PROTOTYPE DATA WAREHOUSE EN BUSINESS INTELLIGENCE

The Cloud Consultants

Versiebeheer

Versie	Auteurs	Datum	Opmerking
0.1	A. van Dalen S.A. Twardowski	02-11-2020	Eerste versie
0.2	A. van Dalen S.A. Twardowski	26-11-2020	Final draft

Inhoudsopgave

Introductie

Na het uitvoeren van onderzoek naar een data warehouse en business intelligence wordt er een prototype gemaakt om aan te tonen welke waarde deze technieken kunnen bieden voor de organisatie. Er zal worden toegelicht welke keuzes zijn gemaakt voor de gebruikte data, de business intelligence vragen, het gekozen model wat hierbij past en de gebruikte software om het prototype te bouwen.

Brondata

Om een goede voorstelling te maken van een data warehouse en business intelligence is het van belang om een realistische dataset te gebruiken die overeenkomt met een typisch bedrijf. De gekozen dataset voor dit prototype is de AdventureWorks dataset van Microsoft.

De AdventureWorks dataset is speciaal ontwikkeld voor leer- en prototype doeleinden en geeft een realistische voorstelling van een typische database welke binnen een bedrijf wordt gebruikt. De dataset bevat data over bijvoorbeeld personeel, winkels, klanten, producten, en orders. Tevens bevat de dataset data over verschillende jaren waardoor deze data goed is in te zetten voor een prototype van een datawarehouse.

De AdventureWorks dataset is in verschillende formaten te verkrijgen waaronder een operationele database en een datawarehouse. Voor het prototype maken wij gebruik van de operationele database als brondata. Deze data wordt met behulp van ETL-tools getransformeerd en geïmporteerd in een nieuwe datawarehouse.

Business intelligence vragen

Om een goed beeld te krijgen wat business intelligence voor de organisatie kan betekenen is het van belang de juiste informatie vragen te stellen. Wat van belang is bij het opzetten van de vragen is dat de brondata aansluit bij de te verkrijgen informatie. Wanneer de bron niet over de juiste data beschikt moet de vraag worden aangepast of moet er meer brondata worden verzameld welke aansluit bij de informatiebehoefte van de organisatie. Tevens moeten ook de vragen natuurlijk aansluiten bij de informatiebehoefte van de organisatie.

Om een goed beeld te geven wat business intelligence kan betekenen voor de organisatie hebben wij de volgende vragen opgesteld welke nieuwe inzichten kunnen geven aan het management:

Vraag 1: Welke producten en productcategorieën worden per periode het meest verkocht per locatie?

Deze business vraag zal inzicht geven in welke producten populair zijn op bepaalde periodes en bepaalde locaties. Deze vraag bevat meerdere dimensies: productcategorie, tijd en plaats.

De tweede business vraag die we zullen behandelen is:

Vraag 2: Wat is de omzet per winkel, per periode per categorie?

Deze informatie kan inzicht geven in welke winkels bepaalde categorieën producten afnemen en hier bijvoorbeeld kortingsacties of marketing op aanpassen. Ook deze vraag bevat meerdere dimensies: winkel, tijd en productcategorie.

Modellen

Om goed antwoord te geven op de verschillende vragen vanuit de business is het belangrijk de vraag te analyseren en te bepalen welke data hierbij nodig is. Tevens is het van belang een juist schema te maken wat bij de vraag vanuit de business aansluit om zo de juiste informatie uit het systeem te krijgen. Voor beide vragen wordt de benodigde data en het gemaakte schema besproken.

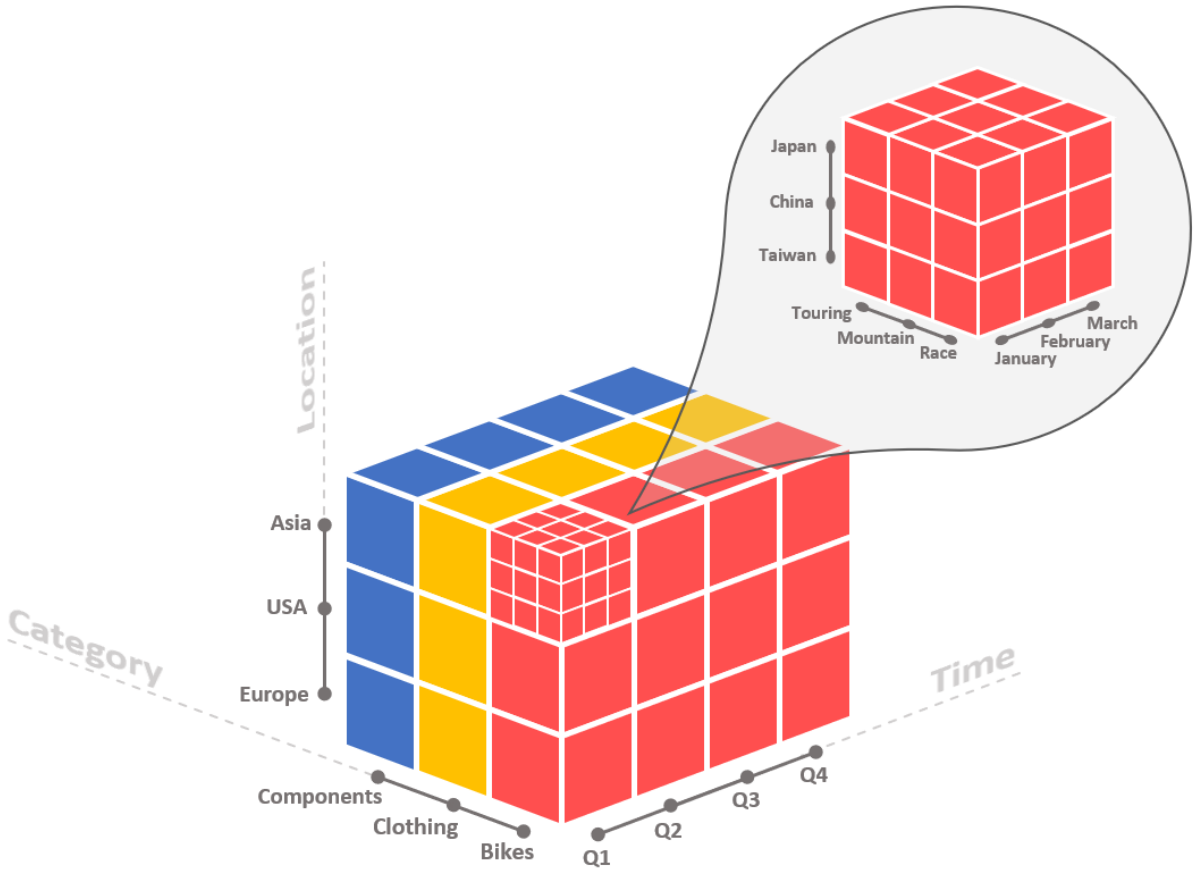
Vraag 1: Welke producten en productcategorieën worden per periode het meest verkocht per locatie?

