

Enterprise warehouse architectuur

Het definiëren van een informatiearchitectuur in de praktijk.

Inleiding

Over datawarehousing wordt veel geschreven vanuit de toepassing en helaas minder vanuit de architectuur. Een goede informatiearchitectuur is het fundament voor elk informatiesysteem, zo ook voor een datawarehouse. Onderdeel van de informatiearchitectuur is definitie van de wijze waarop het datawarehouse een bijdrage levert aan het realiseren van de doelen die de organisatie zich stelt. Voor Centraal Boekhuis zijn dit externe doelen zoals het leveren van een bijdrage aan de versterking van de keten, maar ook interne doelen, zoals het verbeteren van de bedrijfsprocessen door inzicht te krijgen in zaken als artikelstromen en geleverde diensten.

Het is niet altijd eenvoudig vanuit bedrijfsdoelen de vertaalslag te maken naar het logische deel van de informatiearchitectuur. Toch is het noodzakelijk de logische inrichting van het datawarehouse onder architectuur te definiëren alvorens te starten met de bouw van een datawarehouse. Belangrijke keuzes en afwegingen die vooraf moeten en ook kunnen worden gemaakt leiden tot een gedegen basis voor de langere termijn. De kans dat vervolgens een datawarehousetraject faalt wordt daarmee kleiner.

Informatieprobleem

Binnen een organisatie is er informatie in verschillende afdelingen en op verschillende niveaus. Daarnaast is er op één niveau of afdeling informatie in grote omvang en in verschillende vorm. Hierdoor kan de samenhang tussen de onderlinge informatiebronnen zoekraken. De organisatie is niet meer in staat deze overvloed aan informatie onderling te relateren en te interpreteren, en het wordt daardoor steeds moeilijker de juiste beslissingen te nemen. Informatie implodeert dan als het ware in gegevens, hoewel het tegenovergestelde effect moet worden bereikt. Om dit te voorkomen heeft Centraal Boekhuis de strategische keuze gemaakt voor een enterprise warehouse.

Vaak wordt het begrip datawarehouse gebruikt in de overkoepelende context van **operational datastore**, extractie-, laad- en transformatieprocessen, datawarehouse, datamarts en metadata. Om begripsverwarring te voorkomen gebruiken we voor de overkoepelende context de term **enterprise warehouse**.

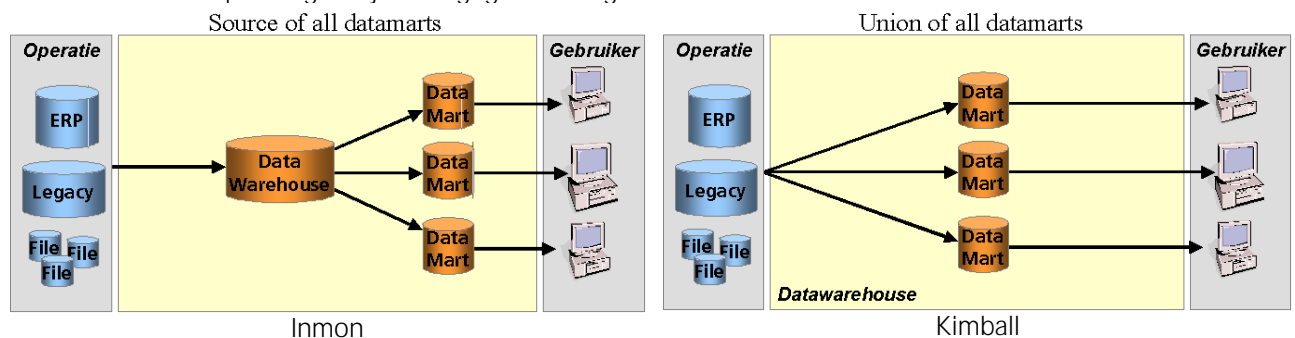
Datamart versus datawarehouse

Datawarehouse en datamart zijn twee basisbegrippen die moeten worden gedefinieerd. Het is makkelijk een definitie over te nemen uit de literatuur. Echter, in de praktijk is het belangrijk een eigen definitie te hanteren. Om een eigen definitie op te stellen zal de theorie moeten worden geraadpleegd. Twee auteurs die naar voren komen zijn Ralph Kimball en Bill Inmon. Zij worden gezien als de goeroes of grondleggers van het kennisgebied datawarehousing. Bill Inmon is 'the godfather' van het begrip datawarehousing en Ralph Kimball is de pionier van het dimensioneel modelleren.

