Bereid uw datawarehouse voor op de toekomst

Datawarehousing op z'n smalst

Jan-Paul Fillié

In het afgelopen decennium zijn bij veel bedrijven in Nederland en Europa datawarehouse-projecten gestart. Bij vrijwel alle grote bedrijven is inmiddels een werkend managementinformatiesysteem aanwezig en voor een select aantal is de rol van dit datawarehouse zelfs cruciaal geworden in de bedrijfsvoering. Gevolg van deze ontwikkeling is dat nu ook de concurrentiepositie van deze bedrijven afhankelijk is geworden van het datawarehouse en met name van de connecties met de buitenwereld. Het valt voor te stellen dat een bedrijf, dat snel en betrouwbaar informatie kan uitwisselen met zijn klanten en leveranciers, een stap voor heeft op een concurrent die nog niet zover is.

Onder druk van nieuwe wetgeving is de rol van het datawarehouse ook aan het veranderen. De informatie die bedrijven moeten verstrekken over met name de financiële huishouding is enorm in omvang toegenomen. Met name herleidbaarheid van de informatie is van belang en betrouwbaarheid en de kwaliteit van de data, die toch altijd al onder druk stonden, moeten hoger worden. Deze 'compliance' stelt nieuwe en veel zwaardere eisen aan de werking van de totale informatievoorziening en de architectuur van het datawarehouse in het bijzonder.

Hoeveel componenten hebben we nog nodig?

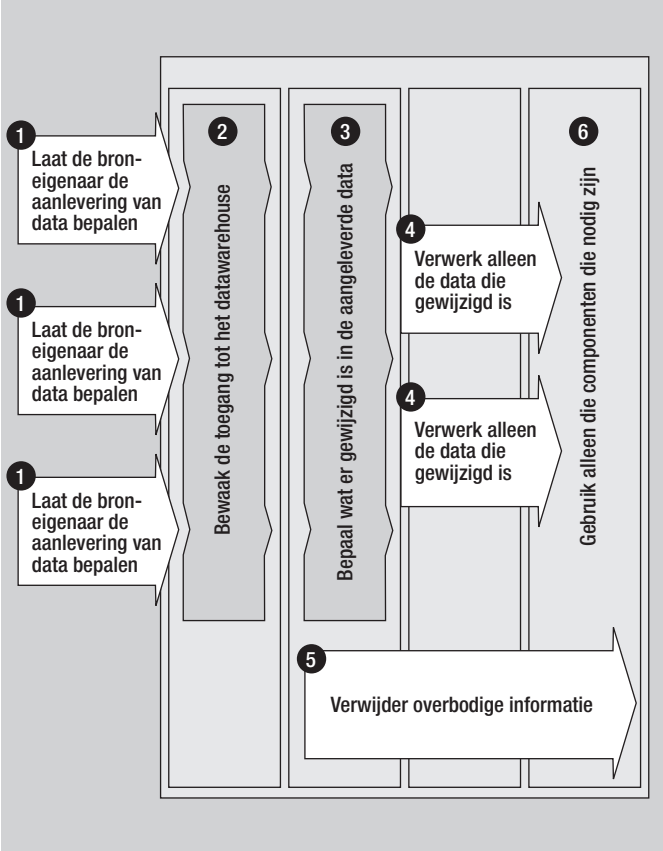
In de huidige datawarehouse-omgevingen kom je veel verschillende architecturen tegen. De combinatie van lagen (componenten van het datawarehouse) varieert van één enkele database tot uitgebreide interfaces per bronsysteem in combinatie met verschillende staging area's, user interfaces, file storage directory's, een ODS (Operational Data Store), een EDWH (Enterprise Data Warehouse-laag) met daarop verschillende business views (gezichtpunten in de vorm van modellen) en tot slot verschillende datamarts. Desondanks kunnen maar weinig van deze datawarehouses goed omgaan met de groei die nodig is om te kunnen blijven voldoen aan de toenemende vraag naar informatie. De kleine datawarehouses kunnen dit niet goed, omdat alle logica in één stap zit, waardoor een enkele aanpassing direct gevolgen heeft voor het geheel. Van de uitgebreide structuren omtbreekt meestal simpelweg het overzicht.
In bijna alle gevallen zie je dat bij toenemende vraag naar functionaliteit, de complexiteit van de ETL (Extraction, Transformation and Loading) programmatuur exponentieel toeneemt. Ook ontstaan bij vrijwel al deze omgevingen op den duur performance-problemen en neemt de kwaliteit van de data af zodra er meer bronnen worden ontsloten in het datawarehouse.
De IT-afdeling die verantwoordelijk is voor het datawarehouse wordt hierop aangesproken door de business. IT zoekt vervolgens de oplossing voor deze problemen in het toevoegen van nieuwe componenten zoals business views of datamarts, waardoor meestal het probleem niet wordt opgelost en de begrijpelijkheid voor de gebruikers alleen maar afneemt.

