Gefaseerde projectaanpak bij vertaling datamodel in een BI-project

Analyseren en modelleren met behulp van matrices

Rob Peters

Waaraan moet de modellering in een Business Intelligence (BI) project voldoen? Ten eerste moet de gewenste informatie worden opgeslagen. Vervolgens moet de opslag zodanig zijn dat de informatie snel in de gevraagde vorm getoond kan worden. De nadruk ligt dus op inhoud en vorm. In een BI-project worden deze inhoud en vorm bepaald en vertaald naar een datamodel. Het GUIDE-model, de voor BI-projecten, ondersteunt en stuurt deze vertaalslag.

De modellering van een BI-project omvat de datamodellen van de onderscheiden architectuurcomponenten. Het datawarehouse vormt de basis voor de BI-omgeving van de organisatie. Daarom wordt juist hierin de informatie zo gedetailleerd mogelijk vastgelegd, inclusief de gewenste historie. Tevens biedt het datawarehouse de mogelijkheid tot uniformiteit door het vastleggen van definities. Het datawarehouse wordt niet benaderd door eindgebruikers met hun analyse-tools. De datamarts daarentegen worden wel benaderd door deze eindgebruikerstools. De modellering moet daarom zijn afgestemd op het optimaal functioneren daarvan. Dit kan bijvoorbeeld betekenen dat data in een datamart moet worden geaggregeerd of dat meetwaarden worden berekend.
De datamarts worden gevuld vanuit het datawarehouse.

Hiërarchische relaties

Gewenste informatie wordt vastgelegd in definities en structuur van meetwaarden en dimensies. Meetwaarden moeten eenduidig voor de gehele organisatie worden gedefinieerd. Zo kan bijvoorbeeld het begrip winstmarge binnen verschillende afdelingen een andere betekenis hebben. Als verschillende definities van marge gewenst zijn, dan moeten deze worden benoemd en gedefinieerd. Bij voorkeur vindt deze naamgeving en definitie eenduidig in het datamodel van het datawarehouse plaats, zodat deze automatisch worden meegenomen in de afgeleide datamarts. De structuur van de informatie wordt in grote mate bepaald door de hiërarchie in de dimensies en de relatie tussen meetwaarden en dimensies. In dimensies kan men vaak een hiërarchische relatie tussen attributen definiëren. Bijvoorbeeld in een tijddimensie bestaat een hiërarchische relatie tussen jaar, maand en dag. Verder wordt per meetwaarde bepaald aan welke dimensies een koppeling mogelijk is, en wat het laagst mogelijk niveau is van die koppeling. Bijvoorbeeld, facturen worden per dag vastgelegd en daarom kan de meetwaarde factuurbedrag worden gekoppeld aan de tijddimensie met als laagste niveau de dag. Bij maandbudgetten daarentegen, kan het budgetbedrag worden gekoppeld aan de tijddimensie en is het laagste niveau de maand. Het is van belang deze relatie tussen meetwaarden en dimensies goed te bepalen zodat vergelijkingen van meetwaarden inzichtelijk en eenvoudig worden.

De gewenste vorm van de informatie heeft belangrijke consequenties voor de datamarts. Eerst moet worden bepaald welke meetwaarden en dimensies men gezamenlijk wil zien en op welk detailniveau. Als meetwaarden uit verschillende omgevingen moeten worden samengevoegd, dan moeten de gevolgen daarvan worden onderzocht. Het samenvoegen van de meetwaarden factuurbedrag en budgetbedrag in een vergelijkend rapport, betekent dat het laagste niveau van koppeling met de tijddimensie maand wordt. Het budgetbedrag is hier beperkend omdat het wat betreft detail niet verder dan maand gaat. Het detailniveau waarop men de informatie wil en kan zien, bepaalt het aggregatieniveau van de datamarts.

> Een groot deel van de gewenste informatie blijkt in de bronsystemen te zitten

Als de gewenste vorm van de informatie een OLAP-kubus is waarin in- en uitzoomen van belang is, dan worden de dimensies beperkt tot duidelijke hiërarchische structuren of in ieder geval classificerende attributen. Zo zal men in een OLAP-kubus van de productdimensie wel de hiërarchie “productgroep, productsubgroep, product” opnemen maar niet het nettogewicht (wellicht wel gewichtsklasse). Echter, in rapporten worden ook beschrijvende attributen opgenomen, zoals het nettogewicht van een product of het telefoonnummer van een klant. Een goede beschrijving van gewenste combinaties meetwaarden en dimensies met detailniveau en attributen vorm de basis van een goed datamart-ontwerp.