De data die bij deze vraag hoort is:

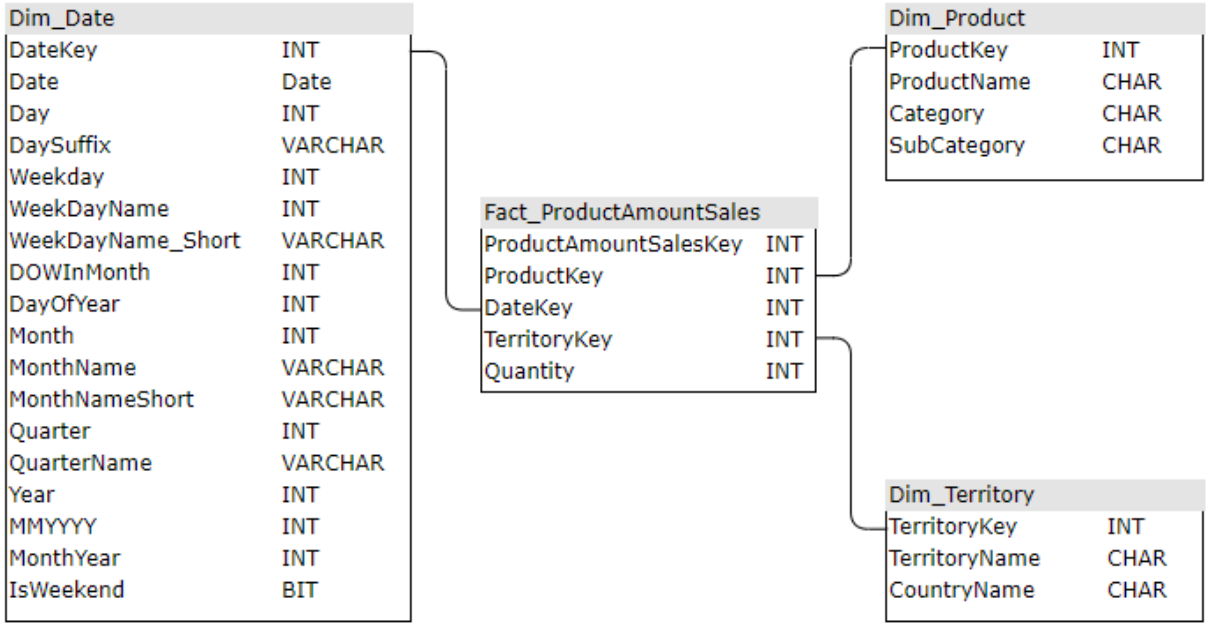
- Producten
- Productcategorieën en sub categorieën
- Tijd
- Locatie

Deze vier soorten data zijn de verschillende dimensies die aan het model worden gegeven. In de feiten tabel worden de aantallen opgenomen om een goed antwoord te kunnen geven op de vraag.

Een mooi beeld over de dimensies van de informatie is te zien in onderstaand kubus diagram.



Het ster schema die bij deze vraag hoort ziet er als volgt uit.



