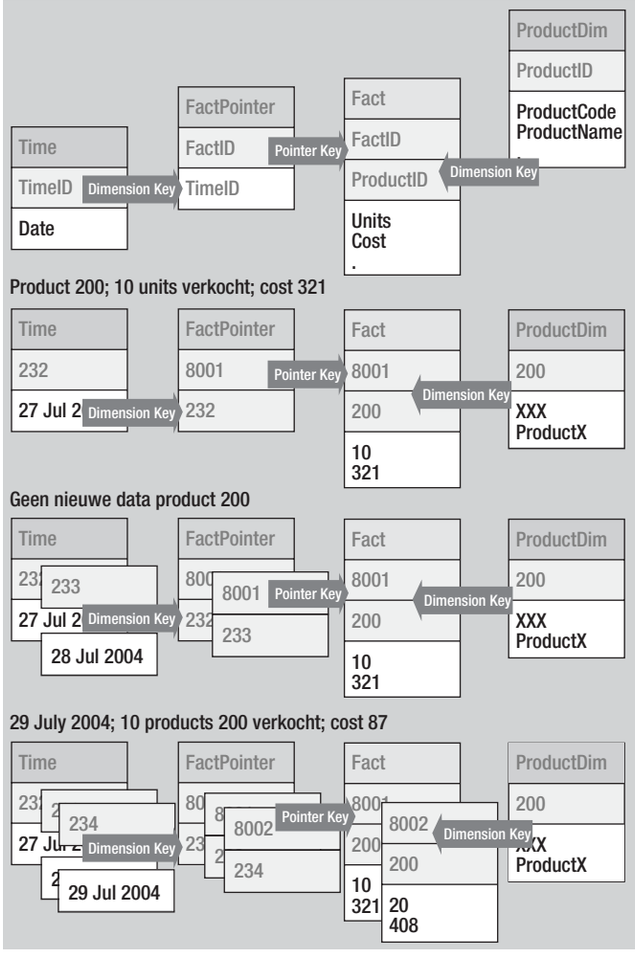
Methoden voor optimale informatievoorziening

Stand- en stroominformatie

Oscar Zonneveld

Er bestaan verschillende soorten informatievoorziening binnen een organisatie. De twee belangrijkste zijn standinformatie en stroominformatie. In dit artikel worden deze twee soorten informatie gepositioneerd en wordt ingegaan op hun eigenschappen in een datawarehouse-omgeving.

Voor het beslissingsproces binnen een organisatie zijn drie verschillende niveaus te onderscheiden, te weten strategisch, tactisch en operationeel. Medewerkers die acteren op deze niveaus hebben behoefte aan verschillende soorten informatie. Zo heeft een medewerker op strategisch niveau meestal overzichten over aantallen en portefeuillestanden op geaggregeerd niveau nodig om strategische beslissingen te kunnen nemen. Om op tactisch niveau beslissingen te kunnen nemen heeft een medewerker ondersteuning nodig van zowel aantallen als inzicht in het proces. De aantallen zijn dan weliswaar minder geaggregeerd dan de aantallen op het strategische niveau. De medewerker op operationeel niveau heeft tenslotte behoefte aan procesinformatie om beslissingen te kunnen nemen.



Afbeelding 1: Opbouw van een pointer-structuur.

Standinformatie

Kijkend naar de informatiebehoefte binnen een organisatie is men vaak geïnteresseerd in de informatie die aangeeft hoe een proces of de organisatie op een bepaald moment opereert. Dat betekent dus dat men op een bepaald moment een *snapshot* wil hebben van de totale hoeveelheid gegevens. Dit snapshot is een verticale doorsnijding van de aanwezige data in het datawarehouse en wordt ook wel standinformatie genoemd. Zo is het bijvoorbeeld mogelijk om de totale stand van een portefeuille op een bepaald moment te analyseren. Bij beschouwing van de karakteristieken van standinformatie wordt duidelijk dat er bij standinformatie een moment aanwezig moet zijn waarop naar de informatie wordt gekeken. Dit moment wordt vaak peilmoment of peildatum genoemd.