Matrix

GUIDE is een projectaanpak voor BI-projecten (voor een bespreking zie het artikel in DB/M 8, december 2002). In GUIDE is praktijkervaring met BI-projecten vastgelegd in standaard-stappen met bijbehorende templates. Een belangrijk onderdeel van deze templates zijn de matrices waarin de relatie tussen meetwaarden en dimensies wordt vastgelegd.

Een aantal stappen in GUIDE levert belangrijke input voor het modelleren van het datawarehouse en de bijbehorende datamarts: de informatie-analyse, de bron-analyse en de output-analyse (zie afbeelding 1). In de informatie-analyse wordt bepaald wat de gewenste informatie is en deze wordt gedefinieerd in termen van meetwaarden en dimensies. In de bron-analyse wordt onderzocht in welke bronsystemen deze meetwaarden en dimensies zijn vastgelegd. De vorm waarin deze meetwaarden en dimensies moeten worden getoond, wordt in de output-analyse gedefinieerd. In iedere stap gaat het om specifieke combinaties van meetwaarden en dimensies. Deze combinaties kunnen heel goed gevisualiseerd worden met behulp van een meetwaarden-dimensie matrix.

In afbeelding 2 wordt getoond hoe de meetwaarden-dimensies matrix wordt toegepast in de verschillende stappen van het GUIDE-model. In de informatie-analyse wordt bepaald dat de meetwaarden verkooporderbedrag en factuurbedrag van belang zijn. Deze meetwaarden moeten worden geanalyseerd met behulp van de dimensies tijd, product en klant. De gebruiker wil de meetwaarden verkooporderbedrag en factuurbedrag vergelijken op het gedetailleerde niveau dag, product en klant. In de volgende stap, de bron-analyse, wordt bepaald of op dat niveau vergelijking mogelijk is. Een groot deel van de gewenste informatie blijkt in de bronsystemen te zitten. Het factuurbedrag is echter niet per product vastgelegd en omdat facturen één maal per week op een vaste dag worden gegenereerd, is de factuur gekoppeld aan een weeknummer in plaats van datum. Dat is een hoger aggregatieniveau dan het gewenste. Bij het modelleren moet hiermee rekening worden gehouden. De gecombineerde informatie uit de informatie-analyse en de bron-analyse vormt dé basis voor het ontwerp van het datawarehouse. Verkooporderbedrag en factuurbedrag zullen zo gedetailleerd mogelijk in aparte feitentabellen worden vastgelegd. Het model van de feitentabellen wordt: verkooporder (datum, product, klant, verkooporderbedrag); factuur (week, jaar, klant, factuurbedrag).

I. Informatie-analyse: Gewenste meetwaarde-dimensie combinaties.

Dimensies		**Meetwaarden**	
Naam	**Laagste niveau van koppeling**	**Verkooporder bedrag**	**Factuurbedrag**
Tijd	Dag	gewenst	gewenst
Product	Product id	gewenst	gewenst
Klant	Klant id	gewenst	gewenst

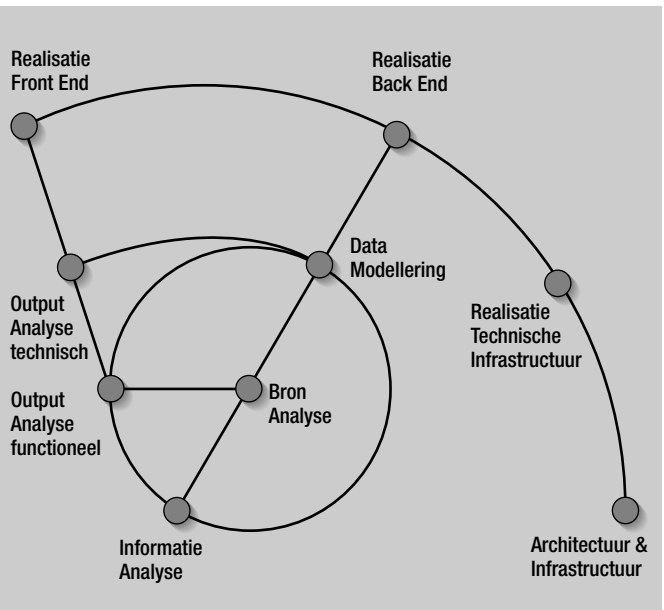
II. Bron-analyse: Gevonden meetwaarde-dimensie combinaties.