> In een ERP-applicatie hebben data in tabellen geen betekenis zonder bijbehorende metadata

Zo werd bij een organisatie in de gezondheidszorg telkens een nieuwe laag toegevoegd aan het datawarehouse, om problemen in de data en in de performance op te lossen. Gevolg van deze ingrepen was dat niet alleen het totale proces steeds langzamer werd, maar dat ook de zevende laag na drie jaar ontwikkelen nog niet tot bruikbare informatie heeft geleid. Daarnaast wist niemand meer welke functie werd vervult door welke component.
Belangrijkste gevolg van dergelijke ingrepen is dat de kloof tussen business en IT, die vaak toch al groot is, nog groter wordt.

Tijd voor een nieuwe datawarehouse-architectuur

De oplossing om het datawarehouse geschikt te kunnen maken voor groei is het toepassen van een geschikte architectuur. Op basis van de ervaring opgebouwd met het ontwerpen en bouwen van datawarehouses is binnen Capgemini de architectuur van een nieuwe generatie datawarehouses ontworpen. Deze referentie-architectuur is onderdeel van Capgemini's BI factory.
De BI factory is een geoptimaliseerde software engineering-methodiek voor Business Intelligence en is gericht op versneld oplossingen leveren van betrouwbare kwaliteit.



Afbeelding 1: Toepassingsgebied en -volgorde van de zes basisprincipes in de referentie-architectuur.

De referentie-architectuur (zie afbeelding 1) bestaat uit de volgende basisprincipes:

1. Laat de broneigenaar de aanlevering van data bepalen;
2. Bewaak de toegang tot het datawarehouse;
3. Bepaal wat er gewijzigd is in de aangeleverde data;
4. Verwerk alleen de data die gewijzigd zijn;
5. Verwijder overbodige informatie;
6. Gebruik alleen die componenten die nodig zijn.

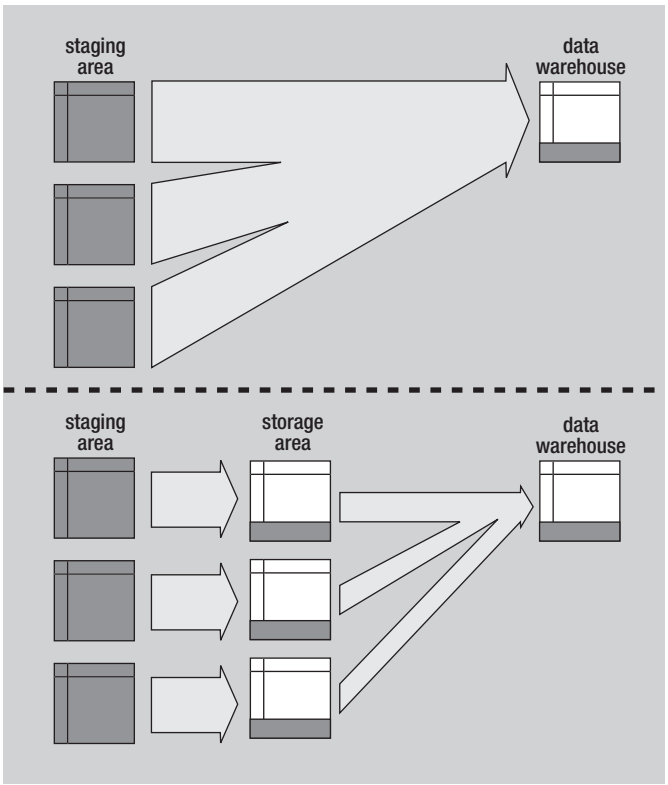
Vernieuwende elementen van deze architectuur zijn onder andere een betere aansluiting op de bronnen, minimalisering van de verwerkingsstroom, gecontroleerd opschonen en het schrappen van overbodige componenten.

