Bouwen enterprise datawarehouse-oplossing is niet altijd nodig

Procesuitvoering, -sturing en strategische informatie

Karien Verhagen

Dit artikel is een reactie op het artikel van Fons Pieters in Database Magazine 3, mei 2011. In dat artikel vraagt hij zich af of het in alle gevallen noodzakelijk is dat het EDWH gemodelleerd moet worden naar de bron.

In het concrete voorbeeld dat dit artikel wil schetsen vindt u overwegingen bij datamodellen voor verschillende doelen en hoe elk model die verschillende doelen het best kan dienen.

Feit of dimensie?

Warren Thornthwaite heeft me eens verteld dat een verhuizing zowel een feit kan zijn als ook een aanleiding tot het opnemen van een nieuwe historische versie van een klant. Als je een verhuisbedrijf bent is een verhuizing een feit in je datawarehouse, als je een retailer bent ligt dat anders. Zo'n opmerking illustreert dat datamodellering samenhangt met het bedrijfsproces. Een datamodel is nooit los te zien van het bedrijfsproces. Ook de functie die het datamodel vervult in het bedrijfsproces (te weten uitvoeren of bijsturen) is medebepalend voor de invulling van het model. De modellering van je datawarehouse hangt samen met de functie die een DWH heeft ten opzichte van de operationele systemen.

Een voorbeeld: een bezorgbedrijf bezorgt pakjes eventueel via verzamelpunten van A naar B en rekent die services af met de betalende klant. De betalende klant is niet noodzakelijkerwijs het bestel-, verzend- of ontvangadres. Die adressen worden ook vastgelegd.

Datamodellen voor snelle registratie

Voor de *uitvoering van de operatie* is het belangrijk dat de aankomsttijd en de vertrektijd snel en eenvoudig worden geregistreerd. Elke bezorging bevat een reeks verplaatsingen. Met bijvoorbeeld een scanner wordt de barcode van de bezorging of levering en de aankomst- of vertrektijd geregistreerd. De locatie wordt ook vastgelegd. Dat kan ook automatisch of eenmalig, voorafgaand aan het scannen. Deze werkwijze, ondersteund met het model zoals te zien in afbeelding 1, is prima voor de *procesuitvoering*.

Vragen als: Waar is mijn pakje nu?, Is het al van deze of gene lokatie vertrokken?, zijn met dit model te beantwoorden. Het zijn vragen om near real-time stuurinformatie. Iets lastiger maar nog steeds mogelijk zijn de volgende vragen: Hoe lang is het al onderweg? Overschrijdt de tijd de afgesproken norm?

Dit soort vragen is nog steeds te beantwoorden door een rapportagesysteem op de operationele database, al zijn voor die laatste vraag ook de afgesproken normtijden nodig.

