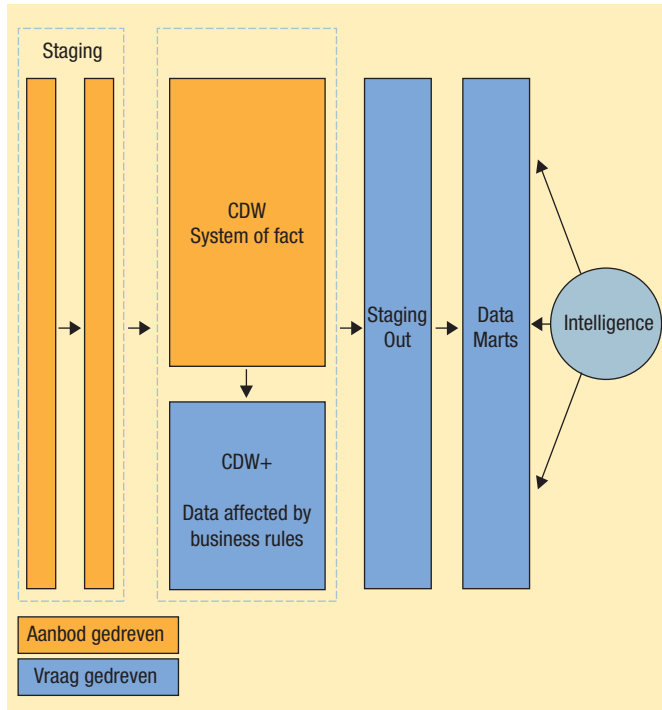
In de CDW+-omgeving brengen we de uitvoering van alle datagerichte bewerkingen onder en bewaren we de resultaten ervan. Aangezien de bewerkingen die we in het CDW+ positioneren geen expliciete eisen stellen aan de modelleerwijze, kunnen we deze zelf bepalen. Met het oog op de voordelen ten aanzien van historie-management en traceerbaarheid kiezen we ook hier weer voor Data Vault modellering.

Inzichtelijkheid

Een gevaar dat altijd op de loer ligt bij het inrichten van de datalogistiek tussen datawarehouse en datamarts, is het ontstaan van onderlinge vervlechting tussen datamarts. Datamarts raken met elkaar vervlochten als een gegeven dat in de ene datamart berekend wordt, hergebruikt wordt in een andere. Hierdoor worden datamarts in hun opbouw gaandeweg steeds verder afhankelijk van elkaar. Deze situatie kan vermeden worden door niet toe te staan dat datamarts elkaars gegevens hergebruiken; zonder aanvullende maatregelen, zou dat echter weer kunnen leiden tot duplicatie van afleidingsregels in meerdere omgevingen en op meerdere punten in de verwerking, met alle nadelige gevolgen van dien.

In de CDW+-omgeving brengen we de uitvoering van alle datagerichte bewerkingen onder

Wanneer we de datagerichte bewerkingen geheel loskoppelen van de presentatiegerichte bewerkingen, en deze in het CDW+ onderbrengen, ontstaat als het ware een centraal datawarehouse voor afgeleide gegevens (zie afbeelding 2). Daarmee worden de mogelijkheden voor hergebruik van (resultaten van) business rules gemaximaliseerd: herbruikbare elementen komen in een zo vroeg mogelijk stadium beschikbaar voor gebruik in andere afleidingen en bij het samenstellen van de datamarts.



Afbeelding 3: Aanbod- en vraaggedreven.

Efficiency en performance

De keuze om bewerkingen die de samenstelling en omvang van de gegevensset beïnvloeden zo vroeg mogelijk in de Big T uit te voeren, is ook met het oog op een efficiënte procesuitvoering zeer verstandig. Immers, hoe minder gegevens er bewerkt worden, hoe beter dat in het algemeen voor de performance is. Doordat de datalogistieke functies naar het CDW+ Data Vault-georiënteerd zijn, wordt optimaal geprofiteerd van de voordelen die een zero-update strategie biedt in termen van parallelisatie en schaalbaarheid.