Voor een datawarehouse hanteert Centraal Boekhuis als definitie: een gegevensverzameling voor informatieverstrekking over onderwerpen van de business en is faciliterend naar de gehele organisatie. Een centrale organisatie is hiervan eigenaar.

Een datamart is een gegevensverzameling over een bepaald onderwerp specifiek voor een afdeling (single line of business) voor beslissingsondersteuning van die afdeling. De afdeling is eigenaar.

Vervolgens dient nagedacht te worden over de wijze waarop een datamart en datawarehouse aan elkaar zijn gerelateerd. In de literatuur uit zich dit vooral in het verschil tussen Kimball en Inmon. Kimball ziet een datawarehouse als 'the union of all datamarts' en volgens Inmon is een datawarehouse 'the source of all datamarts'. Deze opvattingen zijn weergegeven in figuur 1.



Figuur 1

Voor Kimball kan worden gekozen als er op eenvoudige en snelle wijze enkele datamarts moeten worden gerealiseerd. Het nadeel van deze keuze is dat overzichtelijkheid en flexibiliteit afnemen bij latere veranderingen. Naarmate het aantal datamarts toeneemt, neemt het overzicht af. Er zijn meer laadprocessen nodig en daardoor wordt het steeds complexer om de gegevens tussen de datamarts onderling consistent te houden. Dit laatste is noodzakelijk om te voorkomen dat de samenhang zoekraakt. Ook bij een wijziging in een datamart moet op consistentie worden gecontroleerd. Een wijziging in de ene datamart kan leiden tot aanpassingen in andere, vooral als een gegeven in meerdere datamarts voorkomt. De flexibiliteit om datamarts te wijzigen neemt daardoor af.

Met de wens verscheidene datamarts te ontwikkelen, heeft Centraal Boekhuis gekozen voor de centrale positie van het datawarehouse. Daarmee is er één centraal laadproces vanuit de operationele systemen naar het datawarehouse. In het datawarehouse worden de bedrijfsgegevens eenduidig, uniform en consistent vastgelegd. Dit wordt ook wel het corporate datamodel genoemd.

Vervolgens kunnen datamarts op redelijk eenvoudige wijze worden gewijzigd of uit het datawarehouse worden ontsloten. Aanpassingen in het primaire laadproces, operationele systemen naar het datawarehouse, of andere datamarts zijn daarbij niet nodig. Binnen Centraal Boekhuis worden momenteel vijf datamarts gebruikt: twee datamarts die via het internet informatie verstrekt aan klanten, en drie datamarts voor de interne analyse en managementinformatie.

Na deze afwegingen te hebben gemaakt, is het noodzakelijk de logische grondslag te definiëren waarop de informatiearchitectuur van zowel het datawarehouse als de datamart berust. De belangrijkste definities die Centraal Boekhuis hierbij hanteert staan in figuur 2.

<i>Datawarehouse</i>	<i>Datamart</i>
<ul style="list-style-type: none"> • Faciliterend naar de gehele business • Gegevensmodel is een reflectie van de business 	<ul style="list-style-type: none"> • Faciliterend naar één single line of business • Gegevensmodel is een reflectie van de vragen die spelen op de afdeling
<ul style="list-style-type: none"> • De data is relationeel gemodelleerd • Doorgaans genormaliseerd • Detailinformatie • Minimaal geïndexeerd • Omgeving voor professionals • Stabiel qua verandering 	<ul style="list-style-type: none"> • De data is dimensioneel gemodelleerd • Doorgaans gedegenormaliseerd • Geaggregeerde informatie • Maximaal geïndexeerd • Omgeving voor gebruikers • Flexibel qua verandering

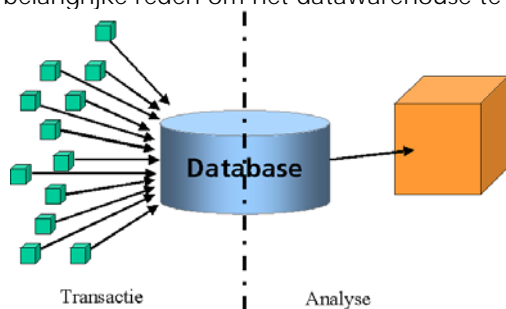
Figuur 2

Elke verandering binnen het enterprise warehouse moet in lijn zijn met de uitgangspunten die in de architectuur zijn gedefinieerd. Een afwijking hierop moet goed worden overwogen en eenduidig worden vastgelegd. Hierin zit namelijk op langere termijn precies het verschil tussen de vruchten plukken of de prijs betalen.