Het datawarehouse voor standinformatie kan op een aantal verschillende manieren worden ingericht. Deze verschillende manieren zijn onder te verdelen in:

1. Totale opslag in feitentabel;
2. Inrichten van pointer-structuur;
3. Feitentabel uitrusten als een 'slowly changing dimension'.

Totale opslag

Om een verticale doorsnijding mogelijk te maken van de informatie en vervolgens deze doorsnijding te relateren aan een moment (peilmoment), is het van belang dat de data volledig aanwezig zijn voor alle peilmomenten. Dit betekent dan ook dat bijvoorbeeld iedere dag de totale set aan data aanwezig dient te zijn in het datawarehouse. Het nadeel van deze methodiek is dat een grote hoeveelheid data opgeslagen dient te worden.

Het voordeel daarentegen is dat het laad- en transformatieproces voor het datawarehouse relatief eenvoudig is.

Bij standinformatie geldt dat niet alle feiten optelbaar zijn. Zo is het totaal aantal verkopen op peilmoment X niet optelbaar met het totaal aantal verkopen op dag X+1. Een voorbeeld: een organisatie heeft 5000 producten op 10 juni verkocht. 11 Juni worden er nog eens 1000 verkocht. Voor de standinformatie bekent dit dus dat op 11 juni 6000 verkochte goederen staan geregistreerd. Zou men besluiten om te aggregeren over de peildatum (in dit voorbeeld 10 en 11 juni) dan zou bij een optelbaar feit het totaal aantal verkochte goederen 11000 zijn. Om het juiste aantal verkochte goederen te rapporteren moet er dus worden gekeken naar het laatste peilmoment. Het totaal aantal verkochte goederen is dan 6000.

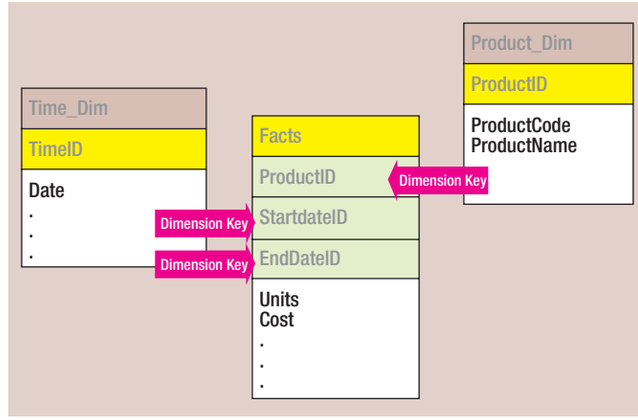
Pointer-structuur

Het nadeel van de totale opslag is de grote hoeveelheid data die opgeslagen wordt. Om de opslag te beperken zijn er andere methodes mogelijk. Een methode die hiervoor gebruikt kan worden is het inrichten van een pointer-structuur. Hierbij wordt extra functionaliteit toegevoegd aan het transformatieproces. Dit transformatieproces zorgt ervoor dat alleen de gemuteerde rijen worden getransformeerd richting het datawarehouse. Het transformatieproces van een pointer-structuur is dus complexer (zie afbeelding 1).

Bij een initiële download (voor de eerste vulling van het datawarehouse) worden de data getransformeerd richting het datawarehouse. In de feitentabel wordt een extra sleutel opgenomen die wordt gebruikt als 'surrogate key' (een nietszeggende sleutel te gebruiken voor referenties). In een pointer-tabel wordt voor ieder feit een rij opgenomen met daarbij de surrogate key van het betreffende feit en een referentie naar het peilmoment (peildatum). Aangezien in de pointer-tabel maar twee attributen zijn opgenomen, worden wel veel rijen in deze tabel opgeslagen, maar is de hoeveelheid bytes beperkt doordat er alleen twee surrogate keys in de betreffende tabel staan.

