Het gebruik van de Data Vault voor Master Data Management

Natural world modeling

Karien Verhagen

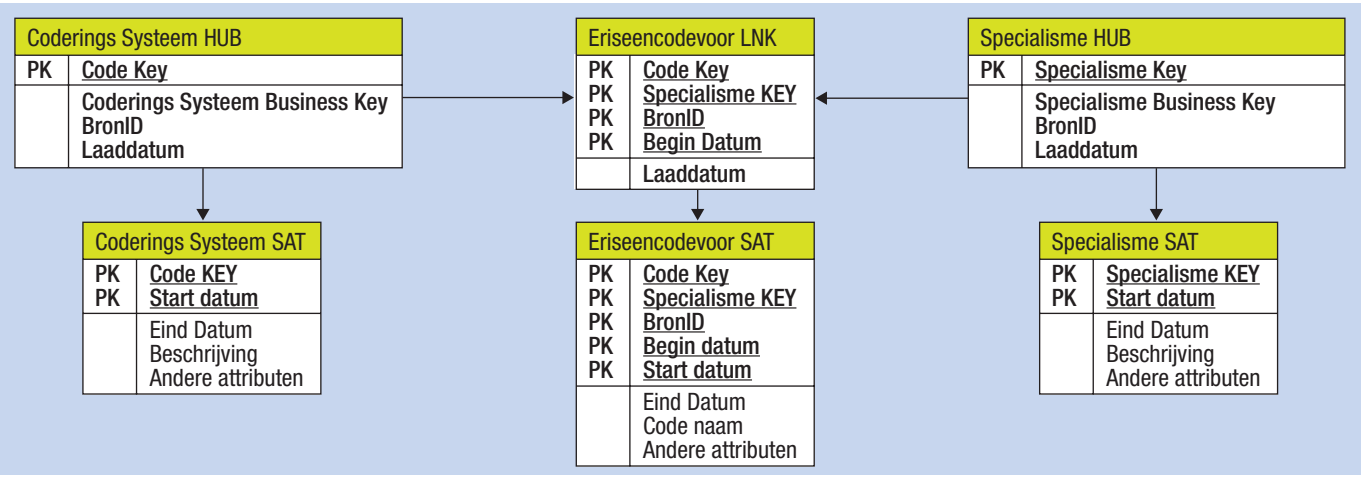
Vraagt u zich wel eens af waarom de Data Vault methodiek van Dan Linstedt zich nu ineens mag verheugen in een snel stijgende populariteit? De methode bestaat in zijn huidige vorm immers al zo'n acht jaar. Misschien is het de 'fit', het goede gevoel dat BI professionals hebben dat deze methode een oplossing biedt voor enkele flinke hobbels die BI-initiatieven in de praktijk moeten nemen, problemen die nu actueel zijn.

Het is niet alleen de theorie uit de boekjes van Linstedt maar vooral ook de filosofie van waaruit hij zijn theorie heeft opgebouwd, die aanspreekt. Dat overkwam ons projectteam bij Prismant. Prismant is een bedrijf dat onder andere door het verzamelen en ordenen van gegevens uit de gezondheidszorg rapporteert en inzage geeft in prestaties van individuele ziekenhuizen. Die prestaties worden gemeten aan die van vergelijkbare instellingen. Deze 'spiegelziekenhuizen' zijn de referentiekaders en worden door de abonnementhouder zelf opgegeven.
De opdracht van de projectgroep was om het bestaande gelaagde datawarehouse te herbouwen zodat het schaalbaar, flexibel en vooral ook tegen minder kosten onderhoudbaar zou worden. Voor twee van de vele problemen die we daarbij tegenkwamen vonden we een aanknopingspunt in de Data Vault theorie. Beide zijn min of meer Master Data Management (MDM) problemen.

Het zijn modelleringscomplicaties als gevolg van het feit dat er niet één basisregistratie is voor ziekenhuizen, noch voor specialismen of specialisten. Omdat onze oplossing de Data Vault theorie interpreteert en vanuit de filosofie breder trekt, hebben we de oplossingen ook met de Genesee Academy van Dan Linstedt besproken. Omdat hij enthousiast is en omdat de problematiek eenvoudig te projecteren is op andere soortgelijke problemen willen we de oplossingen graag met u delen.