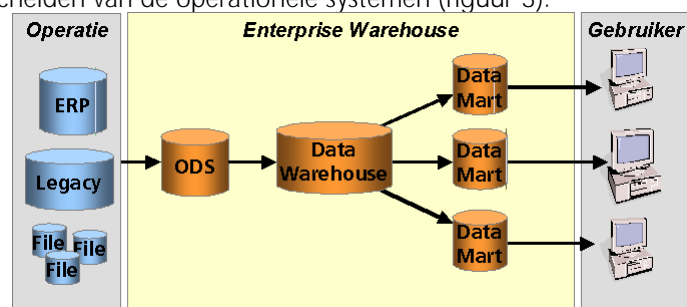
Gegevensstroom naar en binnen het enterprise warehouse

Gegevensextractie

Nadat datawarehouse en datamart zijn onderscheiden, kan nagedacht worden over de wijze waarop beide omgevingen met gegevens worden gevuld. Er zal periodiek een bepaalde extractie moeten plaatsvinden van gegevens uit de operationele systemen die in het datawarehouse worden geplaatst. Echter, de operationele systemen en het datawarehouse zijn twee gescheiden werelden, waarbinnen andere wetten gelden. Zo zijn operationele systemen gericht op uitvoering van transacties. Vele wijzigingen op recordniveau worden achter elkaar in de database verwerkt. Een datawarehouse is gericht op uitvoering van analyses. Hiervoor worden vele records in één keer uit de database gehaald. Dit is een belangrijke reden om het datawarehouse te scheiden van de operationele systemen (figuur 3).



Figuur 3



Figuur 4

Om beide omgevingen onafhankelijk van elkaar te kunnen laten opereren wordt er gebruik gemaakt van een **operational datastore** (ods). Dit is een tijdelijke ruimte waarin de operationele systemen gegevens klaar zetten waarmee vervolgens het datawarehouse aan de slag kan (figuur 4). Het voordeel hiervan is dat de processen waarmee het ods wordt gevuld kunnen worden afgestemd op de regels die gelden voor de operationele systemen en de processen voor het vullen van het datawarehouse op de regels binnen het enterprise warehouse.

Transformatie en laden van gegevens

Van een bepaald soort gegeven dat vanuit de operationele systemen in het ods wordt geplaatst, hoeft de definitie niet altijd bedrijfsbreed overeen te komen. Zo wordt binnen verschillende afdelingen van Centraal Boekhuis het begrip 'afzet' in verschillende context gebruikt. De afdeling Logistiek praat over afzet als afzet van boeken die fysiek vanuit de magazijnen van Centraal Boekhuis zijn gedistribueerd, de afdeling Financiën praat over afzet als de boeken die zijn gefactureerd. In het enterprise warehouse wordt onderscheid gemaakt tussen fysieke en administratieve afzet.

Alle afwijkingen moeten in kaart worden gebracht en uniform worden gedefinieerd. Daarbij is het noodzaak dat de definities door de gehele organisatie worden gedragen. De gegevens in het ods worden conform de definities getransformeerd en geladen in het datawarehouse.

Een tweede vorm van transformatie vindt plaats tijdens het laden van gegevens vanuit het relationele corporate datamodel in het datawarehouse naar het dimensionele vraagmodel in de datamart. Tijdens deze transformatie vinden de extra berekeningen, aggregaties en accumulaties plaats.

Datacleaning

In elk transactioneel systeem komen records voor die niet kloppen. Dit kunnen dubbele records zijn of records die geen presentatie zijn van de werkelijkheid. Dit manifesteert zich vooral bij stamgegevens zoals relatie- en artikelgegevens. Binnen de operationele systemen hoeft dit niet direct tot problemen te leiden. Echter, in het enterprise warehouse kan dit soort 'vuile' data grote gevolgen hebben. Bijvoorbeeld op een rapport waar de afzet per type klant wordt weergegeven, kan een verkeerde beslissing worden genomen als het merendeel van de klanten in het verkeerde type is ingedeeld.

Datacleaning is het proces waarbij data worden geschoond. Dit is een complex, tijdrovend onderdeel; het neemt daardoor de nodige kosten met zich mee. Toch is het niet noodzakelijk om een honderd procent schoon enterprise warehouse na te streven. Het is belangrijker om classificaties, bijvoorbeeld klanttype, op orde te hebben dan detailgegevens. De focus moet daarbij liggen op de classificaties met de hoogste informatiewaarde.

