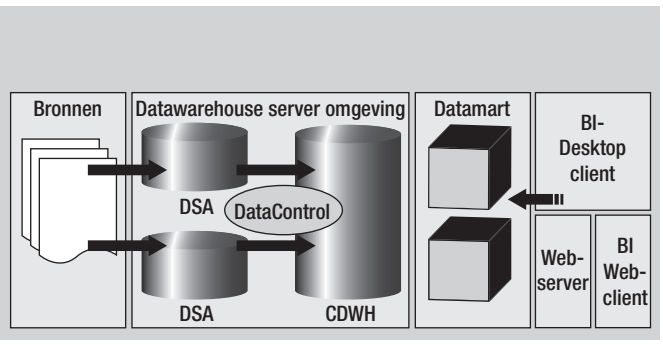
Fraudedetectie met behulp van BI

Data cleansing tegen 'money laundring'

Oscar Zonneveld

De aandacht voor detectie van financiële transacties die mogelijk in verband gebracht kunnen worden met (internationale) 'money laundring' is de laatste tijd sterk toegenomen. Om in omvangrijke dataverzamelingen verdachte patronen te kunnen vinden, zijn kostbare en complexe software-oplossingen beschikbaar. Er bestaat echter een methode om ook op basis van relatief goedkope en algemeen beschikbare datawarehouse-technieken het zoeken naar verdachte transactiepatronen te verbeteren.

De markt voor zogeheten money transfers, transacties waarbij contant geld kan worden overgemaakt naar een willekeurig persoon elders ter wereld, leent zich niet alleen bij uitstek voor het verplaatsen van (zwart) geld, maar is ook een methode om criminele organisaties of personen in hun activiteiten te ondersteunen. Voor money transfers is het hebben van een (traceerbare) bankrekening niet vereist en is bovendien de snelheid waarmee geld zich wereldwijd verplaatst bijzonder hoog. Afhankelijk van de bestemming is een via money transfer overgemaakt bedrag binnen enkele minuten aan de andere kant van de wereld te ontvangen.
Daarom is een meldingsplicht ingesteld waarbij instanties, die financiële diensten aanbieden die zouden kunnen worden misbruikt, aan de hand van objectieve criteria melding kunnen maken van ongebruikelijke transacties. Een dergelijk objectief criterium is bijvoorbeeld de hoogte van het over te maken bedrag. Alle transacties boven een ingesteld limiet worden automatisch gemeld.



Afbeelding 1: Het datawarehouse is opgebouwd uit een drietal lagen.

Patroonherkenning

Naast de objectieve criteria zijn er echter ook subjectieve criteria, op grond waarvan een deskundige op dit gebied zou kunnen besluiten een transactie te melden. Hierbij gaat het vaak om patroonherkenning, zoals bijvoorbeeld een aantal transacties binnen korte tijd op verscheidene locaties in het land uitgevoerd voor bedragen die net onder de objectieve meldingsplicht liggen. In combinatie met het wisselend gebruik van identificatiemiddelen (paspoort, rijbewijs, ID-Kaart) of het per transactie opgeven van een andere naam, is erg lastig en zeer tijdrovend om in traditionele dataverzamelingen een dergelijk patroon te kunnen vinden op het moment dat informatie hierover gewenst is.

> Een via money transfer overgemaakt bedrag is binnen enkele minuten aan de andere kant van de wereld te ontvangen

In de praktijk blijkt dat het in veel gevallen voldoende is om vanuit een bepaald gegeven te redeneren en niet, zoals door andere oplossingen wordt ondersteund, zonder concreet uitgangspunt alle mogelijke verbanden uit een dataverzameling te destilleren in de hoop daarbij iets op het spoor te komen. Aan de hand van het praktische voorbeeld, waarbij door middel van de beschreven wijzigingen in de naam, geboortedatum of het identificatiemiddel een poging wordt gedaan een opvallend transactiepatroon te verdoezelen, wordt uitgelegd hoe met relatief weinig inspanning een datawarehouse geschikt kan worden gemaakt voor het detecteren van verdachte transacties en patronen.
Samengevat komt de methode erop neer dat bij het opbouwen van de dimensies en meetwaarden, gegevens over vooraf gedefinieerde correlaties worden toegevoegd. Bij toepassing van een op deze wijze opgebouwd datawarehouse kan daardoor niet alleen informatie worden verkregen op grond van de aanwezige, traditionele dimensies, maar ook op basis van gecorreleerde data-elementen.