> Alleen de gemuteerde rijen worden overgezet naar het datawarehouse

Bij een volgende download wordt er in het transformatieproces, uitgevoerd door een ETL-tool, een schifting gemaakt tussen de rijen die niet gemuteerd zijn en de rijen die wel gemuteerd zijn. Alleen de gemuteerde rijen worden overgezet naar het datawarehouse. Het transformatieproces wordt zodanig ingericht dat na de transformatie van de feiten, de pointer-tabel gevoed wordt. Dit maakt het mogelijk om bij het vullen van de pointer-tabel een dataset samen te stellen die de actuele stand (in de bron-omgeving) representeert. Hierbij wordt dus voor iedere rij de laatste gedaante van de rij geregistreerd in de pointer-tabel voor het betreffende peilmoment.



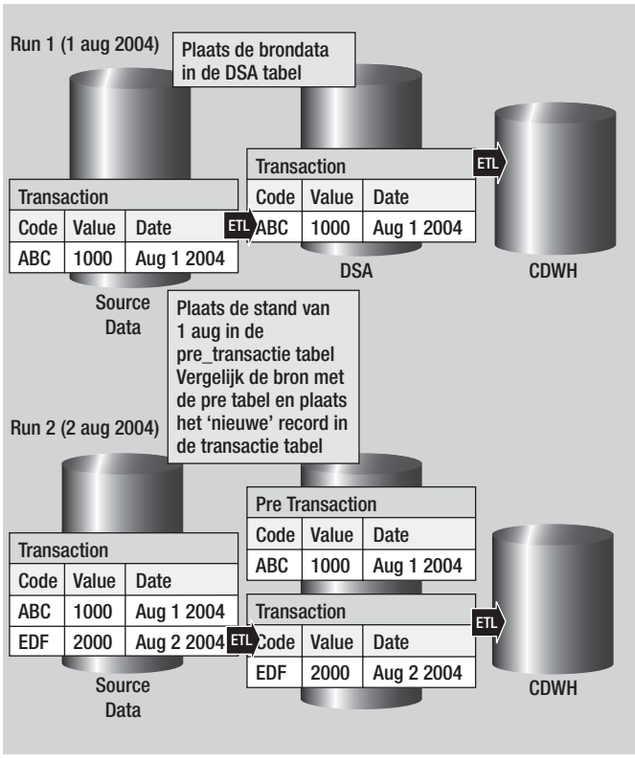
Afbeelding 2: Feitentabel als een slowly changing dimension.

Indien het transformatieproces onverhoopt een dag niet mocht lukken, zijn er vier verschillende scenario's te bedenken:

1. Men besluit om deze dataset leeg te laten;
2. Het is mogelijk om de dataset alsnog te transformeren;
3. Er is een extrapolatieproces om te bepalen wat het had moeten zijn;
4. De dataset van de voorgaande dag wordt gepresenteerd.

Feitentabel uitrusten als een slowly changing dimension

Een andere manier van registratie van standen is het uitrusten van de feitentabel met extra attributen, als het ware een slowly changing dimension. Door het toevoegen van zowel een begin- als einddatum is het mogelijk om voor een feit aan te geven wat de 'levensduur' van een bepaald feit is. Wanneer men dus de data wil bekijken is het van belang dat het peilmoment ligt tussen de begindatum en einddatum van een betreffend feit, zie afbeelding 2.
Bij de initiële download (voor de eerste vulling van het datawarehouse) worden de data getransformeerd richting het datawarehouse. Tijdens deze transformatie wordt de feitentabel gevuld met alle waarden en wordt er een begin- en einddatum vastgesteld. De einddatum ligt ver in de toekomst (bijvoorbeeld 31-12-9999). Wanneer de tweede datastroom komt, bepaalt het transformatieproces of de betreffende feiten reeds geregistreerd zijn. Indien dit het geval is wordt niets gedaan. Indien dit niet het geval is wordt de rij afgesloten (de einddatum wordt gevuld met de begindatum van het volgende feit). Bovendien wordt er een nieuwe rij in de feitentabel opgevoerd met als begindatum de datum vanaf wanneer deze geldig is (deze is dus gelijk aan de einddatum van het vorige voorkomen van het feit) en een einddatum die ver in de toekomst ligt.
Het nadeel van het gebruik van deze methodiek is dat een update in de feitentabel plaatsvindt. Deze update zorgt ervoor dat de feitenrij wordt 'afgesloten'. Aangezien een update-actie 'duurder' is dan het opvoeren van een nieuw record, moet worden onderzocht of een update 'performance-technisch' acceptabel is. Het voordeel bij deze methode is dat de dataopslag in de feitentabel