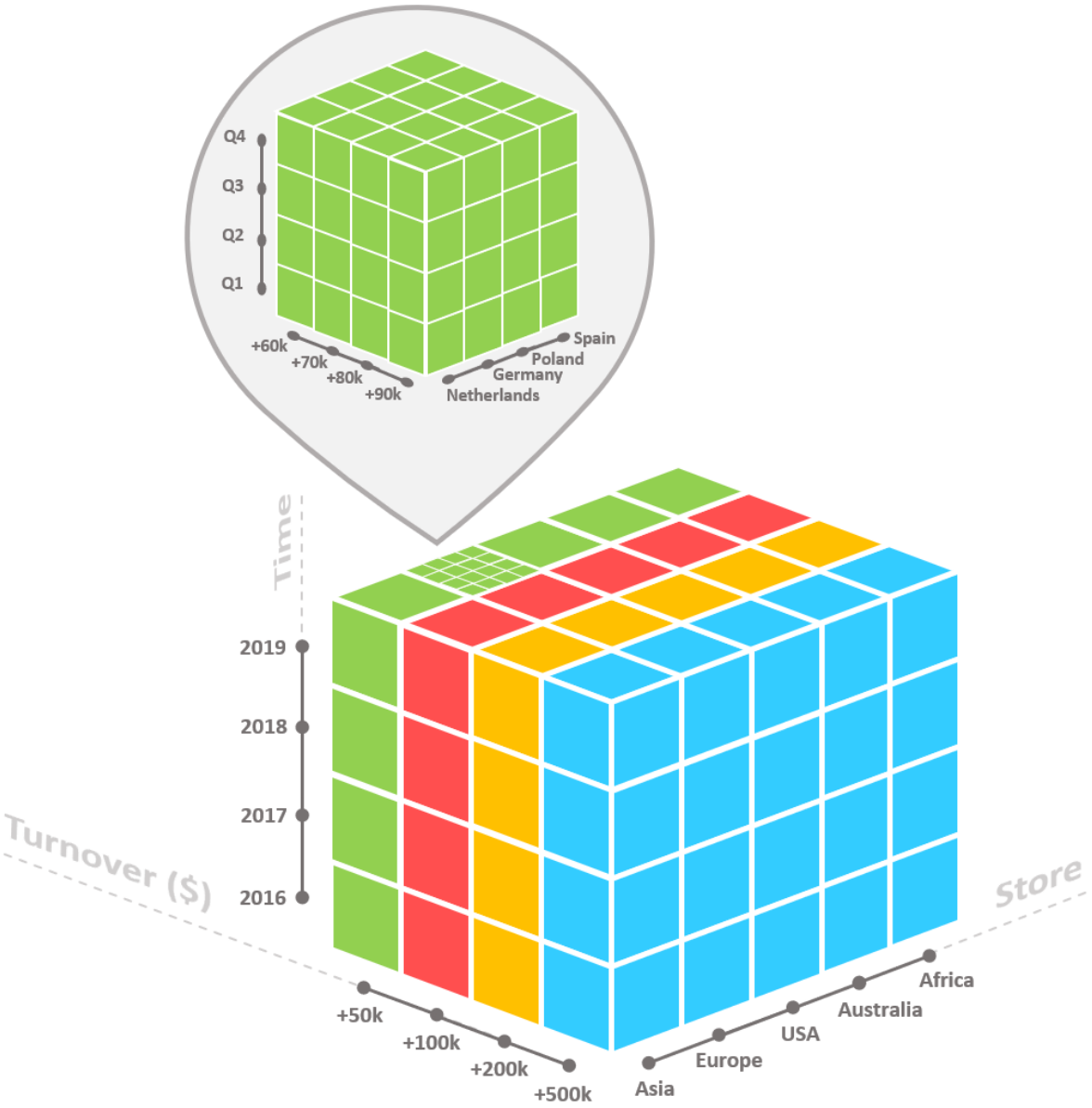
Vraag 2: Wat is de omzet per winkel, per periode per categorie?

De data die bij deze vraag hoort is:

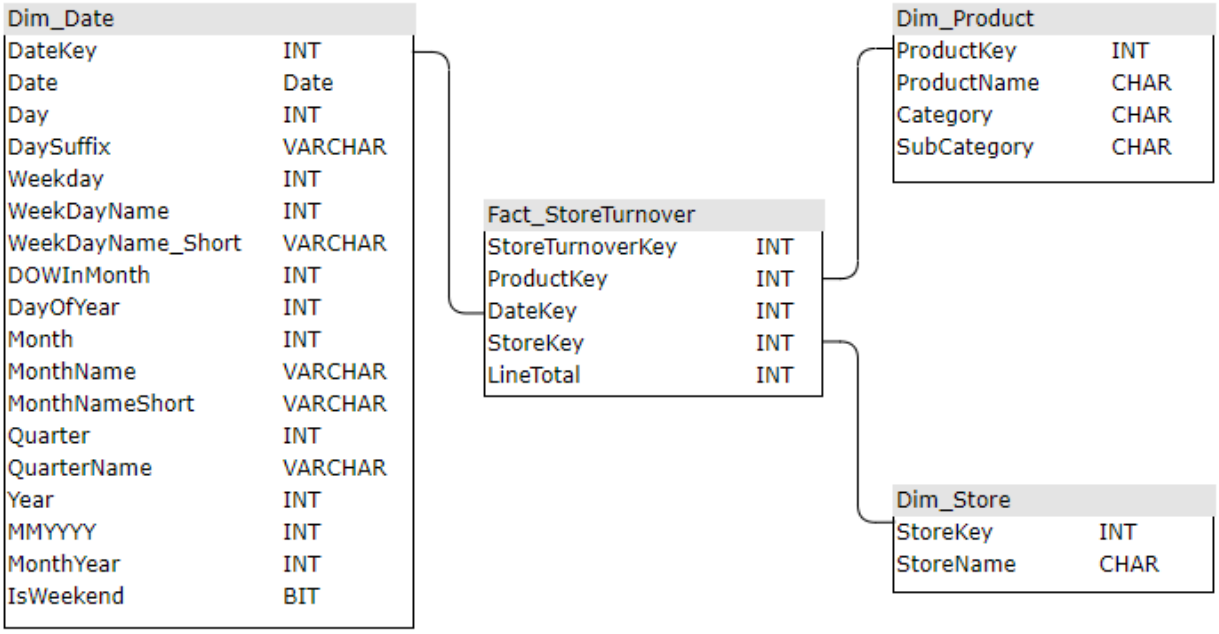
- Omzet
- Productcategorie en subcategorie
- Tijd
- Winkel

Deze vier soorten data zijn de verschillende dimensies die aan het model worden gegeven. In de feiten tabel worden de aantallen opgenomen om een goed antwoord te kunnen geven op de vraag.

Een mooi beeld over de dimensies van de informatie is te zien in onderstaand kubus diagram.



Het ster schema die bij deze vraag hoort ziet er als volgt uit.



Proces

Om het prototype mogelijk te maken is er gebruik gemaakt van verschillende softwarepakketten om het volledige proces van importeren tot visualisatie uit te kunnen voeren. Het proces is op te delen in de volgende stappen:

ETL – in deze stap wordt de brondata geëxtraheerd, getransformeerd en geladen in de data warehouse.