Dimensies		**Meetwaarden**	
Naam	**Laagste niveau van koppeling**	**Verkooporder bedrag**	**Factuurbedrag**
Tijd	Dag	gevonden	
	Week		gevonden
Product	Product id	gevonden	
Klant	Klant id	gevonden	gevonden

III. Output-analyse: In het rapport toegepaste meetwaarde-dimensie combinaties.

Dimensies		**Meetwaarden**	
Naam	**Laagste niveau van koppeling**	**Verkooporder bedrag**	**Factuurbedrag**
Tijd	Dag		
	Week	rapport	rapport
Product	Product id		
Klant	Klant id	rapport	rapport

Afbeelding 2: Voorbeeld van gebruik van de meetwaarden-dimensies matrix.



Afbeelding 1: De voor modellering relevante stappen in het GUIDE-model.

Output-analyse is de volgende stap. Nu moet worden bepaald welke combinaties van meetwaarden en dimensies gezamenlijk geanalyseerd gaan worden en dus gezamenlijk in een output vorm krijgen. In stap II, de bron-analyse, is getoond welke van de gewenste meetwaarde-dimensie combinaties mogelijk zijn. De keuzes van de output-analyse moeten binnen die mogelijkheden vallen. De meetwaarden-dimensies matrix geeft deze mogelijkheden duidelijk weer (zie afbeelding 2). Omdat de gebruiker verkooporderbedrag en factuurbedrag wil vergelijken op een zo laag mogelijk aggregatieniveau, is gekozen voor een rapport waarin dat mogelijk is. Verkooporderbedrag is geaggregeerd per week om het op hetzelfde niveau als factuurbedrag te krijgen.

Verder is product als invalshoek komen te vervallen omdat het niet aan factuur gekoppeld is. In de meetwaarden-dimensies matrix worden deze keuze en de consequenties voor het data-model inzichtelijk. De matrix laat zien dat in het datamart-model gekozen kan worden voor één feitentabel: verkoop-factuur (week, jaar, klant, verkooporderbedrag, factuurbedrag). Concluderend mag worden gesteld dat de meetwaarden-dimensies matrix de relatie tussen de gewenste informatie, de gevonden bronnen, het gewenste rapport en de datamodellen inzichtelijk toont.

Hoger aggregatieniveau

Het volgende voorbeeld betreft een logistiek bedrijf dat inzicht wil krijgen in de financiële positie. Omdat het bedrijf grote voorraden met daaraan verbonden kosten heeft, wordt er speciale aandacht aan besteed. Verder wordt geanalyseerd of het bedrijf op koers is met de behaalde omzet, door deze te vergelijken met het budget. De invalshoeken van waaruit deze meetwaarden moeten worden onderzocht zijn product, klant en tijd. Hoewel men op hoofdlijnen inzicht wil krijgen, wil men de mogelijkheid hebben de detail-gegevens achter de hoofdlijnen te zien. Daarom moet analyse tot op het niveau van omzet per dag, per product en per klant mogelijk zijn. De ontwikkeling van de datawarehouse- en data-mart-datamodellen wordt getoond via de meetwaarden-dimensies matrix.

In de stap informatie-analyse wordt bepaald dat de meetwaarden voorraadwaarde, omzetbedrag en budgetbedrag van belang zijn. Deze meetwaarden moeten worden geanalyseerd met behulp van de dimensies tijd, product en klant. Er zijn echter verschillen tussen de meetwaarden. Omzet en budget in relatie met de tijdsdimensie worden anders bepaald dan voorraad. Omzet en budget worden getotaliseerd per tijdsperiode (dag, maand, enzovoort), terwijl voorraad wordt bepaald op een vast moment, bijvoorbeeld de dagvoorraad dagelijks om 24.00 uur of de maandvoorraad iedere eerste dag van de maand. De relatie tussen voorraad en klant heeft in dit voorbeeld geen betekenis, dus die combinatie valt af.