Probleem 1

Eerste probleem is: er zijn allerlei naamgevingsconventies en even zoveel conversietabellen. Er zijn zeer veel coderingen voor allerlei soorten entiteiten in de gezondheidszorg: voor specialismen, voor ziekenhuizen, voor diagnoses enzovoort. Dan Linstedt leert ons dat daar een valide reden voor is. De academische ziekenhuizen hebben behoefte aan een andere indeling van hun 'business' dan kleine regionale ziekenhuizen. Het doceren bijvoorbeeld, het uitdragen van de kennis en het opleiden van specialisten is een wezenlijk onderdeel van hun taak. Het aanbod in medische zorg is ook anders. Daartussenin bevinden zich de topklinische ziekenhuizen, de hele grote, zoals het Antonius ziekenhuis of het Jeroen Bosch. Grote ziekenhuizen hebben andere en veel meer specialisten in dienst en willen ze ook op een andere manier clusteren. De rapportages van academische ziekenhuizen zullen ook een andere focus hebben. VWS, andere regeringsorganen en instituten zoals het CBS hebben echter behoefte aan



Afbeelding 1: Het ontkoppelen van de business key en naamgevingconventie.

Specialisme	ABC code	Prismant Code
Polyklinische Specialismen	12	*nvt*
Snijdende Specialismen	10	B
Beschouwende Specialismen	3	A-7
Radiologie	154-3	C3
Radiodiagnosticus	*Nvt*	C4
Nucleaire geneeskunde	*nvt.*	C18
Klinische cytologie	*nvt.*	C21
Pathologie	18	D1
Medische Microbiologie	21	F5
Immunologie	23	F8
Parasitologie	n.v.t.	F38

Afbeelding 2: Een conversietabel van ABC codering (fictief) naar Prismant codering.

één groot overzicht over de gezondheidszorg waarbij de specialist of het specialisme belangrijke assen of dimensies zijn. Zij hebben dus baat bij een universele registratie. De regering stimuleert een systeem van basisregistraties met duidelijke verantwoordelijkheden in de door haar gepropageerde NORA architectuur (www.e-overheid.nl/atlas/referentiearchtiectuur). Een dergelijke unieke en alles omvattende registratie voor specialismen, specialisten of ziekenhuizen is er nog niet.