Publicatie – er worden kubussen gemaakt om de juiste data in de data warehouse te publiceren om zo de juiste informatie boven water te krijgen

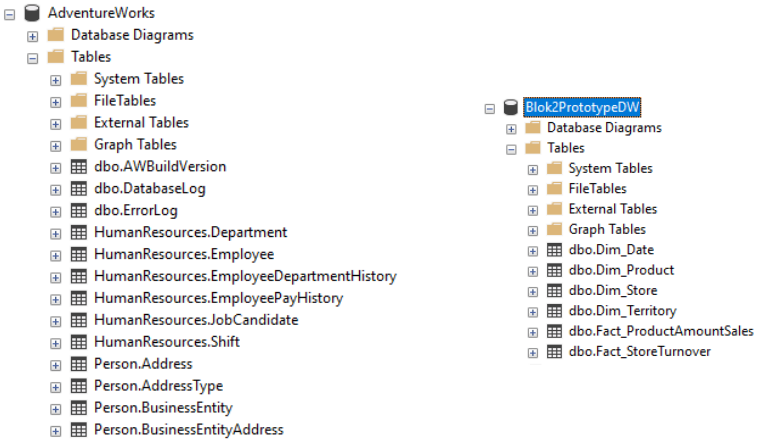
Visualisatie – in deze laatste stap wordt de informatie weergegeven die antwoord geeft op de gestelde business vraag. In deze stap kunnen dimensies worden gekozen, granulatie worden gekozen en informatie worden weergegeven.

Software

Voor alle stappen in het proces is software benodigd die aansluit bij de behoefte van de opdrachtgever. Naast deze software is er ook gekozen voor een database softwarepakket. Voor het prototype kan het zijn dat het slimmer is om bijvoorbeeld gratis of offlinesoftware in te zetten. In een productieomgeving kan het zijn dat andere softwareoplossingen beter passen bij de wensen van de opdrachtgever. De software die in dit hoofdstuk wordt genoemd is gebruikt voor het maken van het prototype.

Database

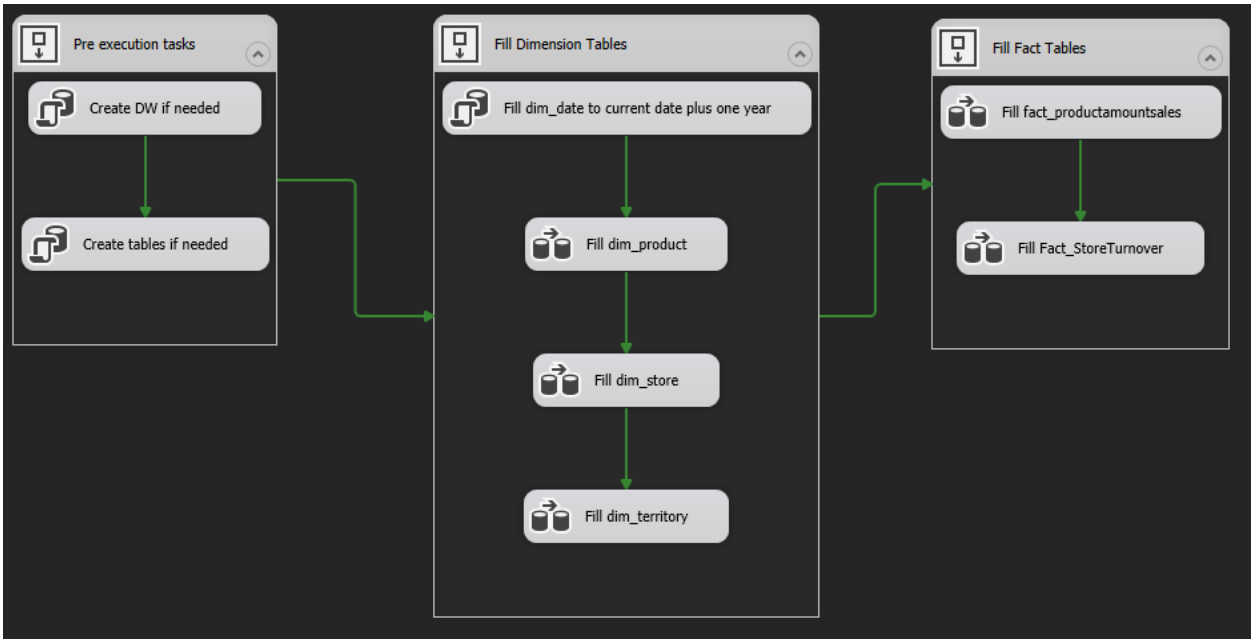
Voor de database software is gekozen voor SQL Server van Microsoft. Deze software is gratis te gebruiken tot een bepaalde grootte en is ruim voldoende voor het prototype. Er is al veel kennis van dit product wat heeft geholpen bij de doorlooptijd van de ontwikkeling van het prototype. De software wordt gehost op een lokale machine. Zowel de brondata als de data warehouse worden gehost op dezelfde SQL Server.