Afbeelding 3: De volledige transformatie.

beperkt blijft. Een ander voordeel is dat wanneer het transformatieproces om wat voor reden dan ook niet gelukt is, het toch mogelijk is om een beeld te schetsen van de stand voor die dag. Deze stand is echter niet de actuele. Door de gebruikers dient vervolgens te worden bepaald of dit acceptabel is.
Voor alle drie bovengenoemde methoden geldt dat voor de gebruiker het peilmoment (peildatum) één van de belangrijkste attributen is om te gebruiken in zijn of haar rapporten.

Stroominformatie

Indien men geïnteresseerd is in het verloop (workflow) van processen dan betekent dit dat men eigenlijk geïnteresseerd is in informatie met betrekking tot stroominformatie. Stroominformatie is de informatie die een gebruiker van de BI-applicatie inzicht geeft in de processen die men wil bewaken. Dit geeft de gebruiker vervolgens de mogelijkheid om mutaties van de afgelopen periode te monitoren en de voortgang binnen het proces zichtbaar te maken.
Vanuit de bronnen zijn twee verschillende soorten aanleveringen te onderscheiden, namelijk:

1. Volledige aanleveringen (alle informatie wordt doorgestuurd);
2. Incrementele aanlevering (alleen de mutaties worden doorgestuurd).

Volledige aanlevering

Indien er een volledige dataset wordt aangeboden, zorgt het ETL-proces ervoor dat vanuit de dataset de mutaties worden achterhaald. Afhankelijk van de dataset zijn hiervoor twee mogelijkheden. Indien de set een kenmerk bevat waarbij de incrementen te herkennen zijn (zoals bijvoorbeeld een mutatiedatum), dan kan dit worden gebruikt om het increment te bepalen. Indien dit niet het geval is, wordt er vaak gebruik gemaakt van een voorportaal. Bij een initiële download wordt de dataset in de uiteindelijke doeltabel geplaatst. Bij een volgende download wordt de huidige dataset (die bewaard blijft in de DSA-omgeving) in de DSA (Data Staging Area) getransformeerd naar het voorportaal (in afbeelding 3 de pre-omgeving). De nieuwe download wordt vervolgens in het ETL-proces vergeleken met de huidige set.
De mutaties die worden onderkend, worden opgeslagen in de 'originele' DSA-omgeving. Door het toepassen van deze techniek blijft de te transformeren dataset naar het datawarehouse beperkt.
Een voorbeeld: bij een initiële download wordt de data van 1 augustus vanuit de bron getransformeerd naar de DSA om vervolgens verder te worden getransformeerd. Bij elke volgende download wordt de vorige download (1 augustus) in zijn geheel in het voorportaal geplaatst en vergeleken met de actuele set.
De mutaties (nieuwe of verschillende records) worden vervolgens geplaatst in de DSA en daarna kan de beperkte set (mutaties) worden getransformeerd richting het datawarehouse.
Dit betekent dat de ETL-programmatuur in het transformatieproces moet gaan bepalen of het nieuwe gegevens zijn.
Het bepalen van de mutaties vereist veel meer resources in termen van disk- en geheugencapaciteit. Door het toevoegen van extra datavelden aan de records (zoals een hash-total) is het mogelijk om dit proces te optimaliseren.