Probleem 2

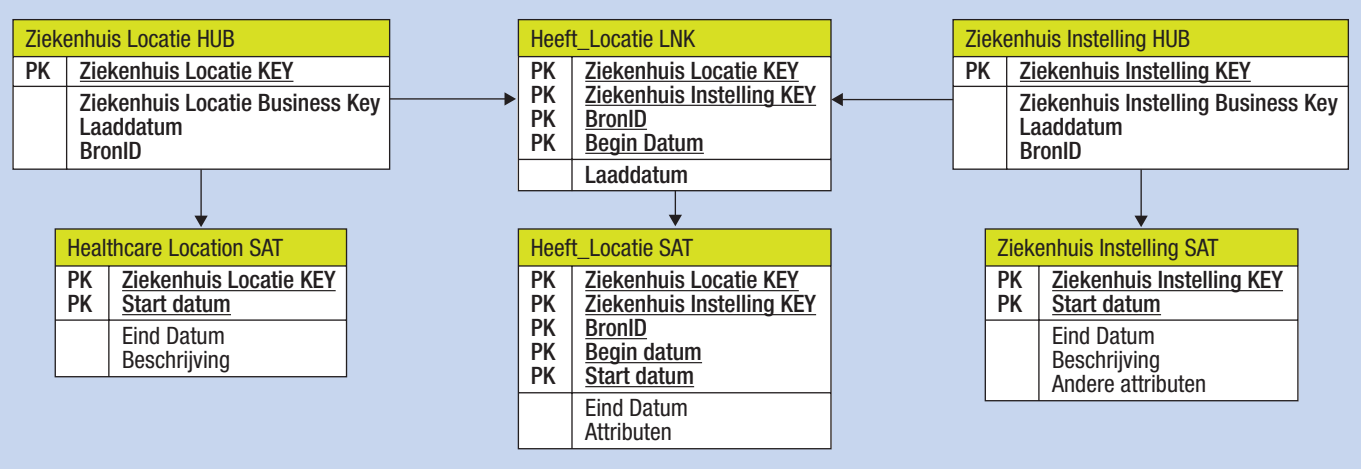
Tweede probleem is: fusies en samenwerkingsverbanden verstoren het zicht op historische ontwikkelingen.
Gezondheidszorginstellingen, maar ook zorgverzekeraars hebben de neiging om voortdurend te fuseren of kleinere partijen over te nemen om de toenemende druk van de administratieve lasten het hoofd te kunnen bieden, om efficiënter te kunnen werken. Rapporten, bedoeld om een correcte inzage te geven in ontwikkelingen op een tijdslijn, zijn daardoor gecompliceerd. In drie jaar tijd kan de organisatie van de instelling diverse malen wijzigen. De instelling kan nieuwe locaties krijgen door nieuwbouw of overnames. Steeds weer nieuwe manieren van samenwerking zorgen voor caesuren in het overzicht, het plotseling wegvallen of opduiken van een nieuwe eenheid in een historische lijn. Het ontsluiten is lastig en de database van het enterprise datawarehouse zal bij elke nieuwe vorm van samenwerking die door de gezondheidsmarkt wordt uitgevonden weer gewijzigd moeten worden. Dat is verre van ideaal. Het datamodel van een enterprise datawarehouse wijzigen tijdens de rit leidt vaak tot suboptimalisatie. De soms duizenden gebruikers moeten immers gewoon door kunnen werken.

Is Data Vault de oplossing?

Hoe gingen we deze problemen bij Prismant te lijf en op welke wijze heeft de filosofie van Dan Linstedt's Data Vault daarbij geholpen?
Stap 1: Vind de echte Business Key (Linstedt) en scheidt deze van de coderingssystemen en de codes voor die entiteit (projectteam Prismant).
Er zijn allerlei soorten specialismen en vele manieren om ze te coderen. Die codes bevatten overlappingen, ze hebben allemaal een ander doel maar geen van de coderingen dekt het hele gebied van bestaande specialismen. Kindchirurgie of vaatchirurgie kunnen een aparte code hebben, andere systemen hebben alleen een code voor chirurgie in het algemeen. Weer andere systemen hebben verschillende coderingen voor specialisten/docenten of specialisten in opleiding en 'normale' specialisten. Soms maakt men onderscheid tussen beschouwende specialismen (zoals pathologie of radiologie) en contactspecialismen, uitgeoefend door specialisten die rechtstreeks te consulteren zijn (zoals neurologen of KNO-artsen).
De Data Vault leert ons om *as-is* te registeren. Dan Linstedt noemt dat *natural world modeling*. We moeten daarvoor eerst de echte business key vinden, de naam onder welke iedereen in de gezondheidszorg het begrip herkent en identificeert. In dit geval is dat de volledige naam van het specialisme. We hebben deze namen gescheiden van de coderingssystemen en hebben links

