

Migratieplan

Groep 3c

Teo Curcic

Joshua van Ee

Cas Ros

Stan Boeter

Inhoudsopgave

[Kwaliteit van de gegevens 3](#_Toc509826155)

[ETL Tools 4](#_Toc509826156)

[SQL Server Intergration Services 4](#_Toc509826157)

[CloverETL 5](#_Toc509826158)

[Apatar 6](#_Toc509826159)

[Mapping 7](#_Toc509826160)

[Order 7](#_Toc509826161)

[Training (cursus) 8](#_Toc509826162)

[Retailer\_Site 8](#_Toc509826163)

[Retailer 9](#_Toc509826164)

[Department & Branch 9](#_Toc509826165)

[Customer 10](#_Toc509826166)

[Bonus 10](#_Toc509826167)

[Product 11](#_Toc509826168)

[Order Item 12](#_Toc509826169)

[Product Type 12](#_Toc509826170)

[Product Line 13](#_Toc509826171)

[Employee 13](#_Toc509826172)

[Sales Target 14](#_Toc509826173)

[Position 14](#_Toc509826174)

[Aanpak 0](#_Toc509826175)

[Testen 0](#_Toc509826176)

# Kwaliteit van de gegevens

Als we kijken naar de verschillende databronnen van de twee organisaties dan zien we dat er veel gegevens erin verwerkt staan. Het is dus ook onmogelijk om te voorkomen dat er geen fouten in zitten. Om een succesvolle migratie uitvoeren, proberen we eerst alle fouten uit de oude databronnen te halen, zodat die niet terug te zien zijn in de nieuwe database van Outdoor Paradise.

Door de databronnen grondig te onderzoeken, zijn er verschillende fouten geconstateerd:

* Inconsistente naamgeving: De verschillende attributen in de databasen zijn zo geformuleerd dat ze niet logisch benoemd zijn en is het niet duidelijk wat de attributen betekenen.
* Veel dubbele waardes: In de databasen zitten zoveel waardes verwerkt dat er ook veel dubbele waarden in zitten. Voor het onderhoud van de databasen is dit zeer onoverzichtelijk.
* Foute gebruik van opslagtypen: De meeste attributen zijn als een varchar() opgeslagen, terwijl het alleen getallen zijn of een karakter is. Dat wil je niet in je database hebben omdat je dan onjuist gebruik maakt van de verschillende datatypen.
* Geen relatie tussen verschillende tabellen: Er zijn tabellen die los van elkaar staan en met geen enkele andere tabel een relatie hebben. Het is niet de bedoeling dat er tabellen los van elkaar komen te staan.
* Te uitgebreide tabellen: Een tabel met te veel attributen en te veel waardes is niet overzichtelijk om te onderhouden. De tabellen moeten zo compact mogelijk gemaakt worden en op een logische manier verdeelt zijn.
* Verouderde data: Er is veel data dat verouderd is en niet meer gebruikt gaat worden door de organisatie. Je wilt zoveel mogelijk up-to-date data in je databasen hebben zodat de database optimaal te gebruiken is.

Door alle fouten te onderzoeken, gaan we ervoor zorgen dat er consistente naamgevingen gebruikt gaat worden, dat alle dubbele waarden uit de database komt en dat alle tabellen juist samengevoegd wordt, zodat ze allemaal een relatie krijgen.