Waarheid

Vergeleken met een vorige-generatie EDW benadert het CDW+ nog het dichtst het ideaal van het geïntegreerde, geschoonde datawarehouse. De vergelijking gaat echter niet helemaal op. Zo hoeven gegevens uit het CDW niet per se eerst door het CDW+ gehaald te worden voor ze naar een datamart gebracht worden. Slechts als een bewerking van de ruwe data vereist is, wordt een tussenstop in het CDW+ gemaakt. De aanwezigheid van het CDW+ betekent dus bij lange na *geen* verdubbeling van de storage-capaciteit!

Ook bevat het CDW+ niet meer automatisch one single version of the truth. Doordat de opslag van brongegevens en de bewerking ervan uit elkaar zijn getrokken in een CDW en een CDW+, is de single version of the truth niet meer *hard-wired* in het ontwerp van het datawarehouse, maar komen ten aanzien daarvan meerdere alternatieve strategieën in beeld. In plaats van te streven naar één versie (waarbij een switch naar een andere versie altijd mogelijk blijft omdat de ruwe data nog steeds beschikbaar zijn), kunnen ook meerdere gelijkwaardige versies

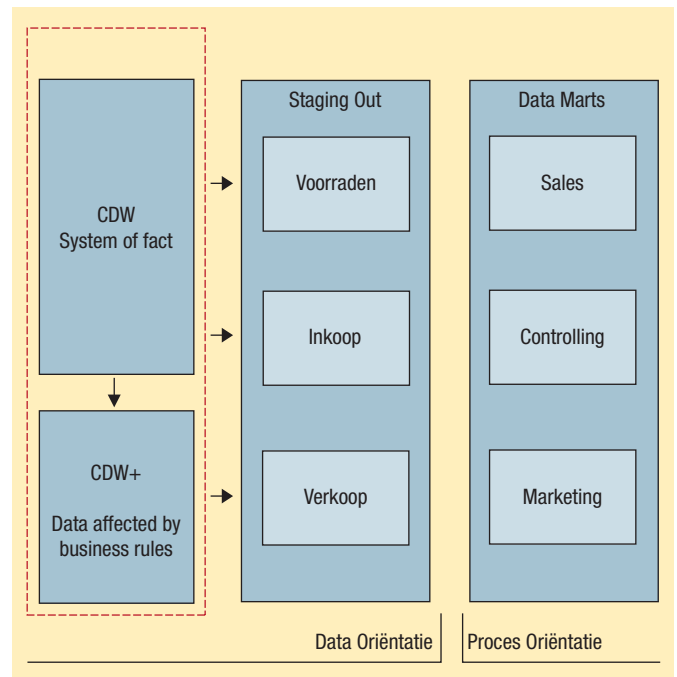
naast elkaar bestaan; of er kan een 'common version of the truth' als enterprise-brede versie naast alle lokale versies ondersteund worden.

Standaardisatie

Doordat zowel voor CDW als CDW+ de Data Vault-modellering wordt toegepast, liggen ze in elkaars verlengde. In de praktijk zal CDW+ vaak niet eens als fysiek aparte omgeving bestaan naast het CDW, maar zal daar simpelweg een uitbreiding op zijn. Zo kan aan een CDW-hub een CDW+-satellite gehangen worden waarin een bepaald afgeleid gegeven bewaard wordt. Met het oog op beheersbaarheid van het voortbrengingsproces, alsmede een strakke domeinscheiding tussen registratie en verrijking omwille van traceerbaarheid, verdient het echter wel aanbeveling om de twee omgevingen logisch gescheiden te houden. Bij de overgang van CDW/CDW+ naar Staging Out is er dus voor de programmatuur geen verschil of iets als CDW- of CDW+-entiteit benoemd is: het maakt bijvoorbeeld voor het opbouwen van een dimensie technisch geen verschil of de dimensiegegevens uit een CDW- of een CDW+-hub gehaald worden – hub is hub. Daardoor kunnen de transformaties van CDW/CDW+ naar Staging Out in hoge mate gestandaardiseerd worden.