Incrementele aanlevering

Wanneer de data vanuit de bron reeds incrementeel worden aangeleverd kan er op twee verschillende manieren mee worden omgegaan:

1. De DSA wordt steeds geschoond en het increment wordt hierin geplaatst;
2. Het increment wordt in de DSA-omgeving bijgelezen.

Indien voor de eerste mogelijkheid wordt gekozen kunnen de data relatief eenvoudig worden getransformeerd richting het datawarehouse. Het nadeel van deze techniek is dat detailinformatie (die opgeslagen is in de DSA) verloren kan gaan. Een voordeel is hierbij de beperkte opslag. Indien voor de tweede mogelijkheid wordt gekozen, wordt vaak extra informatie geregistreerd om de incrementen eenvoudig te herkennen en te kunnen transformeren richting het datawarehouse.

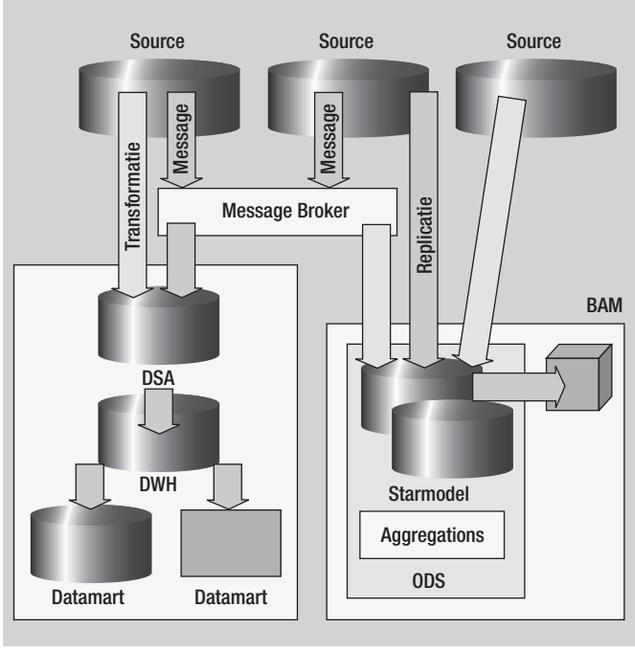
Bijvoorbeeld, bij een eerste download zijn er 100 transacties vanuit het bronsysteem getransformeerd in de DSA-omgeving. Door het opnemen van een extra veld dat automatisch nummert en door de minimale en maximale waarde op te slaan in een aparte transformatietabel, wordt geregistreerd 'hoe groot' het increment was. Wanneer een tweede increment wordt getransformeerd richting de DSA, wordt opnieuw de maximale waarde bepaald en is de grootte van het increment bekend.

Transformatie naar het datawarehouse

Nu de technieken om de incrementen te bepalen bekend zijn, moeten de data richting het datawarehouse worden getransformeerd. In tegenstelling tot het standen-datawarehouse heeft een stroom-datawarehouse geen peildatum. Bij stroom spreekt men van de transactiedatum (veelal een kalenderdatum).
Bijvoorbeeld: een organisatie heeft 5000 producten op 10 juni verkocht. 11 Juni worden er nog eens 1000 verkocht. Voor de stroom bekent dit dus dat op 11 juni 1000 verkochte goederen staan geregistreerd. Aangezien deze feiten optelbaar zijn, kunnen ze worden geaggregeerd.
Als we naar de feiten kijken die voor stroominformatie worden geregistreerd, worden deze doorgaans op een relatief laag detailniveau opgeslagen. Hierbij moet worden gedacht aan dagniveau of voor doorlooptijden zelfs minuut- of milliseconden-niveau. Bovendien zijn hier de feiten veelal wel optelbaar, in tegenstelling tot standinformatie.