Stuurinformatie uit 2 gescheiden systemen

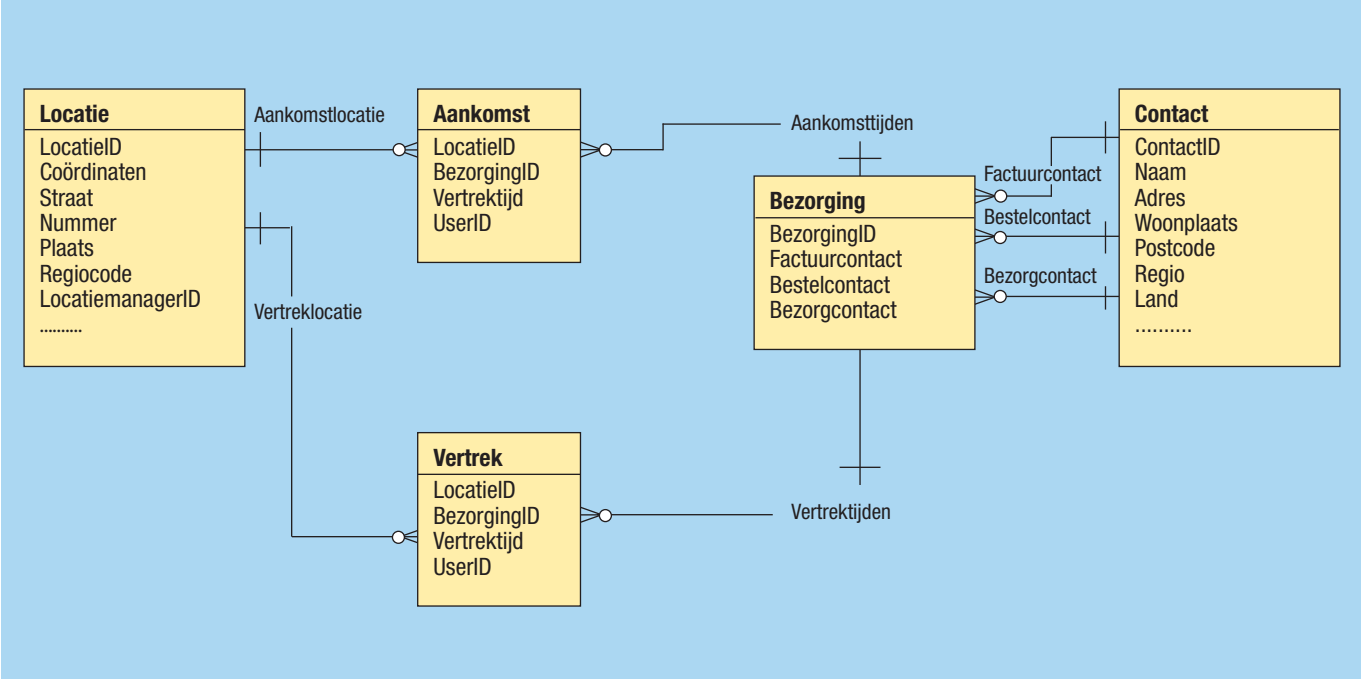
Stel nu dat dit systeem gekoppeld moet worden aan een van dit systeem gescheiden klachtensysteem om vragen voor de helpdesk op te lossen als: Is mijn pakje nog steeds niet vertrokken?, Ik heb daar een klacht over in gediend, wat is daarmee gebeurd?

Dan is een Operational Datastore met de gegevens van de gekoppelde operationele systemen voor actuele operationele stuurinformatie het overwegen waard. Dat heeft voordelen:

- een ODS zit de operatie niet in de weg;
- in een ODS kunnen het klachtensysteem en het logistieke registratiesysteem worden gekoppeld;
- het ODS wordt near real-time geladen en kan actueel zijn. Het kan ook een in-memory database zijn, het werkt bijvoorbeeld met EAI, niet met (batch-georiënteerde) ETL;
- het ODS-model is bedoeld voor raadplegingen en het datamodel kan daartoe worden geoptimaliseerd;
- als er (eventueel later) een datawarehouse wordt gebouwd kan dit vanuit het ODS geladen worden en zijn daarvoor de bronnen niet meer nodig;
- er is geen tijd voor bewerking van de data tijdens het transport, dat is voor actuele stuurinformatie ook niet nodig;
- theorieën omtrent het modelleren naar de bron zoals Data Vault zijn goed toepasbaar.

Als de actuele stuurinformatie uit meer systemen moet komen is een ODS te overwegen

Het ODS heeft in tegenstelling tot bijvoorbeeld een Data Staging Area altijd een gebruikersinterface en houdt die ook als er daarnaast een datawarehouse via datamarts wordt ontsloten.



Afbeelding 1: Optimaal model voor de procesuitvoering.

Strategische en tactische processturing

Pas als er vragen komen voor strategische sturing heeft een datawarehouse zin. Denkt u dan aan vragen als:

- Hoeveel pakjes zijn op tijd afgeleverd;
- Welke klanten worden slecht bediend;
- In welk segment bevinden die klanten zich;
- Welke locatiemanagers doen het goed en wat is afgesproken over hun prestaties;
- Hebben wij meer bezorgd tijdens de staking van de concurrent;
- Hoe is de ontwikkeling van het aantal en de zwaarte van de klachten uitgesplitst per locatie?

Een essentiële vraag voor het strategisch management is bijvoorbeeld ook de gemiddelde doorlooptijd per bezorging of verplaatsing.

Wanneer het strategisch en tactisch management dat soort vragen beantwoord wil zien, moet de realisatie gekoppeld worden aan de norm. Die normen hebben waarschijnlijk een tijdslijn. Er moet historie van worden bijgehouden. Op de operatie is de actuele norm voldoende.

