3/22/2018

Patrick van Batenburg (2125056)

Een datawarehouse voor Kramse

Het nieuw te ontwikkelen datawarehouse

Inhoudsopgave

[Samenvatting 2](#_Toc509665245)

[1. Inleiding 2](#_Toc509665246)

[2. De TPS omzetten naar een datawarehouse 2](#_Toc509665247)

[2.1. De gegevens importeren en integreren in een database 2](#_Toc509665248)

[2.2. Het ontwerp van het ster-schema 3](#_Toc509665249)

[2.3. Het implementeren van het sterschema 5](#_Toc509665250)

[2.4. Het vullen van de ODS-database vanuit de PSA-database 5](#_Toc509665251)

[3. Resultaten 5](#_Toc509665252)

[4. Conclusie 5](#_Toc509665253)

[5. Discussie 5](#_Toc509665254)

[Bijlage 1: Het SQL-script van de ODS-database 6](#_Toc509665255)

# Samenvatting

Het containervervoer bedrijf Kramse wilt graag de gegevens die ze hebben omzetten naar een datawarehouse. In dit verslag wordt beschreven welke stappen zijn ondernomen en de keuzes daarin om een datawarehouse te kunnen realiseren voor Kramse. Er zijn elnkele bevindingen gevonden tijdens het realiseren van de datawarehouse voor Kramse. Deze zijn onder meer welke dimensie- en feiten-tabellen zijn opgenomen en welke attributen zijn opgenomen in deze tabellen. Het belangrijkste zijn de keuzes die gemaakt zijn en die uitgelegd worden hoe deze keuzes tot stand zijn gekomen. Deze bevingen zijn voornamelijk het belangrijkst en cruciaal voor Kramse, omdat deze de structuur van het nieuwe datawarehouse beschrijven en laat zien hoe deze datawarehouse zou moeten gaan werken. Verder worden de keuzes die in de bevindingen plaats vinden onderbouwd over hoe deze tot stand zijn gekomen waar Kramse het eens of oneens kan zijn.

# Inleiding

Het bedrijf Kramse verzorgt het vervoer van zeecontainers tussen havens in Europa. De hiervoor benodigde gegevens komen uit een aantal bronnen. De kerngegevens komen uit een Microsoft Access database. De gegevens over opdrachtgevers komen uit een Excel-bestand. De gegevens over soorten containers komen uit een tekstbestand.

In grote lijnen wil het bedrijf Kramse een operationele database met enkele bijbehorende andere bestanden. De TPS moet dus omgezet worden naar een datawarehouse op basis van SQL Server. De hoofdvraag luidt daarom als volgt: “Kan er een Datawarehouse voor Kramse gerealiseerd worden?”

De functie van dit verslag is om in beeld te brengen hoe de opdrachten zijn uitgevoerd, opgenomen zijn. Wat verder belangrijk is zijn de keuzes die daarin naar voren gekomen zijn en hoe deze tot stand zijn gekomen.

# De TPS omzetten naar een datawarehouse

In dit hoofdstuk gaan er de bevinden bij het maken van de PSA-database en de ODS-database voor Kramse gedocumenteerd worden. In paragraaf 2.1 zal er beschreven worden op welke manier de 3 bronbestanden in een database geïntegreerd zijn. Na deze paragraaf komt paragraaf 2.2. In deze paragraaf zal er beschreven worden welke dimensie- en feiten-tabellen en welke attributen in welke tabel opgenomen zijn in het ster-schema. Hierbij zal er verder uitgelegd worden hoe deze keuzes gekomen zijn. De paragraaf die hierna volgt is paragraaf 2.3. Hierin zal beschreven worden hoe de ODS-database (structuur) gegenereerd kan worden.De paragraaf die daarna volgt is paragraaf 2.4. Hierin zal beschreven worden welke opschoning en transformaties van gegevens (Cleaning, Transform) er toegepast is op de gegevens. Verder zullen deze uitgelegd worden met hoe de keuzes daarin gekomen zijn. Als laatst wordt er besproken hoe en waarom Slowly Changing Dimensions (SCD) ingericht/toegepast zijn in de tabellen.

## De gegevens importeren en integreren in een database

Als allereerst ging ik de **2007 Office System Driver: Data Connectivity Components** installeren die het mogelijk maken om andere applicaties te laten communiceren met Microsoft Access databases.

Nadat deze installatie voltooid was opende ik het .accdb bestand in Microsoft Access. Hierna ging ik in het lint naar “Database Tools” en ging ik vervolgens op “Access Database” te klikken. Hiernavolgende de wizard om de database te splitten en wat het dus makkelijker zou gaan maken om de database te kunnen importeren. Nadat de wizard voltooid was en het een nieuw Access database bestand aangemaakt had die de gesplitte database bijhoudt, ging ik Visual Studio openen om vervolgens een Integration Services Project aan te maken. In dit project maakte ik drie SSIS-pakketten aan voor het Access, Excel en tekstbestand dat ik aangeleverd had gekregen. Door de wizard te volgen kon ik makkelijk de SSIS-pakketten aanmaken voor ieder bestand. Vervolgens werden deze pakketten uitgevoerd, waardoor de gegevens die in de bronbestanden waren, terecht zouden komen in de KramsePSA database die hiervoor special aangemaakt was in SQL Server.