Het verschil tussen stroominformatie en BAM

Als men globaal naar de doelstelling kijkt van stroominformatie en Business Activity Monitoring (BAM) lijkt dit op het eerste gezicht hetzelfde te zijn. Op detailniveau zijn er wel degelijk verschillen te onderkennen. Stroominformatie heeft als voornaamste doelstelling om trends te signaleren in het proces. Bovendien geeft het op een bepaald aggregatieniveau inzicht in het proces. De verversingsgraad van stroominformatie is dan ook vaak dagelijks of zelfs op een nog lager detail niveau, zoals wekelijks of maandelijks. BAM daarentegen geeft een getailleerd inzicht in de bedrijfsvoering op 'near real-time' niveau. Dit betekent dus dat op een near real-time basis de data vanuit de bronnen, door bijvoorbeeld replicatie, worden getransformeerd naar de BAM-omgeving.



Afbeelding 4: Voorbeeld BAM-omgeving.

Deze BAM-omgeving maakt het de eindgebruikers mogelijk om op near real-time basis inzicht te verkrijgen in de voortgang van het proces en daardoor om op korte termijn te kunnen bijsturen. In BAM-applicaties wordt dan ook veel gebruik gemaakt van verschillende alerts, die signalen afgeven naar de gebruiker indien een vooraf gestelde waarde is bereikt of overschreden.
Het verschil in het transformatieproces tussen BAM en stroominformatie ligt vooral in de tijdigheid die het transformatieproces verlangt. BAM-transformatieprocessen zijn doorgaans gestoeld op replicatietechnieken.
De methodiek van de BAM-omgeving wijkt af van de methodiek voor stroominformatie. Waar de stroominformatie voor een langere tijd wordt opgeslagen (voor bijvoorbeeld trends), worden voor BAM de data vanuit de bron (door bijvoorbeeld replicatie) getransformeerd naar een ODS (Operational Data Store) en daarin gedurende een korte periode opgeslagen.
De data worden door middel van BAM-applicaties aan de gebruikers beschikbaar gesteld. Dit zijn meestal specifiek ontwikkelde applicaties. In de markt worden steeds meer producten uitgerust met BAM-functionaliteit. Een voorbeeld hiervan is Microsoft BizTalk Server 2004 met Microsoft Excel als front-end. Voor stroominformatie kan gebruik worden gemaakt van de vaak al aanwezige BI-suites.

Conclusies

Om te bepalen welke soort informatievoorziening en welke daarbij behorende methodiek dient te worden gekozen, is een aantal zaken van belang. Zo is het van belang om te onderkennen wie de gebruiker is. Indien de gebruiker zich op strategisch niveau bevindt, is de vraagstelling veelal gebaseerd op standinformatie. Het gaat hierbij vaak om einde maand- of einde jaar-standen. Voor het tactische niveau is standinformatie alleen niet voldoende. Gewoonlijk is op tactisch niveau ook behoefte aan stroominformatie om de juiste beslissing te kunnen nemen. Op tactisch niveau worden doorgaans de lange termijnbesluiten uitgevoerd en wordt de korte termijnstrategie ontwikkeld. Voor het ontwikkelen van deze strategie is doorgaans standinformatie en inzicht in de voortgang van de processen nodig.
Op operationeel niveau wordt de korte termijnstrategie uitgevoerd en worden ad hoc-besluiten genomen. Voor het onderbouwen van deze ad hoc-besluiten wordt grotendeels stroominformatie gebruikt. Deze stroominformatie moet de gebruiker inzicht geven in de voortgang van de processen.
Om tot de optimale informatievoorziening te komen zijn twee punten van essentieel belang: enerzijds het vaststellen van het beslissingsniveau van de eindgebruiker (strategisch, tactisch of operationeel) en de daarbij behorende informatievoorziening (stand- of stroominformatie) en anderzijds het vaststellen van de optimale methodiek ter ondersteuning van de gekozen informatievoorziening.

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