Datamarts (STO en DM), aanbod- en vraaggedreven

Zoals is te zien in afbeelding 3 is de datalogistiek en de data-opslag tot en met het CDW aanbodgedreven. Dat wil zeggen dat de bron in grote mate bepaalt wat wordt opgeslagen en op welke wijze het wordt opgeslagen. Na het CDW verandert dit echter: er moet een concrete – door de eindgebruiker geformuleerde – requirement zijn om bouw van datalogistiek en data-opslag na



Afbeelding 4: Staging out en datamarts.

het CDW te rechtvaardigen. Tot en met het CDW+ worden alle gegevens bewerkt en klaargezet. In die zin vormen CDW en het CDW+ tezamen een 'staging-in' area voor het datamart-proces. Zoals in de voorgaande paragrafen is benoemd, is het CDW+ gemodelleerd met de Data Vault-methodiek; de Staging Out en de datamart-laag zijn gemodelleerd om de gekozen end-user tool voor de gebruikersvraag optimaal te bedienen: een query- en rapportage-tool vraagt vaak een dimensionale (Dr. Kimball) opslag, een datamining tool vraagt bij voorkeur een platgeslagen tekstbestand, en een analyse-omgeving floreert met een multi-dimensionale opslagstructuur.

Data- en procesoriëntatie

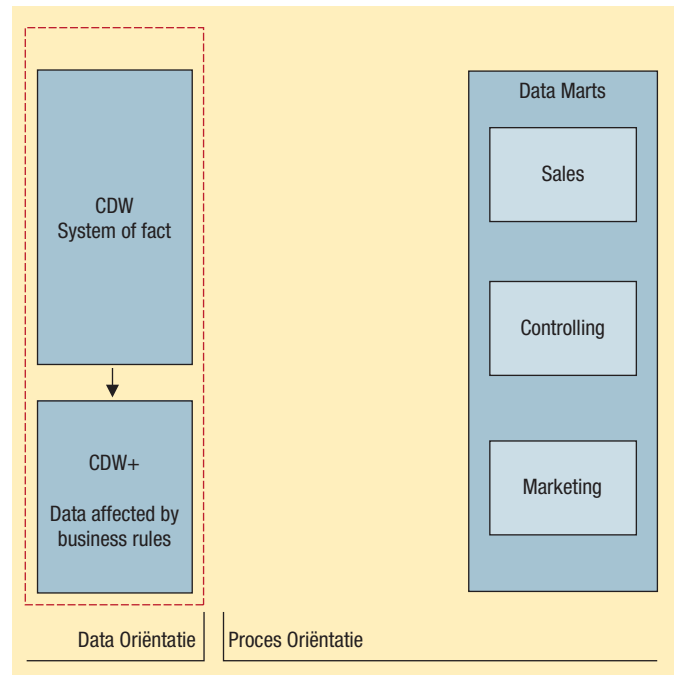
Tot en met de Staging Out worden gegevens datagericht opgeslagen – ongeacht het te ondersteunen business proces. Dit is belangrijk om – zoals eerder beschreven in dit artikel – vervlechting van datamart-structuren te voorkomen. Vanaf de datamart-laag worden gegevens *procesgeoriënteerd* opgeslagen. Hoe de datamart-laag uiteindelijk geïmplementeerd wordt, is afhankelijk van de Business Intelligence tooling en de metadata-laag (bijvoorbeeld de universe in Business Objects) die geboden worden. Bij gebruik van tools met een rijke functionele metadata-laag en geavanceerde codegeneratie (bijvoorbeeld SQL/MDX) hoeft de datamart-laag niet meer te zijn dan de metadata-laag van de betreffende BI-tool. Tools met een arme functionele metadata-laag en dito codegeneratie leiden tot een datamart-laag die geïmplementeerd moet worden in het (R)DBMS. In zo'n geval kan de datamart-laag naar onze mening uitstekend geïmplementeerd worden als een (gematerialiseerde) view-laag op de Staging Out.

Staging Out en datamart-laag zijn gemodelleerd om de gekozen end-user tool optimaal te bedienen

Dit onderscheid in Staging Out (dataoriëntatie) en Datamart (procesoriëntatie) geeft enorme mogelijkheden, waaraan in datawarehouse-architecturen helaas te weinig aandacht wordt geschonken. Afbeelding 4 en afbeelding 5 geven het onderscheid aan.

