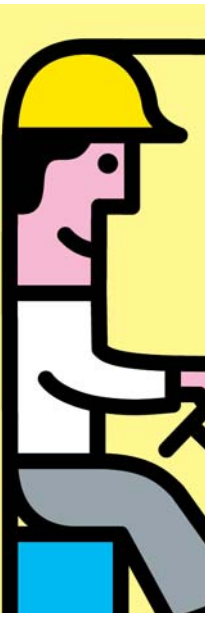


Drielagen-architectuur Kadenza in lijn met eigentijdse vereisten

Datawarehousing kan sneller en flexibeler

Marianne Kompagne

Organisaties willen snel kunnen beschikken over de juiste informatie ter ondersteuning van hun strategische doelen. Daarvoor is een eigentijdse datawarehouse-architectuur nodig. Kadenza ontwikkelde een DWH-referentiemodel dat vele voordelen biedt ten opzichte van de oplossingen van Kimball en Inmon.

De referentiearchitectuur van Kadenza is opgebouwd uit drie onderdelen: de stage area, de datawarehouse area en de informatie area. Per specifieke laag wordt de invulling van de verschillende functionaliteiten beschreven.
In de stage area laag kunnen gegevens uit interne en externe bronnen worden ontvangen, zowel synchroon als asynchroon. Verder wordt bijgehouden of het om nieuwe of gewijzigde data gaat. De stage area voedt het datawarehouse en mogelijk de Operational Data Store. Om correcte, eenduidige en consistente gegevens door te geven, vindt bewerking, integratie en schoning van de gegevens plaats.
In het datamodel in deze laag worden gegevens uit verschillende bronnen geïntegreerd (met onderlinge relatie) weergegeven. Van stamgegevens wordt een actuele versie vastgehouden. De andere gegevens hebben een beperkte historie: er zijn transacties uit het verleden, maar er vindt geen historische stapeling plaats. Alle gegevens van primaire bedrijfsprocessen worden vastgelegd, tenzij aangegeven is dat deze niet gewenst zijn. Dit datamodel is incrementeel uitbreidbaar. De gegevens in de staging area zijn niet toegankelijk voor eindgebruikers.

De datawarehouse area laag is bedoeld om gegevens voor langere tijd vast te houden. Het kan gezien worden als het geheugen voor lange termijn en beschikt over een archieffunctie: gegevens worden gestapeld opgeslagen. Het is bedoeld om de datamart laag van geaggregeerde feiten en dimensies te voorzien. Dit dient herhaaldelijk en met verschillende mate van detaillering te gebeuren. Uit deze laag kan ook ondersteuning voor data mining en ad hoc analyse plaatsvinden.
Het datamodel van de datawarehouse area is een bedrijfsbreed informatiemodel dat incrementeel uitbreidbaar is. Doordat het geënt is op informatie en niet op transacties, is het model toegankelijk en snel. Alle gegevens van primaire bedrijfsprocessen met hun historie (ook dimensies) worden in deze laag vastgelegd, tenzij aangegeven is dat deze niet gewenst zijn.
De informatie area laag is bedoeld om de kennisintensieve processen van de organisatie te ondersteunen. Het ontwerp van de datamart laag is gebaseerd op de informatiebehoefte van een bepaald bedrijfsonderdeel. Het dient ondersteuning te bieden aan voorgedefinieerde business-vragen, gebaseerd op *measurable facts* (KPI's en/of CSF's) afgeleid van de business-processen van de organisatie.
De data in de datamart betreffen een andere representatie van de historische laag, weergegeven in een model dat begrijpelijk is voor eindgebruikers. Samen met performance voldoet het dan aan de belangrijkste vereisten. De datamarts zijn (relatief) eenvoudig te wijzigen, op te bouwen of incrementeel uit te breiden. Deze laag is toegankelijk voor eindgebruikers.