Methodiek

Het is lastig om te bepalen of een begunstigde daadwerkelijk de

beoogde begunstigde is, of dat we te maken hebben met een ander rechtelijk persoon. Om dit te kunnen vaststellen is van belang regels op te stellen die bepalen of een persoon daadwerkelijk mag worden gezien als één en dezelfde persoon (of in ieder geval met een grote waarschijnlijkheid). Er mag worden aangenomen dat twee transacties toebehoren aan dezelfde persoon indien geldt dat:

1. het ID (Paspoort, Rijbewijs, ID kaart, etc.) en geboortedatum gelijk zijn;
2. het ID (Paspoort, Rijbewijs, ID kaart, etc.) en achternaam gelijk zijn;
3. de achternaam en geboortedatum gelijk zijn.

Hier wordt een vierde criterium aan toegevoegd, namelijk een persoon die een transactie doet is minimaal gelijk aan zichzelf. Dit lijkt evident, echter in de oplossing is dit een belangrijk criterium. De bovenstaande criteria zijn (tijdens de informatie-analyse) opgesteld door de business.

Op basis van de bovenstaande criteria wordt ieder persoon gecorreleerd aan minimaal 1 ander persoon, namelijk op basis van het vierde criterium. Indien er extra requirements gelden worden ook deze correlaties geregistreerd. De OLAP Client maakt het vervolgens mogelijk om analyses uit te voeren op de verschillende correlaties. Zo is het bijvoorbeeld mogelijk om te analyseren hoe het transactiegedrag van een persoon er uit ziet. Naast het transactiegedrag van een persoon kan ook worden bekeken of de betreffende persoon in een andere gedaante (op basis van de onderkende correlaties) een transactie heeft uitgevoerd. Op basis van deze patronen worden vervolgens de reeds eerder beschreven meldingen gedaan.

Implementatie en architectuur

Voor de implementatie van de methode voor het onderkennen van de fraude is gekozen voor een datawarehouse. Het datawarehouse draagt er zorg voor dat de data vanuit de bronnen op een gecontroleerde en eenduidige manier beschikbaar worden gesteld aan de gebruiker. Bovendien zorgt de datawarehouse-oplossing voor het vastleggen van historische data, zodat trendanalyse en patroonherkenning vanuit het verleden kunnen plaatsvinden.
Het datawarehouse is opgebouwd uit een drietal lagen, te weten DSA (Data Staging Area), het Central Datawarehouse (CDWH) en datamarts (DM), zie afbeelding 1 voor de architectuur.

Het transformatieproces wordt in detail beschreven. De data worden vanuit de bron naar de DSA-omgeving getransformeerd. In de DSA-omgeving kunnen zaken als cleansing worden doorgevoerd. Onder cleansing van data wordt verstaan dat data vanuit de bron op gecontroleerde wijze geschoond worden. Indien bijvoorbeeld een geboortedatum van 31 februari 2005 wordt ingegeven kan deze worden gecorrigeerd in een voor de business logische waarde. Dit kan dus betekenen dat de waarde wordt gecorrigeerd met de vermoedelijk beoogde waarde (bijvoorbeeld 28 februari 2005 of 1 Maart 2005) of dat het betreffende attribuut wordt gezet op een default waarde (vaak wordt hiervoor een waarde gebruikt die overduidelijk niet reëel kan zijn, zoals in dit voorbeeld 1 januari 1800). De geschoonde rijen worden in een aparte omgeving (tabel) geregistreerd met hun oorspronkelijk ingevoerde waarde en de geschoonde en/of default waarde.
Op deze manier is het gewaarborgd dat de 'originele waarde' behouden blijft.
In de CDWH-omgeving worden meestal twee verschillende modellen gebruikt. Er kan worden gekozen voor een CDWH op basis van een relationeel model (de methodiek die Bill Inmon omarmt) of een gedimensioneerd model (de methodiek die Ralph Kimball onderschrijft). In dit artikel is gekozen voor een gedimensioneerde oplossing volgens de methodieken van Kimball, omdat in deze casus gebruik wordt gemaakt van Microsoft Analysis Services.