AdventureWorks
Database Diagrams
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.AWBuildVersion
dbo.DatabaseLog
dbo.ErrorLog
HumanResources.Department
HumanResources.Employee
HumanResources.EmployeeDepartmentHistory
HumanResources.EmployeePayHistory
HumanResources.JobCandidate
HumanResources.Shift
Person.Address
Person.AddressType
Person.BusinessEntity
Person.BusinessEntityAddress
Blok2PrototypeDW
Database Diagrams
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.Dim_Date
dbo.Dim_Product
dbo.Dim_Store
dbo.Dim_Territory
dbo.Fact_ProductAmountSales
dbo.Fact_StoreTurnover

ETL

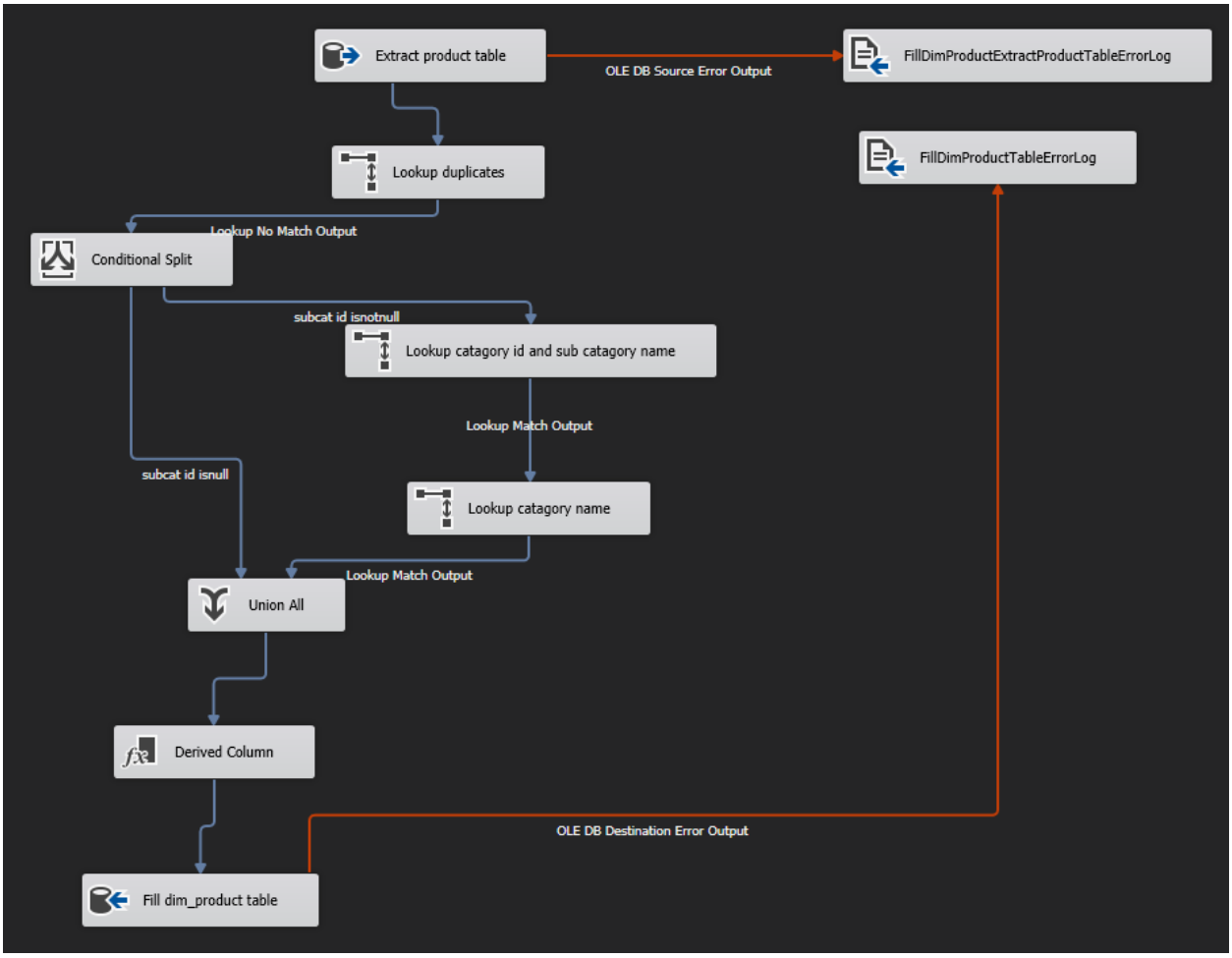
Voor deze stap is de keuze gevallen op SQL Server Integration Services van Microsoft. Deze software sluit goed aan bij de gekozen database en is gratis te gebruiken. Voor het maken van een prototype is dit een goede keuzen. Er kunnen binnen SSIS verschillende bronnen worden aangesproken, voor het prototype maken we gebruik van een SQL-database bron welke gehost wordt op een MSSQL-server.

SSIS voorziet in de import, de transformatie en het laden van de data in de data warehouse. Het ETL-proces ziet er als volgt uit.



In de pre execution fase wordt gecontroleerd of de data warehouse bestaat en of de tabellen ion de warehouse bestaan. Indien dit niet het geval is worden de data warehouse of de tabellen aangemaakt.

In de fill dimension fase worden de dimensie tabellen gevuld. Dit wordt gedaan doormiddel van data flows. Een data flow is een sub proces met een bron, een verwerking en een doel. Hieronder wordt de Fill dim_product data flow weergegeven.



Het proces begint bij de bron, in dit geval een SQL database. Vervolgens worden in de lookup, conditional split, union all en derived column stappen transformatie uitgevoerd. Als laatste wordt die nieuwe data weggeschreven in de data warehouse. Regels met errors worden weggeschreven naar een log bestand. Naast het loggen van afzonderlijke elementen wordt het complete proces ook gelogd naar een csv bestand.