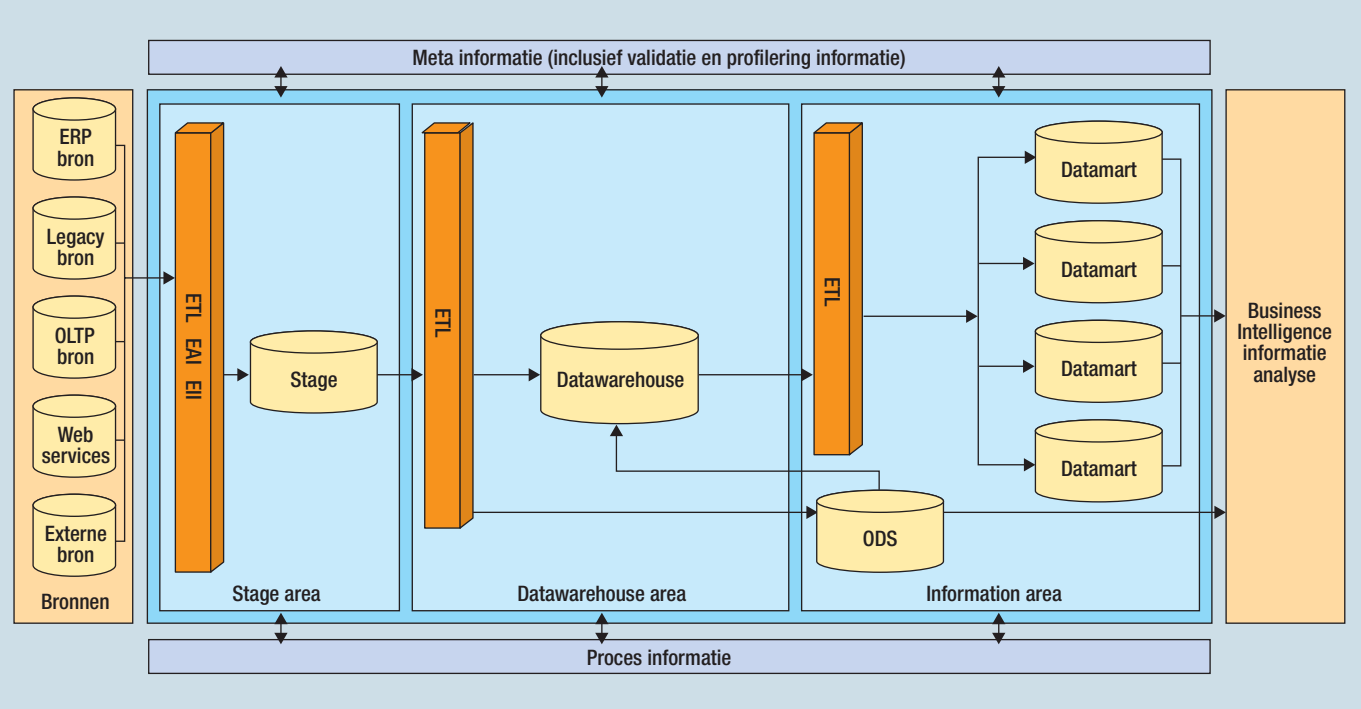
Aanvullende vereisten

De Operational Data Store is bedoeld om gegevens uit verschillende externe en interne bronnen op een geïntegreerde wijze weer te geven en te bevragen. De gegevens worden met een zo klein mogelijke vertraging aan de eindgebruikers ter beschikking gesteld. Gegevens uit verschillende bronnen worden geïntegreerd weergegeven in voornamelijk 3NF.
De procesinformatie, informatie voor de centrale administratie en registratie van procesgegevens, dient om de processen toegankelijk en inzichtelijk te maken. Zo kan bijvoorbeeld de bron getraceerd worden. Hier worden ook uitzonderingen en fouten afgehandeld.
Metadata worden niet alleen gebruikt ter informatie van de beheerder en bij het ETL-proces, ze worden ook toegepast voor de gebruikers van het datawarehouse. Tijdens het opslagproces worden data opgeslagen na validatie en profilering. Belangrijk daarbij zijn natuurlijk compleetheid, correctheid en accuraatheid.