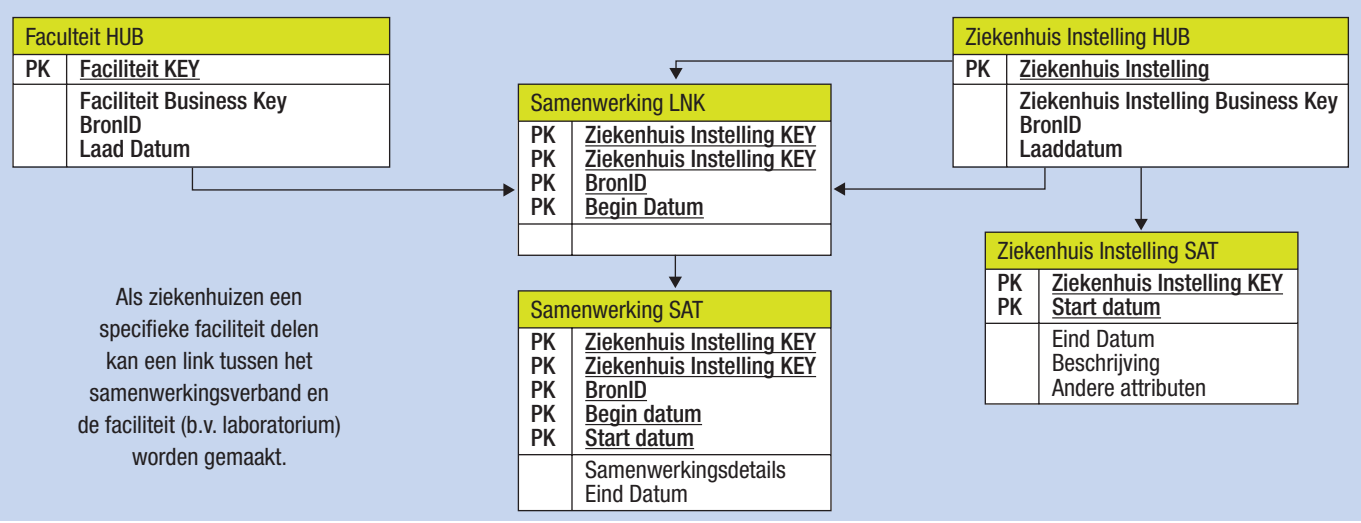
Code Key	Specialisme Key	BronID	Begin Datum	Start Datum	Eind Datum	Codenaam
22	124	Prismant	21-6-2009	22-6-2009	99-99-9999	12
22	125	Prismant	21-6-2009	22-6-2009	99-99-9999	10
23	125	Prismant	21-6-2009	22-6-2009	99-99-9999	B
22	126	Prismant	21-6-2009	22-6-2009	99-99-9999	3
23	126	Prismant	21-6-2009	22-6-2009	99-99-9999	A-7
22	127	Prismant	21-6-2009	22-6-2009	99-99-9999	154-3
23	127	Prismant	21-6-2009	22-6-2009	99-99-9999	C3
23	128	Prismant	21-6-2009	22-6-2009	99-99-9999	C4
23	129	Prismant	21-6-2009	22-6-2009	99-99-9999	C18
23	130	Prismant	21-6-2009	22-6-2009	99-99-9999	C21
22	131	Prismant	21-6-2009	22-6-2009	99-99-9999	18
23	131	Prismant	21-6-2009	22-6-2009	99-99-9999	D1
22	132	Prismant	21-6-2009	22-6-2009	99-99-9999	21
23	132	Prismant	21-6-2009	22-6-2009	99-99-9999	F5
22	133	Prismant	21-6-2009	22-6-2009	99-99-9999	23
23	133	Prismant	21-6-2009	22-6-2009	99-99-9999	F8
23	134	Prismant	21-6-2009	22-6-2009	99-99-9999	F38

Afbeelding 3: De rijen in de Eriseencodevoor SAT-tabel.



Afbeelding 4: Een instelling kan meer locaties hebben.