> Een techniek als fuzzy logic zou niet misstaan om te bepalen welke procedures dienen te worden toegevoegd

Het transformatieproces tussen de DSA-omgeving is opgesplitst in twee delen. Als eerste worden de data voor de onderkende dimensies getransformeerd. Eén van deze dimensietabellen is de personentabel. In deze personentabel zijn alle personen die ooit een transactie hebben gedaan aanwezig. Ieder persoon wordt in deze tabel uniek opgeslagen. Dus wanneer er bij de invoer al dan niet bewust incorrecte gegevens worden ingevoerd, zoals bijvoorbeeld de geboortedatum, en deze persoon heeft reeds een transactie gedaan, wordt de betreffende persoon alsnog opnieuw opgevoerd. Door het toepassen van de reeds beschreven correlatiemethodiek wordt de betreffende 'vervuilde rij' gecorreleerd aan een rij die op basis van de gestelde criteria 'gelijk' is.

Dim Persoon

Doel database	Doel attribuut	Type	Sleutel
DWH..dim_Persoon	Per_SID		PK
DWH..dim_Persoon	Per_Achternaam	SCDT02	FK
DWH..dim_Persoon	Per_Initialen	SCDT02	FK
DWH..dim_Persoon	Per_Geboorte_datum	SCDT02	FK
DWH..dim_Persoon	Per_Type_ID	SCDT02	FK
DWH..dim_Persoon	Per_ID	SCDT02	FK

* Per_SID is de surrogate key die gebruikt wordt in de dimensie.
* SCDT02 is het type Slowly Changing Dimensie dat wordt gebruikt.

Aangezien de correlatie plaatsvindt tussen voorkomens van personen uit dezelfde tabel, wordt er eerst een view gecreëerd op de reeds eerder bestaande dimensietabel.

```
Create View dim_Correlatie_persoon
AS select * from dwh_db..dim_Persoon
```

Deze view en de originele personentabel worden vervolgens met elkaar vergeleken op basis van de eerder gestelde criteria.
De resultaten met daarbij het geldende criteriumnummer geeft aan welke correlatie een record van een persoon uit de personen-tabel heeft met een record van een persoon uit de dim_Correlatie_persoon View.
Voor het uitvoeren van de correlaties wordt voor iedere correlatie afzonderlijk een procedure ontwikkeld. Deze procedure (mits uitgevoerd) draagt er zorg voor dat de betreffende correlatie wordt geregistreerd in de eerder beschreven correlatietabel.

Controle	StoredProcedure	Uitvoeren
CorrelatieType0	Stp_correlatie_type_0	J
CorrelatieType1	Stp_correlatie_type_1	J
CorrelatieType2	Stp_correlatie_type_2	J
CorrelatieType3	Stp_correlatie_type_3	J

Door het vastleggen van de benodigde procedures in een tabel is het mogelijk om deze naar wens uit te breiden en de correlaties zo strikt mogelijk te leggen. Technieken als fuzzy logic zouden hier dan ook niet misstaan om te bepalen welke procedures dienen te worden toegevoegd om de nauwkeurigheid van de correlatie-methodiek te verhogen. De resultaten van de bovenstaande procedures worden weggeschreven in de onderstaande tabel.