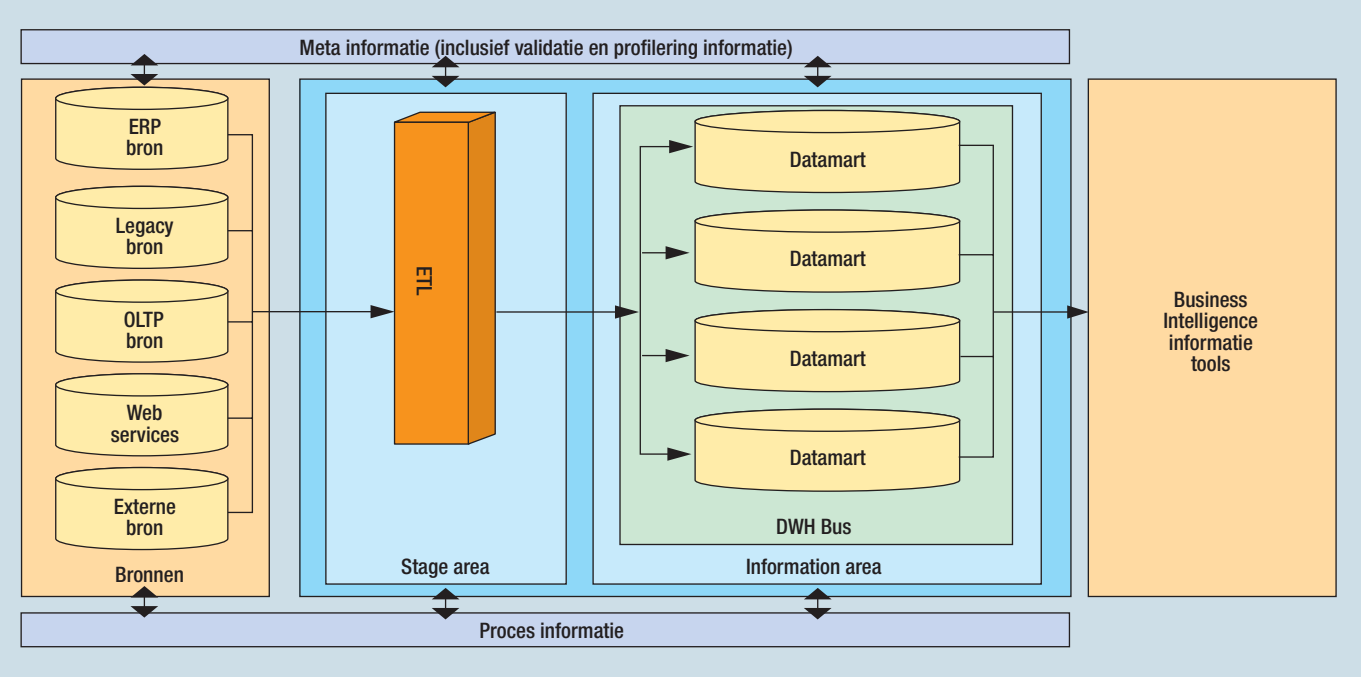
Wanneer we datawarehouse-architecturen vergelijken met de gebruikelijke aanpak volgens Kimball en Inmon, wordt allereerst gekeken naar de ondersteuning van strategische bedrijfsdoelen en informatievereisten. Het datawarehouse ondersteunt het besluitvormings- en informatieproces op operationeel, tactisch en strategisch niveau. Daarnaast ondersteunt het een bedrijf vanuit verschillende perspectieven en over alle bedrijfsonderdelen.
Vanuit een bedrijfsplan (strategisch) en de performance indicatoren (KPI's) wordt immers de basis gelegd voor een business informatiemodel en het definitieve ontwerp van een datawarehouse-architectuur.
De Kadenza DWH referentiearchitectuur verschilt met die van Kimball en Inmon in de laag waar de historisch correcte informatie bewaard wordt. Bij Kimball is dit de datawarehouse bus, terwijl Inmon de corporate information factory benadering heeft. In de referentiearchitectuur gaan we uit van een separate datawarehouse laag met informatie opgeslagen in een bedrijfsinformatie datamodel. Deze laag is uitbreidbaar voor toekomstige informatiebehoeften en is geïsoleerd van de informatievoorziening aan eindgebruikers. Vanuit deze historisch correcte informatie kunnen dan de datamarts (eindgebruiker informatie) afgeleid worden.
Er ontstaat een flexibele informatievoorziening doordat ingespeeld kan worden op wijzigingen in de informatiebehoefte. Het is bijvoorbeeld mogelijk het detailniveau van de gegevens en de datamarts (nieuwe informatiebehoefte) te wijzigen en uit te breiden. Enerzijds voorkomen we hiermee dat het detailniveau van de informatie bij het ontwerpen van de datamarts wordt bepaald en dat de historie van informatie in de datamarts plaatsvindt en daarmee de detailgradatie van die historie. Anderzijds werken we zo het enterprise datamodel nader uit, waardoor dat niet langer onderhevig is aan interpretatie. Dit in tegenstelling tot het enterprise datamodel dat Inmon gebruikt; deze is niet uitgewerkt en daarom wel onderhevig aan interpretatie.
In een datawarehouse dient de information area (datamart) eenvoudig toegankelijk te zijn voor eindgebruikers. Het datamodel of de structuur dient ook begrijpelijk te zijn voor de eindgebruiker, daarnaast dient het een toegankelijke interface te hebben. Het is bijvoorbeeld mogelijk standaard software te gebruiken voor rapportage en interactieve analyse (OLAP). Tot slot moet de gemiddelde doorlooptijd acceptabel zijn voor de eindgebruiker.
Deze uitgangspunten van de Kimball benadering worden alom onderschreven, dus ook in de beschreven referentiearchitectuur.
De bedrijfsinformatie is opgeslagen in datamarts. Deze datamarts bevatten de informatie van verschillende bedrijfsprocessen vanuit verschillende invalshoeken (conformed dimensions). Echter de toegankelijkheid van informatie wordt ook bepaald door de flexibiliteit bij het definiëren en wijzigen (bijvoorbeeld het detailniveau) van door eindgebruikers gewenste datamarts. Het is dan mogelijk informatie snel beschikbaar te stellen.