Lessons learned

Informatiearchitectuur is het definiëren, structureel weergeven en beschrijven van de verschillende onderwerpen die betrekking hebben op het enterprise warehouse. Tevens worden daarbij de onderlinge afhankelijkheden en samenhang tussen de onderwerpen weergegeven. Het is een raamwerk waarmee veranderingen kunnen worden beheerd en beslissingen kunnen worden genomen. Dit raamwerk levert een belangrijke bijdrage aan minimalisering van complexiteit, behoud van overzicht, reductie van kosten en de tevredenheid van de gebruiker.

Maar de aanwezigheid van een informatiearchitectuur wil niet automatisch zeggen dat het eenvoudig is een enterprise warehouse te realiseren. Een belangrijke basis is gelegd, maar er zullen projecten moeten worden gedefinieerd om het enterprise warehouse te bouwen.

Je weet niet wat je wil weten

Geen enkele gebruiker binnen de organisatie weet vooraf wat hij achteraf wil weten. Afhankelijk van beschikbare informatie en inzichten zullen altijd weer nieuwe vragen ontstaan. Het is onmogelijk om het hele scala aan vragen binnen een organisatie in kaart te brengen. De focus moet liggen op de belangrijkste vragen. Deze sluiten aan bij de onderwerpen waar het management over praat. Notulen van een managementoverleg zijn goede bronnen ter verificatie.

Het is belangrijk het doel van een enterprise warehouse nooit uit het oog te verliezen, namelijk ondersteuning van de organisatie om doelen te realiseren. Om doelen te realiseren moeten beslissingen worden genomen, gebaseerd op waardevolle informatie. Een gebruiker wil de beschikking hebben over alle gegevens, tot in het laagste detail. Bij elke vorm van informatie moet de analist zich afvragen welke business beslissing aan de hand van de informatie genomen kan worden. Is hierop door de organisatie geen antwoord te geven, dan is de informatie waarschijnlijk niets meer dan een gegeven. Het is dan de vraag of het gegeven in het enterprise warehouse moet worden opgenomen.

Bestemmingsplan of blauwdruk

Realisatie van een enterprise warehouse is geen doel op zich. Het is een middel om doelen te realiseren die de organisatie zich stelt. Datawarehousing is een continu dynamisch proces binnen de organisatie. Het stopt niet zodra het eerste project is opgeleverd. Inzichten van een organisatie zijn aan verandering onderhevig. Het is nagenoeg onmogelijk vooraf een blauwdruk te definiëren van de inhoud van het volledige enterprise warehouse. Het is beter te werken vanuit een bestemmingsplan. Hiermee kunnen kleine overzichtelijke projecten worden gestart waarmee een bepaald deel van het bestemmingsplan wordt gerealiseerd. De kreet 'think big, act small' is hier zeker van toepassing.

Het is niet verstandig eerst het complete datawarehouse te bouwen alvorens een datamart te realiseren. Waarschijnlijk is tegen de tijd dat het datawarehouse gereed is de geldkraan allang dichtgedraaid. Het is beter om samen met de organisatie één onderwerp te kiezen en dit verder uit te werken tot en met de realisatie van één datamart. Dit alles moet wel worden uitgevoerd volgens de regels der architectuur. Zo wordt binnen Centraal Boekhuis het datawarehouse incrementeel gebouwd. Afhankelijk van de behoefte van de organisatie worden nieuwe entiteiten aan het corporate datamodel toegevoegd. De afdeling Architectuur ziet erop toe dat iedere nieuw entiteit voldoet aan de eisen van het corporate datamodel.

Gebruikers

Veel datawarehouse- of datamartleveranciers schrijven over de mooiste toepassingen van **on-line analytical processing** (olap). Echter, veel gebruikers van een enterprise warehouse zijn geen olap'ers. Zij zijn al tevreden als ze eenvoudig een overzichtelijk rapport kunnen raadplegen. Zo is er tijdens één van de datamartprojecten specifiek voor gekozen om geen gebruik te maken van een standaard olap-tool. Zo'n tool bood zoveel opties dat de op te bouwen kennis in geen verhouding stond tot het kleine deel wat ervan zou worden gebruikt. Er is daarom gekozen om de rapporten zelf te bouwen met aanwezige technologische kennis. Dit zonder afbreuk te doen aan de geboden functionaliteit, en naar tevredenheid van de gebruikers.