Broneigenaar bepaalt aanlevering data

Er zijn verschillende mogelijkheden om de aanlevering van brondata naar het datawarehouse te realiseren, zoals via ETL de bronnen direct benaderen en uitlezen of specificaties voor aanlevering opleggen aan de broneigenaar. Deze veel voorkomende methodes hebben echter grote beperkingen, vooral als het gaat om de kwaliteit van de data en de manier waarop geanticipeerd kan worden welke impact de veranderingen in het bronsysteem zullen hebben. Door het opleggen van specificaties aan de bron, wat in een groot deel van de datawarehouses gebeurt, wordt de beheerder gedwongen om transformatieregels te programmeren.

Bij een Europese bank ging dit zelfs zo ver dat de opgelegde specificaties pas na maandenlange ontwikkelingen konden worden geleverd. Dit is een taak die veel beter door de gespecialiseerde ETL-ontwikkelaars en met de juiste hulpmiddelen kan worden uitgevoerd.
De nieuwe manier is om de broneigenaar te laten bepalen op welke wijze wordt aangeleverd. Zo kan altijd een vorm gekozen worden die gelijk is aan de opslag van het bronsysteem zelf, waardoor veranderingen in het bronsysteem direct kunnen worden vertaald naar de aanlevering. Wel is het verstandig om gebruik te maken van eventueel aanwezige metadata, omdat deze beschrijvende informatie de begrijpelijkheid van de data aanzienlijk kan verbeteren. In een ERP-applicatie als SAP hebben de data in de tabellen geen betekenis zonder de bijbehorende metadata.

Met de eigenaar van de bron moet een OLA (Operational Level Agreement) worden afgesproken, waarin de specificaties van de aanlevering worden bevroren. Bij een verandering in het bronsysteem is het, door de goede aansluiting op zijn bronmodel, voor de eigenaar direct duidelijk welke gevolgen dit heeft voor de aanlevering aan het datawarehouse. Alleen op deze wijze kan het datawarehouse meebewegen met de bronsystemen.Veel te vaak wordt nu nog vanuit het datawarehouse gedacht dat de operationele systemen volledig stabiel zijn. Nu leidt iedere verandering in de bron direct tot enorme data-integriteitsproblemen, omdat een verandering pas wordt gesignaleerd als het kwaad al is geschiedt.



Afbeelding 2: Verschillen tussen een normale verwerking (boven) en volgens het principe van alleen de gewijzigde data (onder).