Analysemogelijkheden

Een datawarehouse dient data-analyses te ondersteunen voor bijvoorbeeld data mining. Hiervoor moeten gegevens beschikbaar zijn op een gewenst gedetailleerd niveau voor eindgebruikers. Ook is het belangrijk dat de informatiebehoefte gebaseerd is op een bedrijfsbreed begrippenkader en begrippenstructuur en willen we analysemogelijkheden hebben op historische data. Daarnaast moet de betekenis van de gegevens zowel duidelijk en begrijpelijk zijn, als eenduidig en consistent. Ook is het van belang dat er met geschoonde gegevens gewerkt wordt.



Afbeelding 1: DWH referentiemodel van Kadenza.



Afbeelding 2: De Kimball architectuur

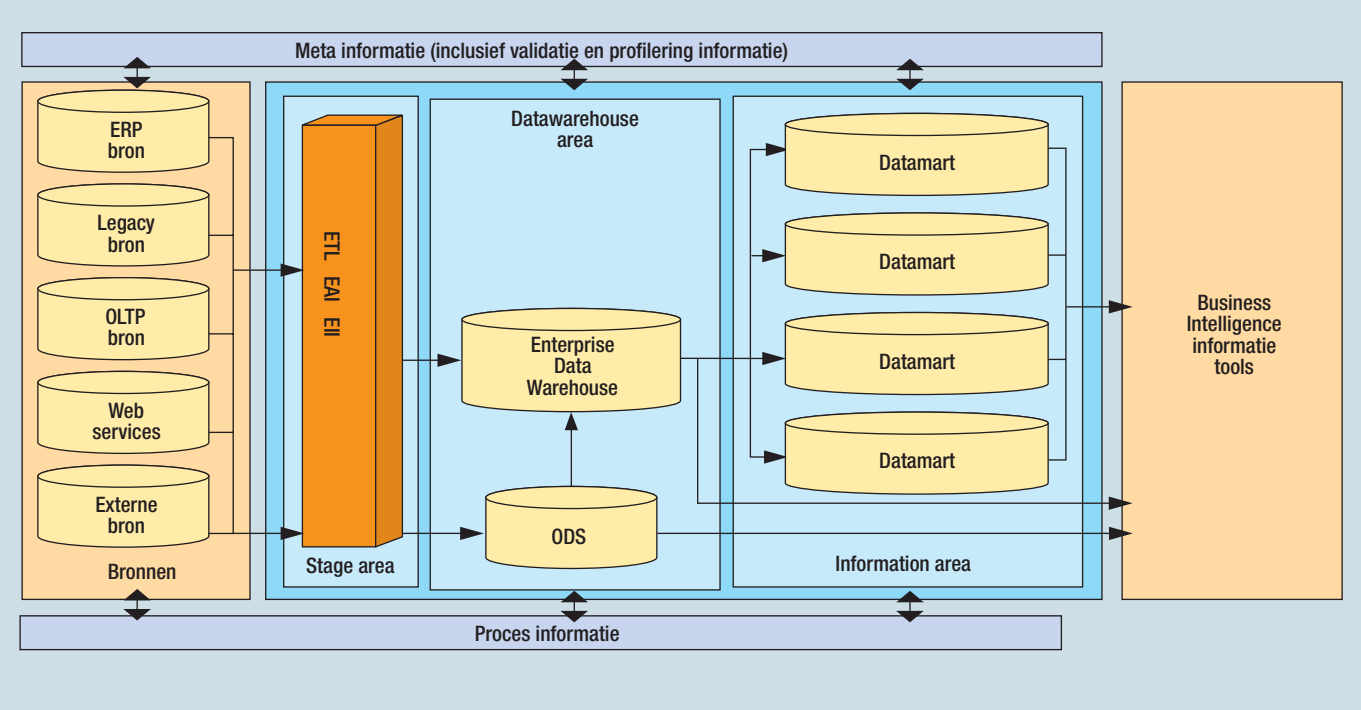
Dit geldt voor alle datawarehouse-architectuur benaderingen. Echter, er zal tijdens het opslagproces bepaald moeten worden wat de betekenis van de gegevens is en hoe en in welke mate gegevens geschoond gaan worden. Hoe beter een organisatie hierop een antwoord kan geven, des te sneller komt informatie beschikbaar voor eindgebruikers.
Een datawarehouse dient bedrijfsbrede behoeften aan operationele data te ondersteunen. Zo moet het datamodel toegankelijk en eenvoudig te begrijpen zijn en de geïntegreerde operationele bedrijfsvoering reflecteren. Tevens moet, indien gewenst *near real-time*, actuele informatie beschikbaar zijn. Ook moet de opgeslagen informatie een betere kwaliteit hebben dan de operationele systemen.

Vergelijkingen

Alle datawarehouse-architecturen hebben een zogenaamde staging laag (stage area). Hier vindt op basis van de bronmodellen extractie van gegevens plaats. De structuur waarin de gegevens in de staging laag worden opgeslagen, kan per datawarehouse-architectuur echter verschillen. Voor het ontwerp van het extractieproces (ETL/EAI/EII) is de