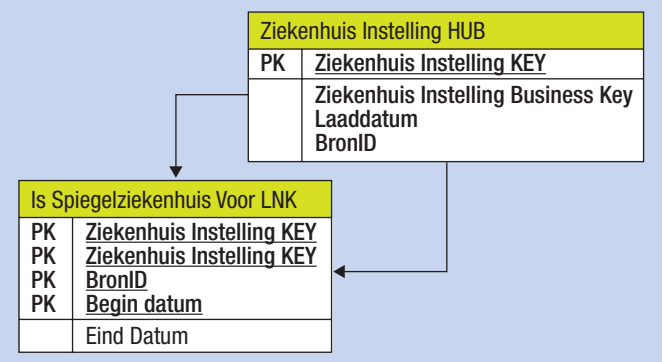
aangebracht daar waar een relatie van toepassing is. Als er een naam voor dit specialisme is in het coderingssysteem voegen we een link toe, als er geen is doen we dat niet, zie afbeelding 1.
Bij veel bedrijven worden voor het vertalen van de ene naar de ander coderegistratie Excel-sheets met overeenkomsten gebruikt, de conversietabellen (zie de tabel in afbeelding 2). De vet gezette specialismen zijn groeperingen van andere specialismen. Dit soort Excel-tabellen is niet erg transparant. Als er dan ook nog wijzigingen in de tijd zijn, bijvoorbeeld een andere groepering of codering vanaf 2008, dan heeft u met uw Excel-conversietabellen een onderhoudsprobleem voor de toekomst geschapen.
Onderhoud op die vele ontoegankelijke tabellen zal kostbaar zijn en steeds lastiger en minder betrouwbaar worden. Het blijft op die manier foutgevoelig mensenwerk.
De modellering van deze Excel-tabel (afbeelding 2) zou volgens de voorgestelde wijze van modelleren leiden tot de volgende rijen (zie de tabel in afbeelding 3). De gegenereerde Code Key voor de ABC codering is 22, die voor de Prismant code 23. Specialisme KEY 125 geeft bijvoorbeeld snijdende specialismen aan. Voor de *nvt*-velden vindt u geen overeenkomstige rij in de Eriseencodevoor SAT tabel. De Eriseencodevoor LNK tabel beschrijft het feit *dat* er een relatie *is* tussen de naamgevingsconventie en het specialisme met ingang van een begindatum. In de SAT-tabel staat de code zelf met de tijdsspanne van geldigheid (van *startdatum*[1] tot einddatum).
Op deze wijze kunt u converteren van elk willekeurig bestand en elk willekeurig nog uit te vinden systeem naar een ander coderingssysteem en op een historisch correcte wijze.
Wijzigingen worden met datums bijgehouden in de SAT-tabel. Er gaat geen informatie verloren. Als er een equivalent is laat u dat zien. Deze manier van modelleren heeft potentie als katalysator van Data Governance.[2] Want u kunt nu rapporten publiceren over de doublures en de gaten in de registraties. De business kan die informatie verwerken door redundantie te knippen en gaten te vullen en zich een beeld te vormen van het systeemoverschrijdende bedrijfsbrede begrip. Op die manier werkt de business langzaam maar bewust aan een overall registratiesysteem, dat correct en betrouwbaar is. Dat centrale systeem wordt impliciet begrepen en gedragen, de business heeft het immers zelf gecreëerd. De randvoorwaarden voor een succesvolle Data



Afbeelding 5: Samenwerkingsverbanden.

Governance is daarmee ingevuld. IT faciliteert het proces maar is er geen eigenaar van. De rapporten met de terugkoppeling over de verbeteringen zullen de cyclische procesgroei naar Data Governance en Master Data Management stimuleren.