Dim Correlatie			
Doel database	**Doel attribuut**	**Type**	**Sleutel**
DWH..dim_Correlatie	Cor_klant1		PK
DWH..dim_Correlatie	Cor_klant2	SCDT01	FK
DWH..dim_Correlatie	Cor_CorrelatieType	SCDT01	FK

De slowly changing dimension voor de 'relatie dimensie' is een type 1. Dit wil zeggen dat er alleen wijzigingen op bestaande rijen worden uitgevoerd en nieuwe rijen worden opgevoerd. In de besproken correlatie-casus worden er echter nooit rijen gewijzigd, immers een correlatie tussen 'twee' personen (al dan niet dezelfde) resulteert altijd in een nieuwe rij. Bij het vullen van deze dimensie is het dan ook legitiem om de betreffende dimensie eerst te trunceren alvorens hem opnieuw te vullen.
Met behulp van dynamisch SQL kan de tabel met de correlatie-procedures worden uitgelezen en worden de verschillende procedures individueel uitgevoerd.

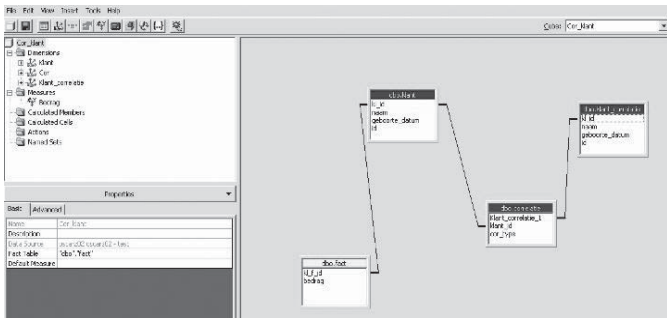
```
DECLARE dataControles CURSOR
FOR
SELECT  dct_ID
,       dct_StoredProcedure
,       dct_Uitvoeren
FROM    DataControle
WHERE   dct_DTSPackageNaam = @DTSPackageNaam
ORDER BY dct_Volgorde

OPEN dataControles
FETCH NEXT FROM dataControles INTO @id,
                                        @sp,@uitvoeren
IF @@FETCH_STATUS <> 0
PRINT '<<Tabel niet gedefinieerd.>>'

WHILE @@FETCH_STATUS = 0
BEGIN
IF UPPER(@uitvoeren) ='J'
BEGIN
BEGIN TRAN
EXECUTE(@sp)
COMMIT TRAN
END

FETCH NEXT FROM dataControles INTO @id,
                                        @sp,@uitvoeren
END

CLOSE dataControles
DEALLOCATE dataControles
```

In de volgende tabellen zijn de procedures voor de verschillende correlaties beschreven.

Stp_correlatie_type_0		
Type:	Stored Procedure	
Doel:	Voer de gegevens t.b.v. Correlatie-type 0 op.	
Beschrijving:	Deze procedure draagt er zorg voor dat alle gegevens t.b.v. Correlatie-type 0 worden opgeslagen in de correlatiedimensie.	
Input:	Parameter:	Beschrijving:
	@DataControlID	Primary key van een datacontrole uit de datacontrole-tabel.
Syntax	Insert into dwh..dim_Correlatie SELECT kl.per_id, kl_cor.Per_id, 0 FROM dwh..dim_persoon kl, dwh..dim_Correlatie_klant kl_cor WHERE kl.Per_id_id = kl_cor.Per_id	
Relatie stp:		
Template:		