frequentie van extractie ofwel data-aanlevering (constante flow of periodieke aanlevering), de extractiemethode (bijvoorbeeld alleen mutaties) en de technologie (zoals webservices, datareplicatie, change data capture) bepalend. In de staging laag vinden tevens databewerkingen (transformatie), gegevenskwaliteitcontrole en schoning van gegevens plaats. De functie van de staging laag is immers het klaarzetten van informatie, om deze vervolgens (historisch) correct op te slaan.
In de referentiearchitectuur kunnen de gegevens behalve naar een datawarehouse laag (fungeert als informatie-archief), ook naar een zogenaamde Operationele Data Store (ODS) geladen worden. De ODS is door Inmon aangedragen en heeft in de referentiearchitectuur dezelfde functie: het bedienen van een operationele informatiebehoefte en het voorzien van operationele processen van informatie. In de referentiearchitectuur verloopt de staging volgens een gedefinieerde methodiek. Deze methodiek is zo gekozen dat bronnen sneller toegevoegd en gewijzigd kunnen worden. Kimball en Inmon dragen hier wel ideeën over aan maar leveren geen methodiek.
Een datawarehouse dient relatief eenvoudig te onderhouden zijn. Dat is belangrijk voor het snel realiseren van wijzigingen van eindgebruikerwensen. Daarnaast willen we de mogelijkheid hebben om de gedetailleerdheid van gegevens voor een eindgebruiker te wijzigen. Ook is het van belang bij het efficiënt en snel structureren en onderhouden van geaggregeerde data, wijzigingen van bronsystemen op te vangen en voor het efficiënt onderhouden van extractie- en transformatieprocessen. Tevens is het belangrijk om een minimale impact te hebben bij veranderingen en isolatie tussen verschillende architectuurlagen om verstrengeling van functionaliteit te voorkomen.