Uitgaande van drie datageoriënteerde datasets (voorraad-, inkoop- en verkoopgegevens) en drie procesgeoriënteerde datamarts (Sales, Marketing en Controlling) is het volgende mogelijk:

- Er kan relatief eenvoudig een vorm van rolgebaseerde beveiliging worden geïmplementeerd. Stel dat marketing wel inzicht in de marges mag hebben, maar geen inzicht in de netto inkooprijzen per leverancier (ook de leveranciersdimensie mag niet getoond worden aan marketing), dan is dat zeer



Afbeelding 5: Geen staging out – alleen datamarts.

eenvoudig te implementeren door Marketing een eigen (gematerialiseerde) view te geven. In het alternatieve scenario (afbeelding 5) zou er voor Marketing een eigen datamart gemaakt moeten worden via ETL.

- De dreiging van proliferatie van datalogistiek en dataopslag wordt impliciet tegengegaan omdat meerdere views die voor verschillende business processen worden ingezet, op een en dezelfde fysieke datamart geïmplementeerd kunnen worden. Dit in tegenstelling tot het creëren van datamarts met een grote mate van conformiteit, die voor verschillende business processen echter net iets anders opgezet moeten worden. In het voorbeeld van afbeelding 4 is het relatief eenvoudig om met een korte doorlooptijd de (gematerialiseerde) views te maken voor Sales, Marketing en Controlling. In het voorbeeld van afbeelding 5 is dit veel problematischer omdat er datamarts op basis van processen moeten worden gerealiseerd.
- Door het introduceren van deze view-laag introduceren we wederom een ontkoppelpunt, dat er mede voor zorgt dat wijzigingen niet altijd volledig door hoeven te resoneren bij alle gebruikers van de gewijzigde datamart. Het beheer van het EDW, en change management in het bijzonder, varen er wel bij. Als er een feit bijkomt in de inkoopgegevens-dataset dan hoeft deze wijziging – per definitie – alleen doorgevoerd te worden in de ETL-stap die deze dataset vult. In het voorbeeld van afbeelding 5 zou het kunnen zijn dat deze wijzigingen in meerdere ETL-mappingen doorgevoerd moeten worden. Met betrekking tot duurzaamheid heeft de opzet uit afbeelding 4 ook een forse invloed op het testbaar houden van de EDW-architectuur.

Tenslotte wordt de mate van vervlechting tot en met de Staging Out bewust onder strakke (architectuur)controle gehouden.

Zoals in het eerste deel aangegeven blijft duurzaamheid ook in dit opzicht een belangrijk kenmerk van een volgende generatie EDW.

Wat nu als de performance van een view in de datamart-laag onvoldoende is? Verschillend type-gebruik (bijvoorbeeld de ene doelgroep is veel meer aan het ad hoc query'en terwijl de andere doelgroep vooral rapportage doet) kan een fors probleem opleveren voor het tunen van fysieke datasets in de Staging Out. De oplossing in de volgende generatie EDW is relatief eenvoudig; compartimentering. Oftewel het isoleren van een groep gebruikers en de data die zij gebruiken in een fysieke omgeving die beschikt over een grote mate van eigen systeem-resources (lees: CPU, memory, storage). Compartimentering in het kader van de volgende generatie EDW kent een aantal gradaties:

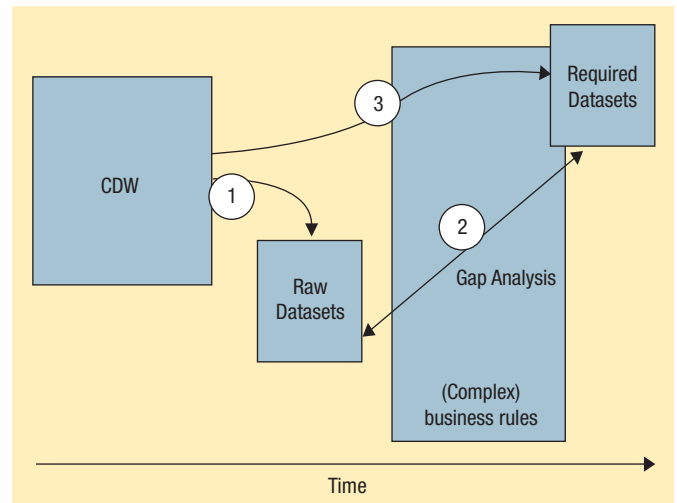
- een datamart view op een of meer fysieke Staging Out datasets materialiseren;
- een datamart view materialiseren en deze fysiek apart zetten op een andere machine;
- misschien wel de mooiste oplossing: de datamart view materialiseren en deze fysiek apart zetten op een *appliance*, een zogenaamde shared-nothing MPP (massively parallel processing) machine (vaak met een kolomgeoriënteerde database), die in staat is om een ongeëvenaarde query performance te bewerkstelligen onder zeer gunstige beheeromstandigheden.