Nu blijkt hoe onhandig het operationele model is voor dit soort vragen. Voor een gemiddelde doorlooptijd per bezorging moeten alle verplaatsingen bij elkaar worden gezocht, de begin- en eindtijden van elkaar afgetrokken en dan door het aantal verplaatsingen worden gedeeld. Een definitie bepaalt vervolgens wat met de wachttijd tussen aankomst en het volgende vertrek moet gebeuren. Ook de zwaarte van de verplaatsing (bijvoorbeeld de totale afstand over de weg) moet worden meegewogen. Dat kan via de normtijden. Dan moet de juiste historische waarde van de normendatabase daartegen worden afgezet. Zelfs Rick van der Lans zou moeite hebben om daar een vlot SQL statement voor te schrijven.

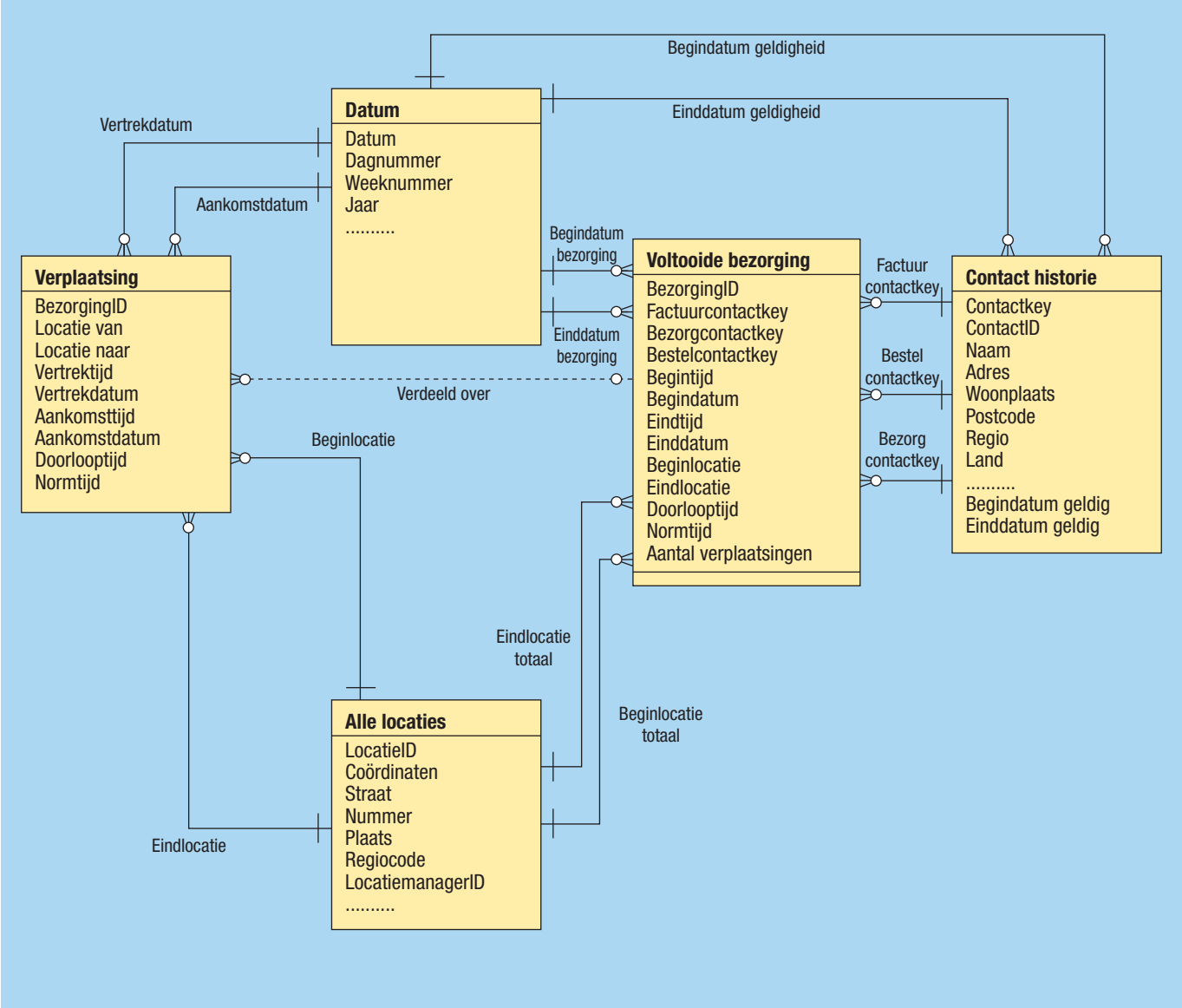
Een model met aankomsttijd en vertrektijd van alle voltooide bezorgingen en de (redundante) doorlooptijd, de start en de eindlocatie, met een optionele koppeling om ook de individuele verplaatsingen te zien van die bezorgingen zou zoveel geschikter zijn. Deze database zou dan alleen de voltooide, de geconsolideerde bezorgingen bevatten. De verplaatsingen zijn wellicht niet eens meer nodig en kunnen bijvoorbeeld met een *drill through* benaderbaar worden gemaakt. Zo'n model zou er uit kunnen zien zoals afbeelding 2 toont. De gewenste kengetallen om het succes van de operatie te peilen zijn met het model in afbeelding 2 veel eenvoudiger te achterhalen. De normtijden zijn hier in het feitenrecord platgeslagen.