> Afhankelijk van de te gebruiken tools worden modelleerkeuzes gemaakt

Omdat bepaling van het budget per dag geen zin heeft, wordt het budget alleen per maand gevolgd. Zo ontstaat een set van gewenste en realistische combinaties die in de matrix getoond worden (zie afbeelding 3: I). De volgende stap is de bron-analyse. Een groot deel van de gewenste informatie blijkt in de bron-systemen te zitten. Het budget is echter niet per klant vastgelegd, en in relatie tot product, slechts op productgroepniveau bepaald. Dat is een hoger aggregatieniveau dan het gewenste. Een ander probleem vormt de voorraad. In het bronsysteem wordt slechts de actuele voorraad bijgehouden. Om de voorraad per dag en per maand te kunnen analyseren, moet in het datawarehouse voorraadhistorie worden bewaard. Door de mogelijke combinaties te tonen in de matrix, ontstaat er een realistisch beeld (zie afbeelding 3: II). Deze gecombineerde informatie uit de informatie-analyse en de bron-analyse vormt de basis voor het ontwerp van het datawarehouse.

Voorraad, omzet en budget zullen zo gedetailleerd mogelijk in aparte feitentabellen worden vastgelegd. Voor voorraad moet nog een keuze worden gemaakt: wordt historie bewaard, en zo ja, op welk detailniveau? Als de keuze per maand is, dan wordt het model van de feitentabellen: voorraad (maand, jaar, product, voorraadwaarde); omzet (datum, product id, klant id, omzetbedrag); budget (maand, jaar, productgroep, budgetbedrag).
Met geïntegreerde informatie-analyse- en bron-analyseresultaten in de matrix, vormt deze nu ook de basis voor het bepalen van output ten behoeve van analyses.

I. Informatie-analyse: Gewenste meetwaarde-dimensie combinaties.

Dimensies		Meetwaarden		
Naam	**Laagste niveau van koppeling**	**Voorraad waarde**	**Omzet bedrag**	**Budget bedrag**
Tijd	Maand	gewenst	gewenst	gewenst
	Dag	gewenst	gewenst	
Product	Productgroep	gewenst	gewenst	gewenst
	Product id	gewenst	gewenst	gewenst
Klant	Klant id		gewenst	gewenst

II. Bron-analyse: Gevonden meetwaarde-dimensie combinaties.

Dimensies		Meetwaarden		
Naam	**Laagste niveau van koppeling**	**Voorraad waarde**	**Omzet bedrag**	**Budget bedrag**
Tijd	Maand	historie nodig	gevonden	gevonden
	Dag	historie nodig	gevonden	
Product	Productgroep	gevonden	gevonden	gevonden
	Product id	gevonden	gevonden	
Klant	Klant id		gevonden	

III. Output-analyse: In drie analyses toegepaste meetwaarde-dimensie combinaties.

Dimensies		Meetwaarden		
Naam	**Laagste niveau van koppeling**	**Voorraad waarde**	**Omzet bedrag**	**Budget bedrag**
Tijd	Maand	A, B	A, B	A
	Dag		C	
Product	Productgroep	A	A	A
	Product id	B	B, C	
Klant	Klant id		C	

Afbeelding 3: Voorraad, omzet en budget in de meetwaarden-dimensies matrix.