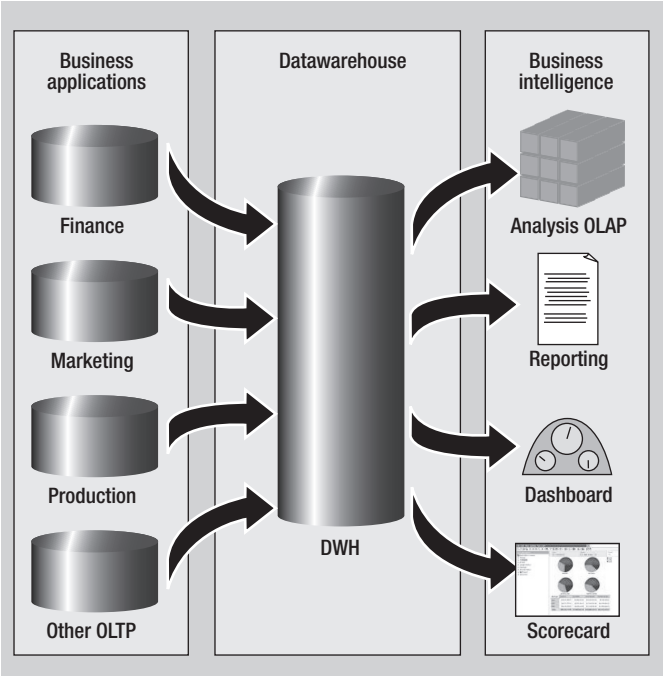
Bewaak de toegang tot het datawarehouse

Na acceptatie van de door de bron aangeleverde specificaties, kan de staging area worden ingericht. Het gebeurt regelmatig dat alleen een deel van de aanlevering wordt overgenomen, omdat alleen dat stuk direct nodig is voor de vereiste functionaliteit. Dit betekent ook dat bij iedere toevoeging of verandering deze selectie weer opnieuw onder de loep dient te worden genomen. In dit voorportaal van de referentie architectuur worden de bronnen één op één en volledig ingelezen om niet alleen flexibiliteit te bieden, maar ook om de beloofde aanlevering te kunnen monitoren. Het is daarom wenselijk om de aanleveringen nauwgezet op de specificaties te controleren in de staging area, want vanaf dit punt is alle data de verantwoordelijkheid van het datawarehouse.

Om extra functionaliteit te kunnen bieden wordt in menig datawarehouse toegang aan gebruikers geboden om de mogelijkheid te geven om referentietabellen bij te werken. Dit zijn meestal gegevens die betrekking hebben op de presentatie, zoals regio- of producthierarchiën. Door foutieve invoer of door een gebrek aan belangstelling, en dus onderhoud, kan deze informatie gemakkelijk ongeldig worden. Als dit gebeurt komt ook de waarde van het datawarehouse ter discussie. Om problemen te voorkomen is het aan te raden om ook de door gebruikers onderhouden referenties te behandelen als een bronsysteem. Alleen op deze manier kan de oorzaak naar de juiste verantwoordelijke worden terug verwezen.

Bepaal wat er gewijzigd is in de aangeleverde data

Bij een ziekenhuis werd dagelijks het datawarehouse opnieuw geladen, zonder dat bekend was welke veranderingen hadden plaatsgevonden. Op het moment dat er een fout optrad in de verwerking kon niet meer worden nagegaan waar dit was opgetreden en wat er tot dan toe was gewijzigd. Het gevolg was dat niet alleen de fout niet kon worden hersteld, maar ook dat het datawarehouse volledig opnieuw moest worden geladen. Een oplossing hiervoor is om de aangeleverde bestanden te vergelijken met de opgeslagen informatie tot dan toe en dit te doen vóór de transformaties.



Afbeelding 3: Basisfunctie van het datawarehouse.

> Laat de business rules alleen los op de gewijzigde records

Dit betekent dat in het datawarehouse een opslag moet zijn gelijk aan het formaat in de bron, alleen op deze wijze is de vergelijking direct en eenvoudig. Stel de verschillen beschikbaar in de storage area als nieuwe, gewijzigde, verwijderde of ongewijzigde items door er een indicatie aan toe te voegen. Deze bijgewerkte storage kan, in tegenstelling tot directe verwerking, vervolgens een totaaloverzicht verschaffen van hetgeen er tot dan toe is geladen in het datawarehouse. Dit geeft tevens een goede recovery-mogelijkheid.