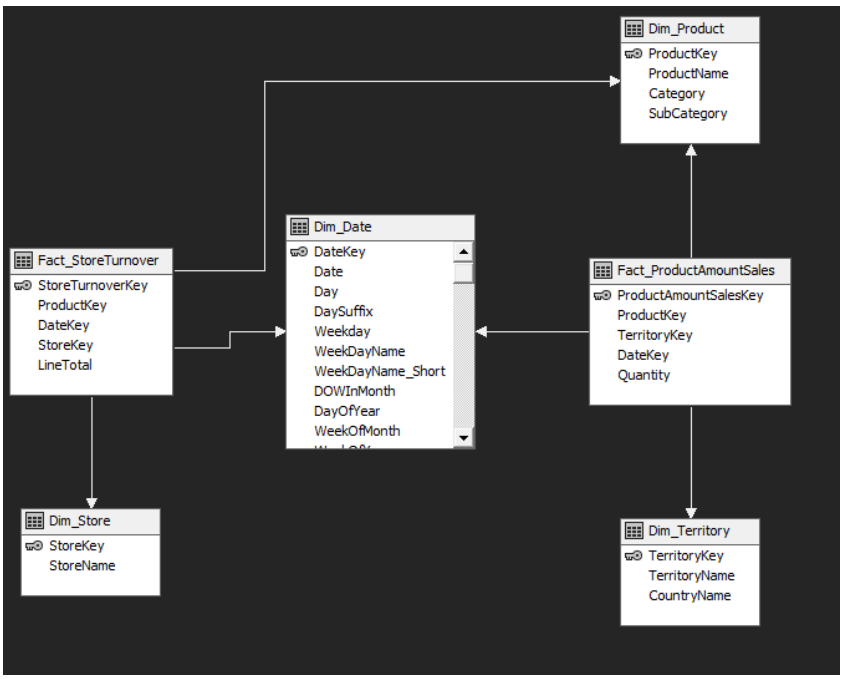
In de fill fact table fase worden de feiten tabellen gevuld met de juiste data. Deze data flows lijken erg op de bovenstaande.

Publicatie

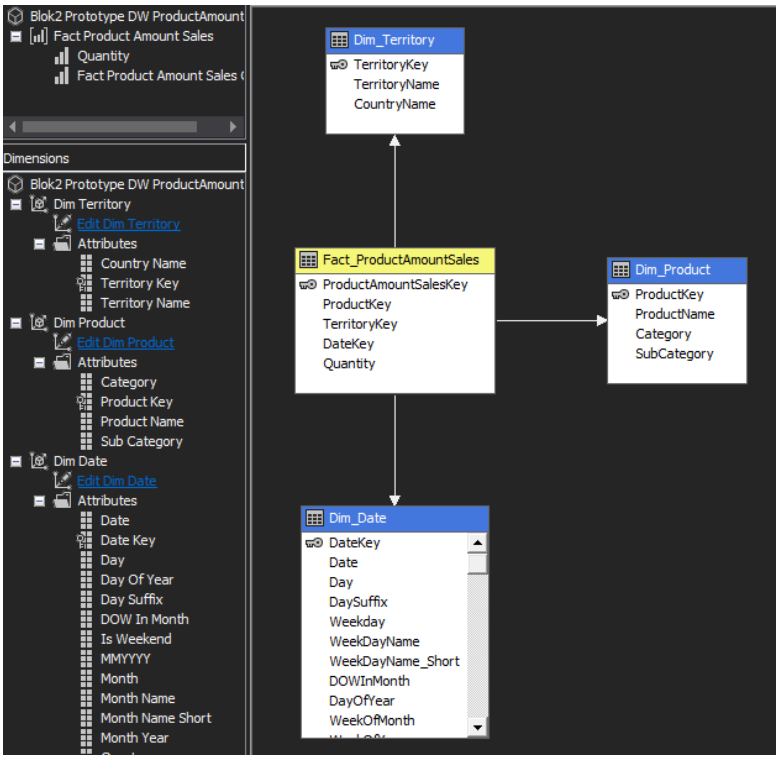
In deze stap wordt de data gepubliceerd zodat de informatie uiteindelijk kan worden gevisualiseerd. Voor de publicatie is gekozen voor SQL Server Analysis services van Microsoft. Deze software sluit goed aan bij de gekozen server software en is gratis in gebruik wat het zeer geschikt maakt voor ons prototype.

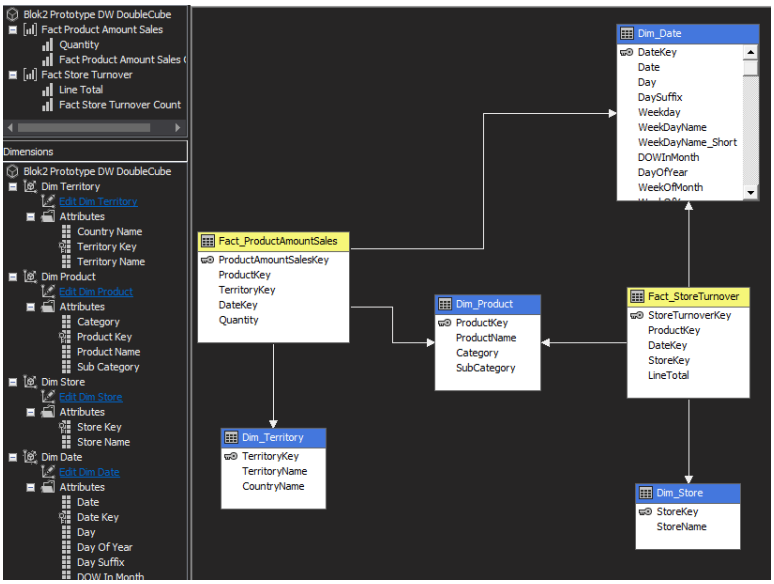
Tijdens het publiceren wordt connectie gemaakt met de data warehouse en worden kubussen gemaakt. De kubussen zijn uiteindelijk in te lezen in de visualisatie software.

In de data source view is de bron data te zien in de vorm van tabellen en relaties. De twee feiten tabellen zijn weergegeven en de bijbehorende dimensies inclusief de attributen die hierbij horen.