De kosten

De meeste kosten van een enterprise-warehouseproject bij Centraal Boekhuis liggen bij het vullen van het datawarehouse. Tijdens deze fase komen complexe onderwerpen aan bod, zoals het corporate datamodel en de inrichting van de extractie-, transformatie- en laadprocessen. Het is verleidelijk om hieraan weinig aandacht te besteden omdat de organisatie wacht op de realisatie van haar datamart. Er bestaat daardoor een grote kans in de eigen valkuil te trappen. In het datawarehouse worden dan beslissingen genomen op basis van de wens van de afdeling waarvoor de datamart wordt ontwikkeld en niet op basis van het belang van de gehele organisatie. Dit druist in tegen de uitgangspunten zoals die zijn gedefinieerd in de informatiearchitectuur.

Het resultaat van zulke beslissingen zal niet direct zichtbaar zijn, maar zal zich uiten tijdens het onderhoud. Vroeg of laat zullen de afdelingspecifieke elementen in het datawarehouse moeten worden omgezet naar uniforme elementen. Investeren in het goed vullen van het datawarehouse verdient zich automatisch terug tijdens creatie van de datamarts. De ervaring leert Centraal Boekhuis, dat hoe beter het datawarehouse is ingericht, hoe makkelijker datamarts eruit kunnen worden ontsloten.

Tevens ervaren we binnen Centraal Boekhuis dat de kosten die worden gemaakt tijdens onderhoud, afhankelijk zijn van de systeemarchitectuur van de operationele systemen. Binnen Centraal Boekhuis is de informatievoorziening merendeel in eigen beheer en vooral onder architectuur ontwikkeld. Twee belangrijke uitgangspunten van de systeemarchitectuur zijn generieke opzet en minimale dataredundantie. Ook de keuze voor één databasemanagementsysteem speelt erbij een belangrijke rol. Het gevolg is dat wijzigingen binnen de operationele systemen nauwelijks leiden tot wijzigingen in het datawarehouse. De onderhoudskosten op het datawarehouse zijn dan ook lager dan die voor de operationele systemen.

Conclusie

Een enterprise warehouse is een middel om de organisatie te verbeteren en de doelen die de organisatie zich stelt te realiseren. De beschikbaarheid van eenduidige, consistente en waardevolle informatie speelt daarbij een cruciale rol. Het is belangrijk om in het bezit te zijn van een architectuur om het enterprise warehouse in lijn te houden met de informatiebehoefte van de organisatie. Daarbij is het vooral raadzaam klein te beginnen en het enterprise warehouse incrementeel op te bouwen.

Iedere architectuur begint bij de definitie en een beschrijving van een aantal essentiële onderwerpen op conceptueel niveau. Vervolgens worden de onderwerpen met elkaar in relatie gebracht en wordt de samenhang beschreven. Het is daarbij zeer belangrijk om niet te ver in detail te treden, om het overzicht te behouden. Binnen de architectuur wordt de focus gelegd op de onderwerpen die belangrijk zijn voor de organisatie. Door deze onderwerpen te projecteren op de praktijk, ontstaat snel een beeld hoe de architectuur er op hoofdlijnen uit moet komen te zien. Architectuur moet niet alleen theoretisch blijven maar toepasbaar zijn in de praktijk. Door architectuur te benaderen vanuit de praktijk, wordt het tastbaar en overzichtelijk. Daarmee bewijst het de toegevoegde waarde voor de organisatie.

Emiel van Bockel

IT-Architect Business Intelligence & Datawarehousing
Centraal Boekhuis
E-mail: e.van.bockel@centraal.boekhuis.nl.

Literatuur

www.dmreview.com
www.inmoncif.com
www.zifa.com

Literatuur

Inmon, W.H., J. D. Welch & K.L. Glassey (1997). *Managing the Data Warehouse*. New York: John Wiley & Sons, Inc.
Kimball, R. (1996). *The Data Warehouse Toolkit*. New York: John Wiley & Sons, Inc.
Gallas, S. (1999). Kimball vs Inmon. In: *DM Review issue september 1999*.
Davenport, T.H., & L. Prusak (2000). *Working Knowledge: how organisations manage what they know*. Boston: Harvard Business School Press.
Nonaka, I., & H. Takeuchi (2003). *De kenniscreërende onderneming: Hoe Japanse bedrijven innovatieprocessen in gang zetten*, Schiedam: Scriptum management.
Porter, M. (2000). *Concurrentie voordeel*. Amsterdam: Contact