De business bepaalt

Eindgebruikers vinden het vaak lastig om 'out of the blue' hun requirements te formuleren. Een *lesson learned* is dan ook dat het vaak beter is om een ruwe datamart op te zetten en die als startpunt te gebruiken om de requirements (lees: de business rules) nader te verfijnen. Met de data op het scherm is een enorme effectiviteitslag te maken in het bepalen van het uiteindelijke datamart-ontwerp.

Maken van ruwe datamarts geeft goede mogelijkheden om de kwaliteit van de data vroegtijdig vast te stellen

De Data Vault-oriëntatie van CDW en CDW+ biedt fascinerende mogelijkheden om een dergelijke ruwe datamart zeer snel te creëren (stap 1, zie afbeelding 6). Kimball-dimensies bestaan uit een groep van HUB's en bijbehorende SAT's, terwijl Kimball-feiten bestaan uit LINK's en bijbehorende SAT's. Wederom kan er dus bij de opbouw van dergelijke datamarts gebruik worden gemaakt van – in grote mate – gestandaardiseerde (en op termijn gegenereerde) laadmechanismen. Met die ruwe datamart kan – samen met de eindgebruiker – op iteratieve wijze de gewenste datamart worden gerealiseerd (stap 2 en 3, zie afbeelding 6). Dit blijkt een in de praktijk vele malen succesvollere methode



Afbeelding 6: Ruwe datasets versus uiteindelijk gewenste datasets.

vergeleken met het op voorhand proberen te verkrijgen van informatiebehoefte en die dan te realiseren. Het maken van ruwe datamarts geeft bovendien goede mogelijkheden om de kwaliteit van de data vroegtijdig vast te stellen. In dit stadium wordt de gebruiker namelijk voor het eerst geconfronteerd met de data. We stippen nogmaals aan dat de volgende generatie EDW het bepalen van slechte datakwaliteit positioneert *na* het CDW. De eindgebruiker bepaalt deze kwaliteit aan de hand van business rules; bijvoorbeeld de business rule 'als ORDERREGEL geen ARTIKELNR dan fout'. Standaard werkwijze binnen de volgende generatie EDW is dat dergelijke datakwaliteitsregels worden gebruikt om zogenaamde *error-marts* te maken waarop de eindgebruiker kan rapporteren en acteren. Expliciet wordt hiermee nogmaals benadrukt dat datakwaliteit *altijd* de verantwoordelijkheid is van de eindgebruiker (lees: business), zowel voor het bepalen van de datakwaliteitsregels als voor het rapporteren op de *error-marts* en het acteren op geconstateerde kwaliteitsissues.

Conclusie

De black box van de volgende generatie EDW is geopend, en de focus is gelegd op het proces na het Centrale Data Warehouse (CDW). Net als het proces voor het CDW wordt ook het proces na het CDW gekenmerkt door een grote mate van standaardisering, traceerbaarheid, auditeerbaarheid, efficiency en performance. We hopen hiermee invulling te geven aan enkele van de witte vlekken in de aanpak rondom het volgende generatie EDW.

Literatuur

1. R. Kimball & J. Caserta (2004), 'The Data Warehouse ETL Toolkit', Wiley.

Lidwine van As (lidwine@grey-matter.nl) is senior consultant op gebied van informatie-architectuur en -management bij Grey Matter. Met dank aan Ronald Damhof (ronald.damhof@prudenza.nl), senior Information management Architect bij Prudenza.