Enterprise datawarehouse-laag altijd nodig?

Als dit model geladen wordt uit het ODS, dan is de vraag of daartussen nog een brongemodelleerd enterprise datawarehouse zou moeten zitten. Wat voor toegevoegde waarde heeft dat dan? Van der Lans' Data Delivery Platform is mede geboren uit onvrede met alle onbezonnen gebouwde lagen met kopieën van operationele gegevens. Ook het artikel van Fons Pieters bepleit een kritische blik ten opzichte van grote tussenlagen met hoofdzakelijk kopieën van operationele data om ze pas daarna te hermodelleren voor een optimaal gebruik als stuurinformatie. Zou het model zoals te zien in afbeelding 2 niet even geschikt zijn als (deel van) een enterprise datawarehouse-model?[1]

Conclusie

Een operationele database heeft procesuitvoering en registratie ten doel en is door haar taak minder geschikt voor de informatievoorziening, die tot doel heeft het proces te verantwoorden, te beheersen en te verbeteren. Het modelleren naar de bron levert



Afbeelding 2: Model geoptimaliseerd voor stuurinformatievoorziening.

een ongekende tijdwinst op bij het ontsluiten van een nieuwe bron als ook bij het periodiek, ongelijktijdig en snel laden van verschillende bronnen. Het hermodelleren van de data voor opvragingen wordt daarmee uitgesteld. Dat moet in een volgende laag alsnog gebeuren. Als zo'n volgende laag vervolgens weer verschillende datamarts moet bedienen is het de vraag of er niet teveel en dus onnodige tussenlagen ontstaan. Een kritische blik is wenselijk. Daar moet, zo beweert ook Pieters, dan wel de tijd voor zijn.

Een enterprise datawarehouse is nodig als er behoefte is aan tactische en strategische stuurinformatie met een bredere horizon en historisch inzicht. Tot die tijd is het operationele systeem zelf met een rapportagetool een prima oplossing. Als de *actuele* stuurinformatie uit meer systemen moet komen is een ODS te overwegen. Het ODS ontsluit near real-time operationele stuurinformatie ook over nog niet afgesloten transacties. Het bouwen van een enterprise datawarehouse-oplossing is (nog) lang niet altijd nodig. Wanneer de vraag naar corporate-brede strategische sturing die behoefte voldoende voedt is het de moeite waard om eerst goed na te denken over het datawarehouse-datamodel.

Discussies over een geschikt schaalbaar en toekomstbestendig datawarehouse-model zijn tijdrovend en intensief en dus duur. Over de voors en tegens van tussenlagen is de beroepswereld nog lang niet uitgepraat. Over het algemeen durf ik de stelling wel aan dat zuinig zijn met tussenlagen meestal onverstandig is. Maar je kunt ook overdrijven.

Noot

1. *Met dien verstande dat in een enterprise-brede oplossing een normendatabase met historie m.i. de voorkeur zou hebben.*

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