Verwerk alleen de gewijzigde data

In bestaande datawarehouses gaat de verwerking van inkomende data volgens het principe van eerst transformeren naar het juiste formaat en vervolgens dit resultaat vergelijken met de doeltabel. Het nadeel wat hieraan kleeft is dat het volledige bronbestand eerst door een zware bewerking moet voordat deze verrijkte, en dus grotere, stroom kan worden vergeleken. Bij een lease-maatschappij blijkt het dagelijks verwerken van alle data langer dan 24 uur te duren, terwijl men deze informatie dagelijks nodig heeft. In afbeelding 2 wordt een dergelijke verwerking duidelijk gemaakt en vergeleken met de referentie-architectuur.

Doordat in het nieuwe datawarehouse nu bekend is welke wijzigingen hebben plaatsgevonden in de verschillende bronnen, kan er bekeken worden hoe deze informatie verwerkt wordt naar het eigenlijke doelmodel. Dit is in de meeste gevallen een datawarehouse-laag, maar dit kan ook eerst een ODS zijn. Nu moet alleen nog bepaald worden welke afhankelijkheden er bestaan tussen de verschillende bronentiteiten behorende bij een doelentiteit. Werk vervolgens de doelentiteiten in de datawarehouse-laag of ODS-laag één voor één af. Laat de business rules alleen los op de gewijzigde records in de storage area en zorg voor een duidelijke afhandeling, als verschillende bronnen elkaar tegen kunnen spreken of als ze met verschillende frequenties verwerkt worden.

Als de verwerking naar het doelmodel is geslaagd kan dit gemakkelijk herhaald worden voor een opvolgende laag, of dit nu een datawarehouse-laag, een ODS of een datamart is. Door het hele datawarehouse heen worden nu alleen maar de data die strikt noodzakelijk zijn meegenomen in de bewerking. Uit

ervaring blijkt dat hierdoor de totale stroom wordt beperkt tot ongeveer een vijfde deel van een normale volledige verwerking. Bij het bepalen van de gebruikersprofielen van een internationaal telecombedrijf kon via deze logica de verwerking van de verkeersdata teruggebracht worden van meer dan een dag naar binnen het uur. Dit sluit ook volledig aan op de werkwijze van een 'real-time' datawarehouse, waarin iedere individuele wijziging leidt tot een verwerkingsstroom.

Verwijder overbodige informatie

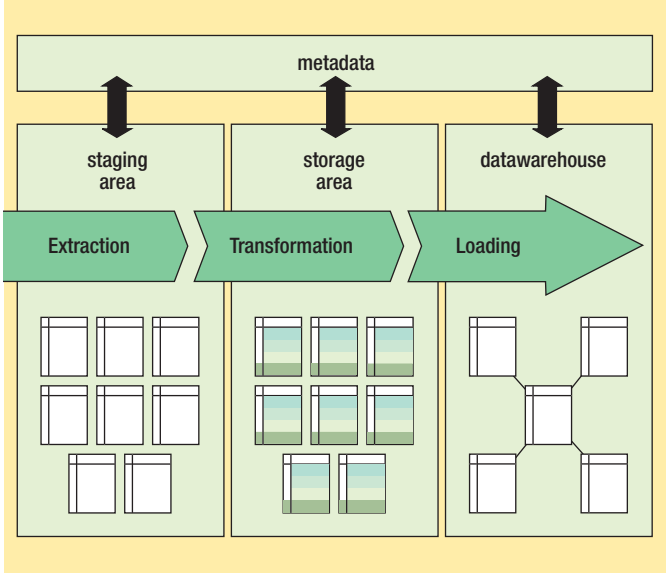
Het toevoegen van informatie is belangrijk voor de waarde van de informatie, maar het verwijderen van verouderde of ongeldige informatie is ook nodig. Verouderde data zorgt niet alleen voor een slechtere performance bij het gebruik van de data, maar leidt ook regelmatig tot foutieve resultaten. Een Nederlandse bank bewaart dertig jaar informatie over betalingsgedrag in het datawarehouse, terwijl voor het doel, de bepaling van het risicoprofiel, zeven jaar voldoende is. Als er al maatregelen worden getroffen dan bestaan deze meestal uit het archiveren van een oud jaar uit de feitentabel. Tijdens deze operatie is het datawarehouse bijna altijd langere tijd niet beschikbaar. Voor de duur van de opslag geldt hetzelfde als voor de verwerking van gewijzigde data: bewaar alleen die informatie die nodig is en verwijs verouderde informatie, in zowel de feiten als dimensies, naar het archief. De kunst is wel om dit met dezelfde frequentie te doen als die van de toevoeging van nieuwe informatie, zodat het gecontroleerd gebeurt en de inhoud van het datawarehouse altijd volledig en actueel blijft.