Stap 2: Vul een ongelijke hiërarchie ook in door LINK-tabellen.
De oplettende lezer ziet dat er in de tabel met specialismen geen hiërarchie is verwerkt. Alle groeperende specialismen en de individuele specialismen op het laagste niveau staan in één tabel. Ze zijn nog niet gegroepeerd. De gangbare manier om specialismen te groeperen is om er 1-op-n relaties van te maken. Probleem is hier dat die groepering soms ook op het laagste niveau gebruikt wordt. Een (soms) groeperend specialisme moet dan verwijzen naar de groepsitems door een recursieve verwijzing naar de specialismentabel. Dat is geen handige manier van modelleren in een datawarehouse, omdat die groeperingen ook nog eens in de historie kunnen wijzigen.
Wanneer drie ziekenhuizen bijvoorbeeld fuseren tot een instelling, zou het ziekenhuis een verwijzing naar zichzelf krijgen, waarbij het nieuwe 'paraplu' ziekenhuis een startdatum krijgt, gelijk aan de einddatum van de samenstellende delen. Dat moet beter kunnen.
De theorie van Dan Linstedt stelt ons in de gelegenheid alle soorten van samenwerking en overnames te modelleren als een expliciete relatie, met behulp van LINK-tabellen. Het model is ook schaalbaar en flexibel, doordat elke manier van samenwerken die nog wordt uitgevonden zal resulteren in een nieuw soort LINK-tabel die eenvoudig kan worden toegevoegd. Relaties tussen ziekenhuizen kunnen ook attributen hebben, die we in SAT-tabellen hebben ondergebracht.
Op zoek naar de business key hebben we voor de ziekenhuizen twee business entiteiten gedefinieerd:
1. De (geografische) locatie, het fysieke gebouw met de afdelingen, een ingang en parkeerplaatsen;
2. De (ziekenhuis)instelling, de meer abstracte administratieve en financiële eenheid (zie afbeelding 4).
Aan die twee eenheden wordt ook vaak met een verschillende naam (lees: business key) gerefereerd. Zo heeft de ISALA Kliniek (Instelling) bijvoorbeeld twee locaties, namelijk Weezenlanden en het Sophia ziekenhuis.
Met de LINK-modellering tussen locaties en instituten kan elke willekeurig relatie worden gemodelleerd, een fusie, een overname, nieuwbouw of zoals in het voorbeeld het delen van een faciliteit, bijvoorbeeld een laboratorium, zie afbeelding 5. U kunt voor elke wijze van samenwerken een nieuw soort LINK maken, een fusie heeft in potentie andere gevolgen en attributen dan een overname.
Ook de zogenaamde spiegelziekenhuizen, de ziekenhuizen waarmee een ziekenhuis vergeleken wenst te worden hebben we geïmplementeerd als LINK (afbeelding 6). Op deze manier ontstaat een groot datamodel. Toch is het zeker niet gecompliceerder dan het met de hand reconstrueren van die relaties uit vele steeds wijzigende spreadsheets. Dit grote datamodel is uitstekend onderhoudbaar vanwege de uniforme wijze van modelleren en de gestructureerde manier van het opslaan van historie. Alle verbanden zijn expliciet gemaakt en benoemd als relaties in LINK-tabellen. Er zijn geen wijzigingen maar uitsluitend toevoegingen nodig bij wijzigende verbanden. Een Data Vault model is een risicomijdende database. Een ander groot voordeel is dat alle wijzen waarop de business naar de gegevens zou willen kijken (door een rode, bruine of groene bril) in principe onaangetast beschikbaar blijven. Voor Dan Linstedt is deze eigenschap randvoorwaardelijk voor een goed Enterprise Datawarehouse.



Afbeelding 6: Spiegelziekenhuizen.

Conclusie

De manier van modelleren die we bij Prismant hebben voorgesteld biedt een uitstekend uitgangspunt voor het opzetten van echte Data Governance. IT rapporteert over de inconsistenties, de overlap, de gaps en toont de verbeteringen in de centrale registraties door de tijd, zodat de business op een natuurlijke manier eigenaar is van de centrale gedefinieerde begrippen. Schoenmaker IT blijft bij zijn leest.
De Data Vault modellering biedt ons een betrouwbare manier om een enterprise datawarehouse te modelleren. Door de regels van *natural world modeling* door te trekken en toe te passen, door relaties te expliciteren en registraties op de echte business key te respecteren creëert u een model dat Master Data Management en Data Governance kan faciliteren, zonder dat IT de verantwoordelijkheid daarvoor overneemt. Metadatarapporten kunnen de evolutionaire groei naar een correcte en centrale allesomvattende registratie tonen en die stimuleren zonder de lokale views aan te tasten en zonder interpretaties door IT.

Noten

1. *Het is goed om een onderscheid te maken tussen een begindatum (de dag dat de relatie aangegaan is) en de startdatum in SAT-tabellen, de start van de geldigheid van de opgesomde attributen.*
2. *Data Governance: a convergence of data quality, data management, business process management, and risk management surrounding the handling of data in an organizations (Wikipedia).*

Karien Verhagen

Drs. C. Verhagen is senior BI consultant bij 4BIS Scholing en Advies.

Met dank aan Henk Wierda en Johan Schreurs.