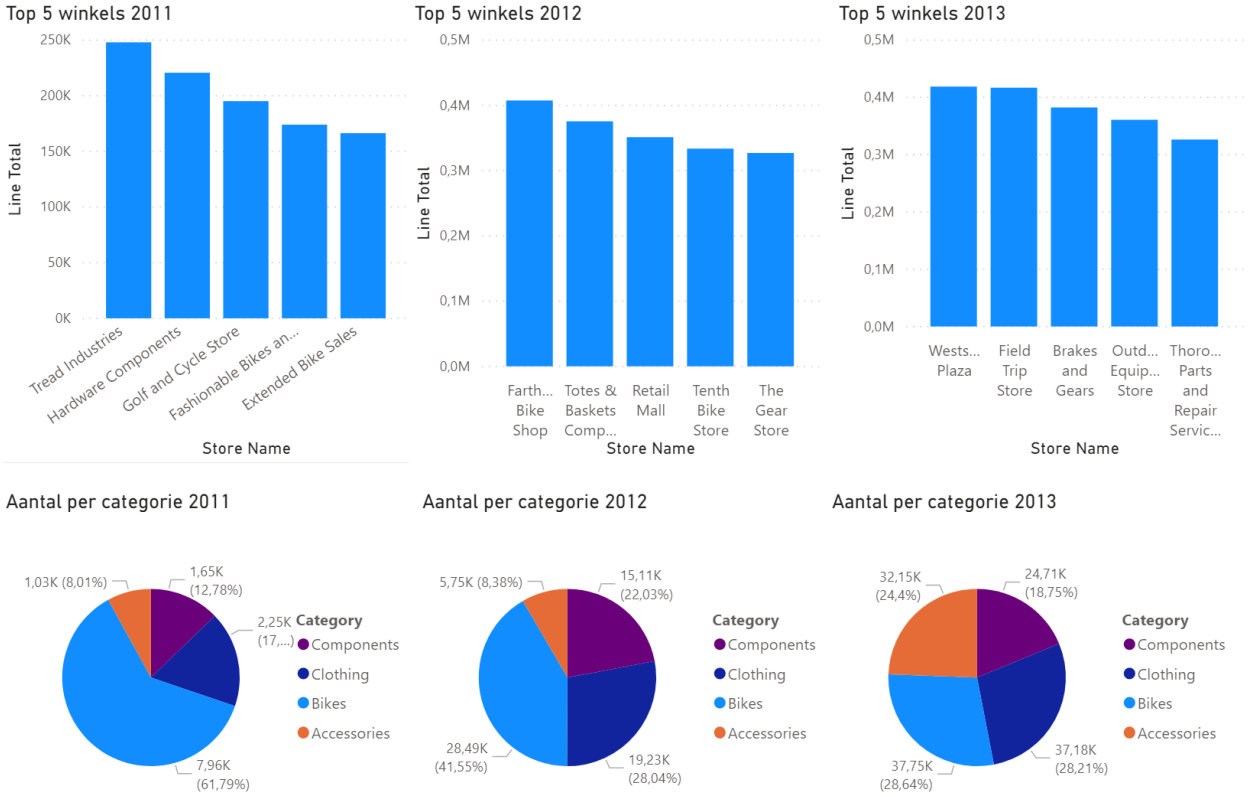
Voor het prototype zijn drie kubussen gemaakt: voor beide vragen een kubus en een gecombineerde kubus. Er is een gecombineerde kubus gemaakt omdat de visualisatie software een beperking heeft voor het verbinden met een enkele kubus. De kubussen worden op een soortgelijke manier weergegeven. Bij een kubus kan zelf worden gekozen welke attributen worden meegenomen. Voor het prototype maken we gebruik van alle attributen.


Blok2 Prototype DW ProductAmount
Fact Product Amount Sales
Quantity
Fact Product Amount Sales
Dimensions
Blok2 Prototype DW ProductAmount
Dim Territory
Edit Dim Territory
Attributes
Country Name
Territory Key
Territory Name
Dim Product
Edit Dim Product
Attributes
Category
Product Key
Product Name
Sub Category
Dim Date
Edit Dim Date
Attributes
Date
Date Key
Day
Day Of Year
Day Suffix
DOW In Month
Is Weekend
MMYYYY
Month
Month Name
Month Name Short
Month Year
Dim_Territory
TerritoryKey
TerritoryName
CountryName
Fact_ProductAmountSales
ProductAmountSalesKey
ProductKey
TerritoryKey
DateKey
Quantity
Dim_Product
ProductKey
ProductName
Category
SubCategory
Dim_Date
DateKey
Date
Day
DaySuffix
Weekday
WeekDayName
WeekDayName_Short
DOWInMonth
DayOfYear
WeekOfMonth