Stp_correlatie_type_1		
Type:	Stored Procedure	
Doel:	Voer de gegevens t.b.v. Correlatie-type 1 op.	
Beschrijving:	Deze procedure draagt er zorg voor dat alle gegevens t.b.v. Correlatie-type 1 worden opgeslagen in de correlatiedimensie.	
Input:	Parameter:	Beschrijving:
	@DataControlID	Primary key van een datacontrole uit de datacontrole-tabel.
Syntax	Insert into dwh..dim_Correlatie SELECT kl.Per_id, kl_cor.Per_id, 1 FROM dwh..dim_Persoon kl, dwh..dim_Correlatie_persoon kl_cor WHERE kl.Per_ID= kl_cor. Per_ID AND kl.Per_Type_ID = kl_cor. Per_ Type_ID AND kl.Per_GeboorteDatum = kl_cor.Per_GeboorteDatum AND kl.Per_ID <> kl_cor.Per_id	
Relatie stp:		
Template:		

Stp_correlatie_type_2		
Type:	Stored Procedure	
Doel:	Voer de gegevens t.b.v. Correlatie-type 2 op.	
Beschrijving:	Deze procedure draagt er zorg voor dat alle gegevens t.b.v. Correlatie-type 2 worden opgeslagen in de correlatiedimensie.	
Input:	Parameter:	Beschrijving:
	@DataControlID	Primary key van een datacontrole uit de datacontrole-tabel.
Syntax	Insert into dwh..dim_Correlatie SELECT kl.Per_id, kl_cor.Per_id, 2 FROM dwh..dim_Persoon kl, dwh..dim_Correlatie_persoon kl_cor WHERE kl.Per_ID= kl_cor. Per_ID AND kl.Per_Type_ID = kl_cor. Per_ Type_ID AND kl.Per_Achternaam = kl_cor.Per_Achternaam AND kl.Per_ID <> kl_cor.Per_id	
Relatie stp:		
Template:		



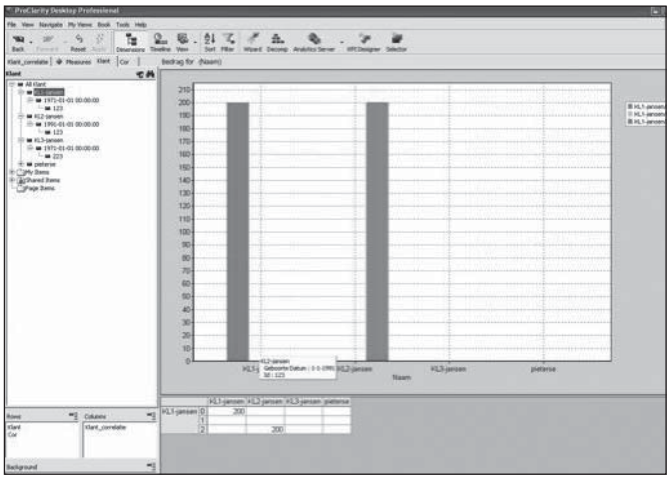
Afbeelding 2: De personendimensie wordt *gejoined* met de correlatiedimensie.

Stp_correlatie_type_3		
Type:	Stored Procedure	
Doel:	Voer de gegevens t.b.v. Correlatie-type 3 op.	
Beschrijving:	Deze procedure draagt er zorg voor dat alle gegevens t.b.v. Correlatie-type 3 worden opgeslagen in de correlatiedimensie.	
Input:	Parameter:	Beschrijving:
	@DataControlID	Primary key van een datacontrole uit de datacontrole-tabel.
Syntax	Insert into dwh..dim_Correlatie SELECT kl.Per_id, kl_cor.Per_id, 3 FROM dwh..dim_Persoon kl, dwh..dim_Correlatie_persoon kl_cor WHERE kl.Per_GeboorteDatum = kl_cor.Per_GeboorteDatum AND kl.Per_Achternaam = kl_cor.Per_Achternaam AND kl.Per_ID <> kl_cor.Per_id	
Relatie stp:		
Template:		

Zoals aangegeven zijn de uit te voeren correlatieprocedures uitbreidbaar en flexibel (door de procedures te registreren in een aparte tabel). Nadat bovenstaande procedures zijn uitgevoerd, is de correlatiedimensie gevuld. De volgende stap in het transformatieproces is het transformeren van de feiten. De feitentabel bevat geen directe relatie naar de correlatiedimensie. De reden hiervoor ligt in het volgende: een transactie waarvan de meetwaarden worden opgenomen als feit, wordt geregistreerd in de feitentabel, dat wil zeggen de meetwaarden worden geregistreerd én de daarbij geldende surrogate keys van de dimensies.