Gebruik alleen die componenten die nodig zijn

Afbeelding 3 geeft een schematische weergave van de basisfunctie van het datawarehouse. Dit bestaat ten minste uit de centrale ondersteuning van alle BI-toepassingen. Voor deze functionaliteit is niet veel nodig. In afbeelding 4 wordt deze functionaliteit verder uitgediept aan de hand van de verschillende onderdelen ofwel lagen:

- de staging area dient als toegangspoort voor de bronaanleveringen;
- in de storage area worden de verschillen tussen de nieuw aangeleverde bronbestanden eerst bepaald en vervolgens beschikbaar gesteld met de juiste indicaties;
- de datawarehouse-laag bevat het juiste doelmodel voor het beschikbaar stellen van informatie aan de gebruikers van het datawarehouse. Een sterke voorkeur gaat uit naar het dimensionele model door zijn begrijpelijkheid.

Naast deze componenten kan het datawarehouse nog worden uitgebreid met andersoortige componenten voor specifieke functionaliteiten. Een voorbeeld hiervan is het ODS, dat de mogelijkheid biedt aan applicaties om gebruik te maken van gedeelde informatie. Omdat van een dergelijke component wordt gevraagd goed aan te sluiten bij de opslagwijze van operationele systemen (meestal relationeel), zal het datamodel waarschijnlijk



Afbeelding 4: Minimale componenten van de referentie architectuur.

sterk afwijken van bijvoorbeeld een datawarehouse-laag. Ook is er geen historie nodig in het ODS. Gebruik deze componenten dus alleen als er nu (of in de nabije toekomst) behoefte aan bestaat in de organisatie, bijvoorbeeld bij een koppeling naar de CRM-applicatie.

Conclusies

Het is en blijft lastig om het datawarehouse zowel onder controle te houden als er regelmatig nieuwe functionaliteiten aan toe te voegen. Door het datawarehouse te ontdoen van overbodige en nodeloos complexe onderdelen, wordt de onderhoudbaarheid een stuk groter en kan ook de mogelijkheid tot gecontroleerde groei worden geboden. En door alleen datgene te verwerken en te bewaren wat nodig is, kan veel beter worden herleid waar informatie vandaan komt en waar eventuele aanpassingen nodig zijn. Dus voorziet u problemen met de instandhouding of groei van uw datawarehouse, dan doet u er verstandig aan om terug te gaan naar de basis. Zonder een geminimaliseerde en gestroomlijnde architectuur komt snel de grens van mens en techniek in zicht waardoor het datawarehouse zal falen. Door over te gaan naar een vernieuwde architectuur kan niet alleen de functionaliteit van het datawarehouse beter worden overgebracht naar de business, maar kunnen ook de doorlooptijden en dus ook kosten van groei, binnen de perken worden gehouden. U weet direct wat iedere functionele verandering of toevoeging gaat betekenen voor de technische oplossing.

Jan-Paul Fillié

Jan-Paul Fillié (janpaul.fillie@capgemini.com) is consultant op het gebied van Business Intelligence bij Capgemini.