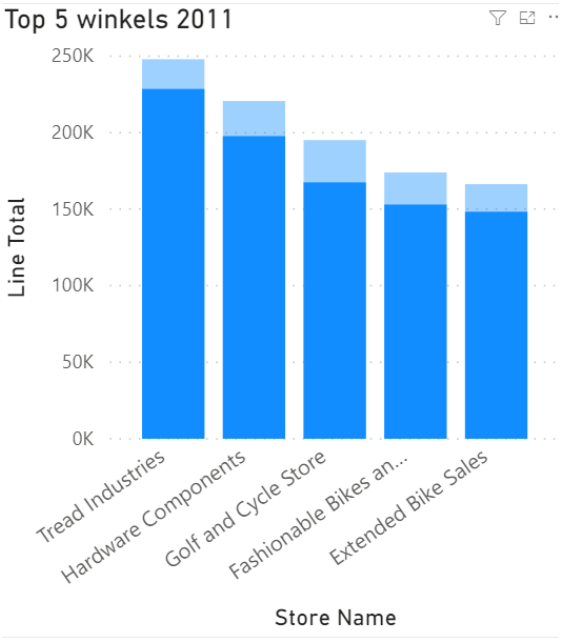

Blok2 Prototype DW DoubleCube
Fact Product Amount Sales
Quantity
Fact Product Amount Sales
Fact Store Turnover
Line Total
Fact Store Turnover Count
Dimensions
Blok2 Prototype DW DoubleCube
Dim Territory
Edit Dim Territory
Attributes
Country Name
Territory Key
Territory Name
Dim Product
Edit Dim Product
Attributes
Category
Product Key
Product Name
Sub Category
Dim Store
Edit Dim Store
Attributes
Store Key
Store Name
Dim Date
Edit Dim Date
Attributes
Date
Date Key
Day
Day Of Year
Day Suffix
Dim_Date
DateKey
Date
Day
DaySuffix
Weekday
WeekDayName
WeekDayName_Short
DOWInMonth
DayOfYear
WeekOfMonth
Fact_ProductAmountSales
ProductAmountSalesKey
ProductKey
TerritoryKey
DateKey
Quantity
Dim_Product
ProductKey
ProductName
Category
SubCategory
Fact_StoreTurnover
StoreTurnoverKey
ProductKey
DateKey
StoreKey
LineTotal
Dim_Territory
TerritoryKey
TerritoryName
CountryName
Dim_Store
StoreKey
StoreName

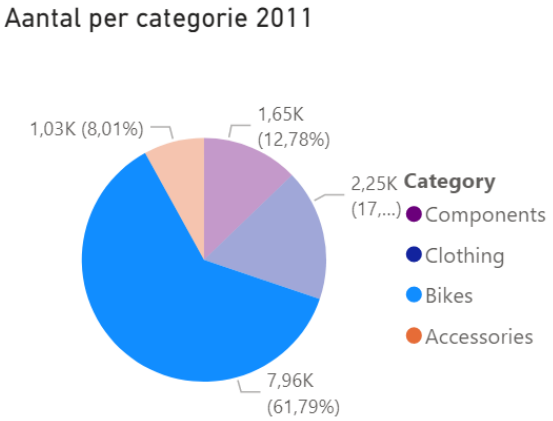
Visualisatie

De laatste stap in het proces is visualisatie. In deze stap komen de antwoorden op de business vragen tevoorschijn. Deze stap laat de kracht van business intelligence goed zien. Voor de visualisatie software is gekozen voor Power BI van Microsoft. Deze software sluit goed aan bij de andere gekozen softwarepakketten en is gratis te gebruiken. Dit maakt het zeer geschikt voor een prototype. Hieronder een voorbeeld van een dashboard gemaakt in PowerBI.

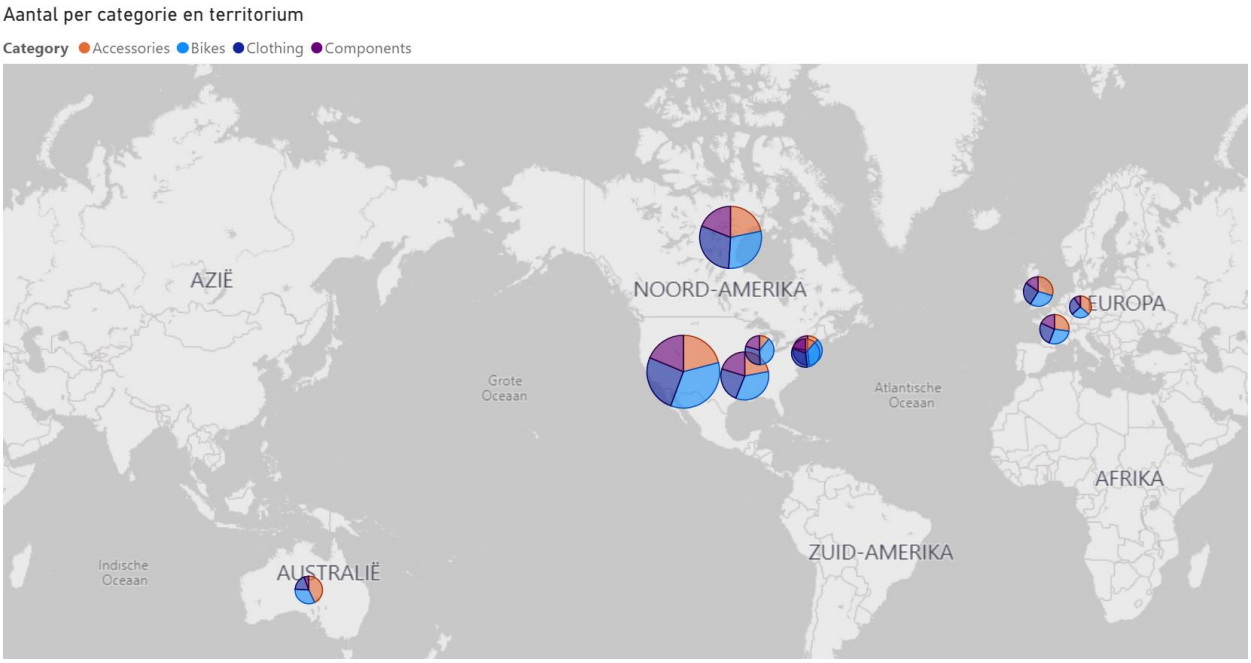


De dashboards zijn interactief. Door te klikken op de bikes categorie wordt boven weergegeven wat het deel aan verkochte fietsen is per winkel.





Er kunnen zeer veel verschillende dashboards worden gemaakt. Hieronder is een dashboard te zien met een landkaart welke aangeeft wat de totaal aantal verkochte producten zijn per regio.



We hebben in het prototype een aantal dashboard opgenomen die inzicht bieden in de mogelijkheden die de visualisatie software biedt.