Fact Correlatie			
Doel database	**Doel attribuut**	**Type**	**Sleutel**
Fact Correlatie	Dim_Tijd.TJD_ID		
Fact Correlatie	Dim_Product.PRD_ID		
Fact Correlatie	Dim_Persoon.PER_DI		
Fact Correlatie	Transactie Bedrag		
Fact Correlatie	Aantal Artikelen		

Wanneer nu ook de correlatie als surrogate key wordt opgenomen in de feitentabel, zou dit kunnen betekenen dat de betreffende transactie meerdere malen wordt opgeslagen (namelijk bij geldende correlaties). Om deze dubbele tellingen te voorkomen wordt niet de correlatiedimensie gekoppeld aan de feitentabel, maar de 'oorspronkelijke' personendimensie. De personendimensie wordt *gejoined* met de correlatiedimensie, die wordt gekoppeld aan de correlatie-view. Hierbij wordt de 'snowflake-methodiek' toegepast. Door het toepassen van deze methodiek kan de gewenste informatie worden samengesteld. Afbeelding 2 toont het zojuist beschreven model.



Afbeelding 3: Voorbeeldrapport.

Afbeelding 3 toont een voorbeeldrapport: aangegeven in de rijen is welke klant er is en de daarbij bijbehorende correlatietype.
Op de kolommen is aangegeven met welke klant de betreffende klant van de rij een correlatie heeft. Het gepresenteerde getal is het bijbehorende bedrag. In dit voorbeeld heeft klant 1 (met geboorte datum 1-1-1971 en ID 123) een type 0 correlatie met zichzelf en een type 2 correlatie met klant 2. De achternaam van klant 2 is evenals klant 1 Jansen en zijn ID is ook 123. De geboortedatum wijkt in dit voorbeeld af (bijvoorbeeld door foutieve invoer).

Conclusie

Het blijkt dat correlaties een duidelijke meerwaarde kunnen bieden ten opzichte van een dataware-house-omgeving waarin deze methode niet wordt toegepast. De voorgestelde methode is bijzonder flexibel, omdat de correlaties eenvoudig kunnen worden uitgebreid of aangepast aan nieuwe wensen. Omdat uitsluitend gebruik gemaakt wordt van bestaande technieken, is iedere database-ontwerper ook zonder specifieke training in staat om, volgens het beschreven stramien, functionaliteit op basis van correlaties toe te voegen. Vanzelfsprekend is het uitgangspunt wel dat een goed ontworpen datawarehouse-omgeving beschikbaar is. Eenmaal geïmplementeerd en gedocumenteerd, is functionaliteit op basis van correlaties goed te beheren en te onderhouden, waardoor de kosten beperkt kunnen blijven. In de markt zijn verscheidene leveranciers actief met andere oplossingen op dit gebied. Deze oplossingen bieden in vele gevallen uitstekende resultaten, echter de kostprijs is navenant. Gezien de grote hoeveelheid aan functionaliteit die wordt geboden, mag worden verondersteld dat er ook geïnvesteerd moet worden in opleidingen om effectief met de geboden functionaliteit te kunnen omgaan en onderhouden. Voor projecten, waarbij een dergelijke oplossing 'overdone' is, kan de correlatietechniek', door zijn flexibiliteit en kostenefficiëntie, een welkome oplossing zijn.

Ing. O. Zonneveld (oscarz@infosupport.com) is consultant bij Info Support.