## Het ontwerp van het ster-schema

Zoals in Figuur 1 te zien is, is de Ship tabel opgenomen als een dimensie en zijn de attributen overgenomen zonder andere aanpassingen aan deze tabel. Hetzelfde geld ook voor de Consignor, Item en Container tabellen. Deze zijn ook opgenomen als een dimensie met dezelfde attributen. Verder zijn de tabellen Port, Voyage en VoyagePort alle drie platgeslagen tot een dimensie tabel, omdat ze relaties met elkaar hadden. Hierdoor zijn alle attributen bij elkaar samengevoegd. Hoewel de attribuut PortOrder niet twee keer voorkomt, ook al was deze in de VoyagePort tabel als in zowel in de Port tabel. Deze attribuut was echter alleen overbodig in de VoyagePort tabel ook doordat hier geen enkele gegevens vermeld waren. Verder is er een tijd dimensie voor de efficiëntie, omdat datums bereken veel tijd kan kosten. Als laatste dimensie is er de landen dimensie die de landen bijhouden doormiddel van afkortingen, landcodes en landnaam in het Nederlands als zowel in het Engels. Dit is vanwege dat de namen van landen op verschillende manieren zijn opgenomen in tabellen, in de eigen taal of Engels, al dan niet afgekort. Voor alle dimensies behalve voor de Calendar en Country dimensies heeft iedere andere dimensie de attributen START\_DATE en END\_DATE gekregen vanwege Slowly Changing Dimensions (SCD). De Shipment tabel werd de feiten tabel, doordat het al veel relaties met andere tabellen had en belangrijke informatie inhoud. De standaard attributen werden behouden en er kwamen natuurlijk veel nieuwe attributen bij die de sleutels van de andere dimensie tabellen bij moeten gaan houden.



Figuur 1: Het ster-schema

## Het implementeren van het sterschema

Om de structuur van de ODS-database te kunnen implementeren zoals in paragraaf 2.2 is beschreven met het ster-schema, is er SQL-script te vinden in bijlage 1. Door deze SQL-script uit te voeren kan de structuur van de ODS-database structuur gemakkelijk gegenereerd worden.

## Het vullen van de ODS-database vanuit de PSA-database

Bij een heleboel tabellen waren de data types van bepaalde vanuit de PSA-database niet correct voor de ODS-database en hiervoor moesten deze attributen geconverteerd worden naar hun goede format en data type waar nodig was. Slowly Changing Dimensions (SCD) werd toegepast op iedere tabel behalve voor de feiten, tijd en landen-tabellen. Attributen die hier van het data type Date en Money waren werden voornamelijk gezien als een Historical attribuut. De meeste attributen van de tabellen werden voornamelijk Historical attributen, maar sommige waarbij sommige attributen weleens konden veranderen werden gezien als Changing attributen en hierdoor zijn er geen Fixed attributen aan de orde gekomen. De tabellen die over de START\_DATE en END\_DATE attributen beschikken zijn de tabellen waar Slowly Changing Dimensions (SCD) op toegepast werd. Natuurlijk werd Slowly Changing Dimensions op deze attributen niet toegepast. Voor het maken van de Calendar dimensie is er een Analysis Servives Multidimensional and Data Mining Project aangemaakt in Visual Studio waarmee je met een dimensie wizard makkelijk een tijd tabel kan aanmaken. Hierbij waren alle standaard waardes gelaten zoals ze waren en waren alle tabellen die genereerd konden worden ook aangevinkt. Verder was er gekozen om records aan te maken tussen de jaren 2015 en 2018.

# Resultaten

De resultaten die hieruit naar voren zijn gekomen is dat een ODS-database voor Kramse gerealiseerd kan worden doormiddel van het SQL-script (zie bijlage 1). Hiermee blijk dus dat we antwoord hebben gekregen op onze hoofdvraag. De hoofdvraag luidde als volgt: “Kan er een Datawarehouse voor Kramse gerealiseerd worden?”. Verder zijn de keuzes die gemaakt zijn terug te vinden in hoofdstuk 2 waar Kramse mogelijk veel aan kan hebben, omdat hierdoor gegevens die in de ODS-database komen van goede kwaliteit zullen zijn en consistent zullen gaan blijven. Verder is in hoofdstuk 2 de structuur van beschreven en visueel weergegeven door een ster-schema.

# Conclusie

Uit de resultaten bij het maken van de ODS-database voor Kramse is gebleken dat deze database structuur voor Kramse makkelijk te realiseren en implementeren valt. Deze resultaten worden onderbouwd door de keuzes die genomen zijn en terug te vinden zijn in hoofdstuk 2 samen met een ster-schema dat de structuur van de ODS-database visueel weer zal geven. De resultaten hebben tot slot uitgewezen dat met de structuur van de ODS-database de gegevens die hierin komen over goede kwaliteit zullen beschikken en consistent zullen zijn. Hieruit is gebleken dat Kramse een succesvolle ODS-database makkelijk kan opzetten met het SQL-stript dat terug te vinden is in bijlage 1.

# Discussie

Doordat de brongegevens eerst geïntegreerd moesten gaan worden in een PSA-database kan op basis hiervan gesteld worden dat bij een herhaling hiervan de resultaten hetzelfde zouden zijn en dat daarmee de resultaten van dit onderzoek valide zijn.

Hierbij moet wel rekening gehouden worden dat dit onderzoek zich uitsluitend heeft gefocust op het maken van een datawarehouse op basis van de ster-schema structuur. Zodra gefocust wordt om enkele tabellen wel te gaan normaliseren dan zouden de resultaten mogelijk anders zijn. Om deze reden kan geen algemene uitspraak gedaan worden over welk alternatief datawarehouse structuur een goede datawarehouse kan bieden voor Kramse.

Het advies voor vervolgonderzoek is dan ook om een soortgelijk onderzoek uit te voeren om te achterhalen of het verschil tussen een alternatief datawarehouse structuur hetzelfde invloed heeft als gefocust wordt op andere factoren dan die in dit onderzoek aan de orde zijn gekomen.

# Bijlage 1: Het SQL-script van de ODS-database

In deze bijlage is het SQL-script te vinden waarmee de ODS-database (structuur) gegenereerd kan worden.