Flexibiliteit

In tegenstelling tot de architecturen van Kimball en Inmon biedt de gelaagde referentiearchitectuur flexibiliteit. Door de isolatie van de verschillende lagen kan de datawarehouse laag met het informatiearchief ongemoeid gelaten worden, terwijl er in de andere lagen wijzigingen doorgevoerd worden. Wijzigingen in bronsystemen kunnen snel en flexibel in de staging laag opgevangen worden en wijzigingen in eindgebruikerwensen in de informatielaag. In combinatie met een standaard schaalbare methodiek voor extractie- en transformatieprocessen, vormt dit een relatief eenvoudig te onderhouden datawarehouse-oplossing.



Afbeelding 3: De Corporate Information Factory van Inmon.

Informatie kan sneller ontsloten en beschikbaar gesteld worden. Daarbij is het wenselijk dat een datawarehouse eenvoudig uit te breiden is om snel functionaliteit te realiseren behorende bij (voor het datawarehouse) nieuwe bedrijfsonderdelen en snel data toe te voegen afkomstig uit (voor het datawarehouse) nieuwe bronnen. Dit kan nog worden versterkt door een uitbreidbaar informatie/datamodel in de datawarehouse laag van de referentie-architectuur.

> Vanuit de historisch correcte informatie kunnen de datamarts afgeleid worden

Op het gebied van schaalbaarheid moet een datawarehouse-infrastructuur kunnen voldoen aan de wens om datagroei (meer bronnen, meer historie) en de groei van eindgebruikeractiviteiten op te vangen. Daarnaast is schaalbaarheid van belang om aan wensen voor hogere performance te voldoen en doorlooptijden te verkorten. Dit geldt voor alle datawarehouse-architecturen, maar is in tegenstelling tot de andere vereisten meer een technische infrastructuur vereiste. Een gelaagde architectuur is op technisch niveau eenvoudiger te ondersteunen.
Een datawarehouse dient vanuit een operationeel gezichtspunt eenvoudig te hanteren te zijn op het gebied van systeembeheer-onderdelen, job scheduling en load balancing tools. Daarnaast is een goede hanteerbaarheid belangrijk voor een eenvoudig gebruik van backup- en restore-procedures.
Voor datawarehouse-architecturen geldt dat we ook met beheer-ders te maken hebben. Vooral de meta-informatie en proces-informatie dienen het gebruik van standaard beheersoftware te ondersteunen. Het belangrijkste punt bij deze vereiste is dat de drielagen-referentiearchitectuur eenvoudiger te implementeren is in een hanteerbare technische architectuur. Het is immers door de verschillende karakters van de datawarehouse-lagen eenvoudiger op technisch niveau te ondersteunen. Zo kan bijvoorbeeld de informatielaag beschikbaar blijven, terwijl de datawarehouse laag geladen wordt. Zo kunnen we stabiliteit en hogere beschikbaar-heid van omgevingen krijgen.

Conclusies

Hedendaagse datawarehouses stellen business-informatie beschikbaar op een betrouwbare, snelle en flexibele manier. De beschreven referentiearchitectuur ontsluit management-informatie voor strategische doeleinden. Het speelt snel en flexibel in op wijzigingen in de informatiebehoefte door het flexibel definiëren en wijzigen van door eindgebruikers gewenste datamarts. Tevens maakt de isolatie van de drie beschreven architectuurlagen het mogelijk bronnen sneller toe te voegen en te wijzigen. Hierdoor scoort de architectuur niet alleen goed op het snel en flexibel beschikbaar stellen van informatie, maar ook goed op de vereiste onderhoudbaarheid, uitbreidbaarheid en schaalbaarheid. Deze worden verder nog ondersteund door de gebruikte ETL-methodiek en de gebruikte datamodellen in de staging laag en de datawarehouse laag. Gecombineerd met een goede hanteerbaarheid zorgt dit voor een relatief goedkope en betrouwbare datawarehouse-omgeving.

Marianne Kompagne is BI/DWH architect bij Kadenza.