In de output-analyse worden gewenste analyses vertaalt naar een output. De mogelijke output-combinaties van meetwaarden en dimensies worden gevisualiseerd in de matrix (zie afbeelding 3, III). Het is belangrijk dat meetwaarden worden vergeleken op gelijke hiërarchische niveaus. Bijvoorbeeld, als budget met omzet vergeleken wordt dan moet dit gebeuren op maandniveau. Immers, omzet per dag kan niet vergeleken worden met budget per maand. In dit voorbeeld is gekozen voor drie mogelijke analyses; A, B, C.
In analyse A worden voorraad, omzet en budget naast elkaar vergeleken per maand en per productgroep. In analyse B worden voorraad en omzet met elkaar vergeleken op maand- en product id-basis. In analyse C wordt alleen omzet gevolgd, en wel per dag, per product en per klant.
In de matrix is te zien dat bij onderlinge vergelijking van meetwaarden vanuit een bepaalde dimensie, altijd de laagst mogelijke gemeenschappelijke deler in die dimensie is gekozen (afbeelding 3, III). De matrix laat zien welke meetwaarden en dimensies gezamenlijk geanalyseerd moeten worden en deze wensen vormen de basis voor het datamart-datamodel.
Afhankelijk van de te gebruiken tools en de gewenste performance worden modelleerkeuzes gemaakt. Om de performance tijdens de analyses te optimaliseren kan men ervoor kiezen voor iedere analyserichting (A, B, C) een feitentabel met het optimale aggregatieniveau te creëren in de datamart. Bij een normale groei van het datawarehouse met analysevragen, zou dit te veel tabellen opleveren. Het andere extreem is één feitentabel die de drie analyserichtingen mogelijk maakt. Deze tabel zou er als volgt uit kunnen zien: feitentabel (dag, maand, jaar, product id, productgroep, klant id, meetwaarde type, bedrag) (zie afbeelding 4).
De velden dag en klant id hebben de waarde '1' voor voorraad en budget, evenals product id voor budget. Het veld meetwaarde type heeft de waarden 'voorraad', 'omzet' of 'budget', afhankelijk van de inhoud van het veld aantal. Het uiteindelijke datamart-model bevindt zich tussen beide uitersten.

Zo zijn in dit voorbeeld informatiewens, realiteit en uiteindelijke invulling onderling verbonden door middel van meetwaarden-dimensies matrices. Problemen door onvolledige compatibiliteit van meetwaarden zijn snel duidelijk. Het proces van modellering laat zich gemakkelijk op hoofdlijnen sturen door wat de matrix aangeeft als zijnde mogelijke combinaties.

Conclusie

De datamodellen in een BI-project moeten aan de voorwaarden van inhoud en vorm voldoen. Deze inhoud en vorm worden bepaald door de vraag van de gebruiker. De projectaanpak GUIDE stuurt hier duidelijk op aan door het gebruik van meetwaarden-dimensies matrices in haar templates. Vooral in de analysestappen zijn deze van belang. De stappen informatie-analyse en bron-analyse bepalen wat gewenste informatie en wat mogelijk beschikbare informatie is.

Integratie van deze informatie in een matrix maakt (on)mogelijkheden inzichtelijk en deze matrix vormt een goede basis voor de datamodellering van het datawarehouse. In de stap output-analyse wordt de matrix verder ingevuld met gewenste analyses bijvoorbeeld in de vorm van rapportages. Dit vormt een overzichtelijke basis voor ontwerpkeuzes in de datamart.
Aan de hand van eenvoudige voorbeelden is het nut van meetwaarden-dimensies matrices voor het modelleren van datawarehouses en datamarts getoond. Door de overzichtelijkheid van de matrices zal bij toename van complexiteit het nut van de matrices alleen maar toenemen.

Rob Peters (rob.peters@ordina.nl) is senior consultant bij Ordina TTI en gespecialiseerd in Business Intelligence en datawarehousing.

Dag	Maand	Jaar	Product id	Productgroep	Klant id	Meetwaarde type	Bedrag
1	1	2003	1	AA	1	budget	2000
1	1	2003	37548	AA	1	voorraad	500
8	1	2003	37548	AA	X8294	omzet	50
23	1	2003	37548	AA	X9203	omzet	75
25	1	2003	37548	AA	X1029	omzet	60
1	1	2003	37279	AA	1	voorraad	200
17	1	2003	37279	AA	X7024	omzet	100
1	2	2003	1	AA	1	budget	2500
1	2	2003	37548	AA	1	voorraad	315
4	2	2003	37548	AA	X2034	omzet	100
12	2	2003	37548	AA	X8294	omzet	50
19	2	2003	37548	AA	X4020	omzet	125
enz							

Afbeelding 4: Voorbeeld van een geïntegreerde feitentabel.