# ETL Tools

## SQL Server Intergration Services

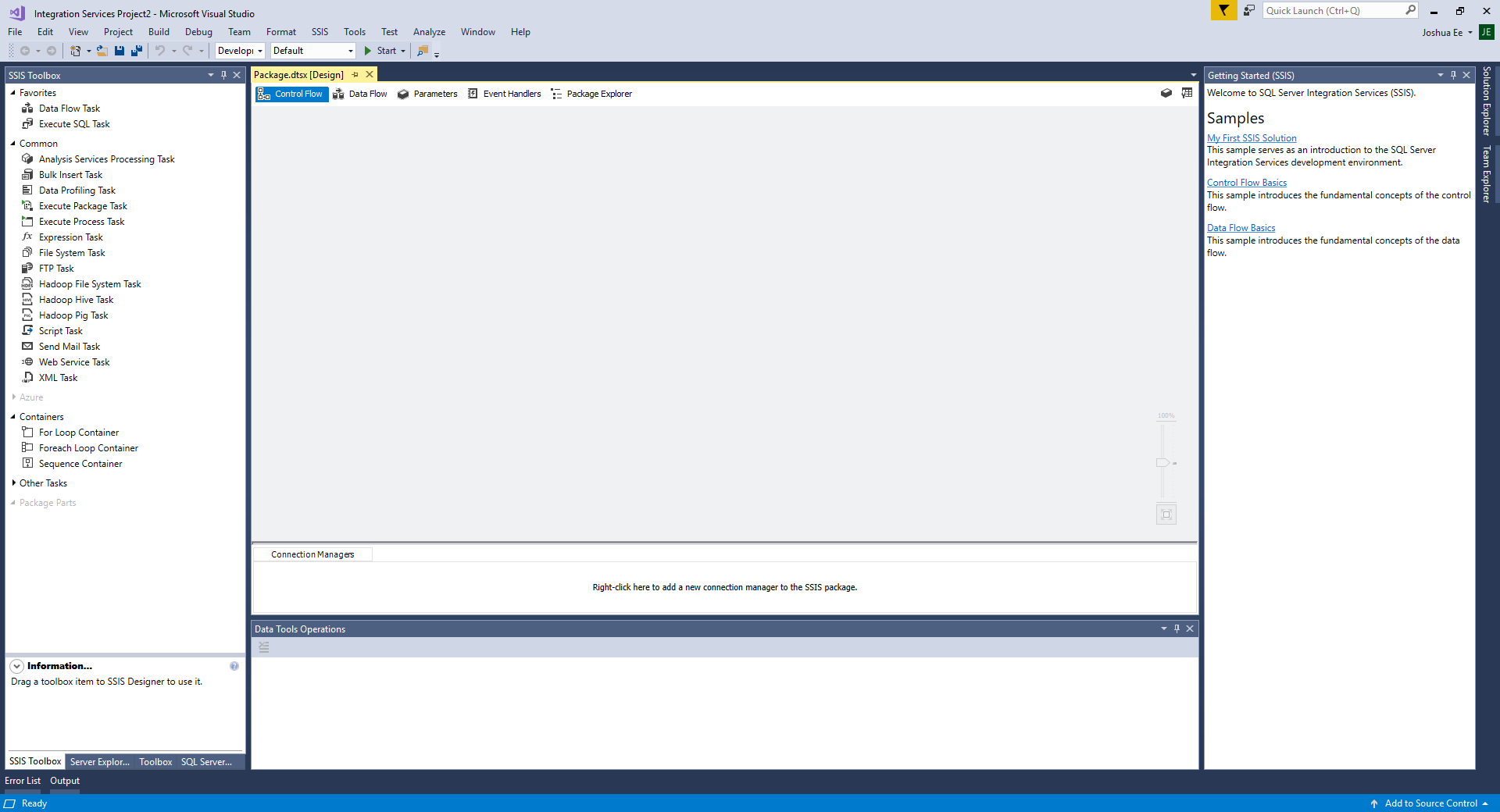
Homepagina: <https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services>

Waarom gekozen om te onderzoeken?

De tool is van Microsoft, en omdat wij al met visual studio werken vinden wij dit erg makkelijk.

Impressie:

SSIS is enorm goed gedocumenteerd. Microsoft zelf heeft een erg gedetailleerde documentatie pagina, de tool heeft zo veel gebruikers dat het internet vol staat met vragen en antwoorden. De tool is erg geavanceerd en heeft een redelijk bekend uiterlijk door Visual Studio. Ook wordt er op de website genoemd dat deze tool ondersteuning heeft voor het updaten van data warehouses, wat wij later nodig zullen hebben. Verder is er geen account nodig voor het gebruik van SSIS, wat ook een pluspunt is. SSIS wordt ook als de snelste commercieel beschikbare ETL tool gezien (Bron: <https://technet.microsoft.com/en-us/library/dd537533(v=sql.100).aspx>)



Conclusie:

SSIS is onze eerste keuze voor ETL tool die wij gaan gebruiken voor het project. Door de grote hoeveelheid documentatie, en het professionele uiterlijk lijkt het ons een fijn programma om mee te werken.

## CloverETL

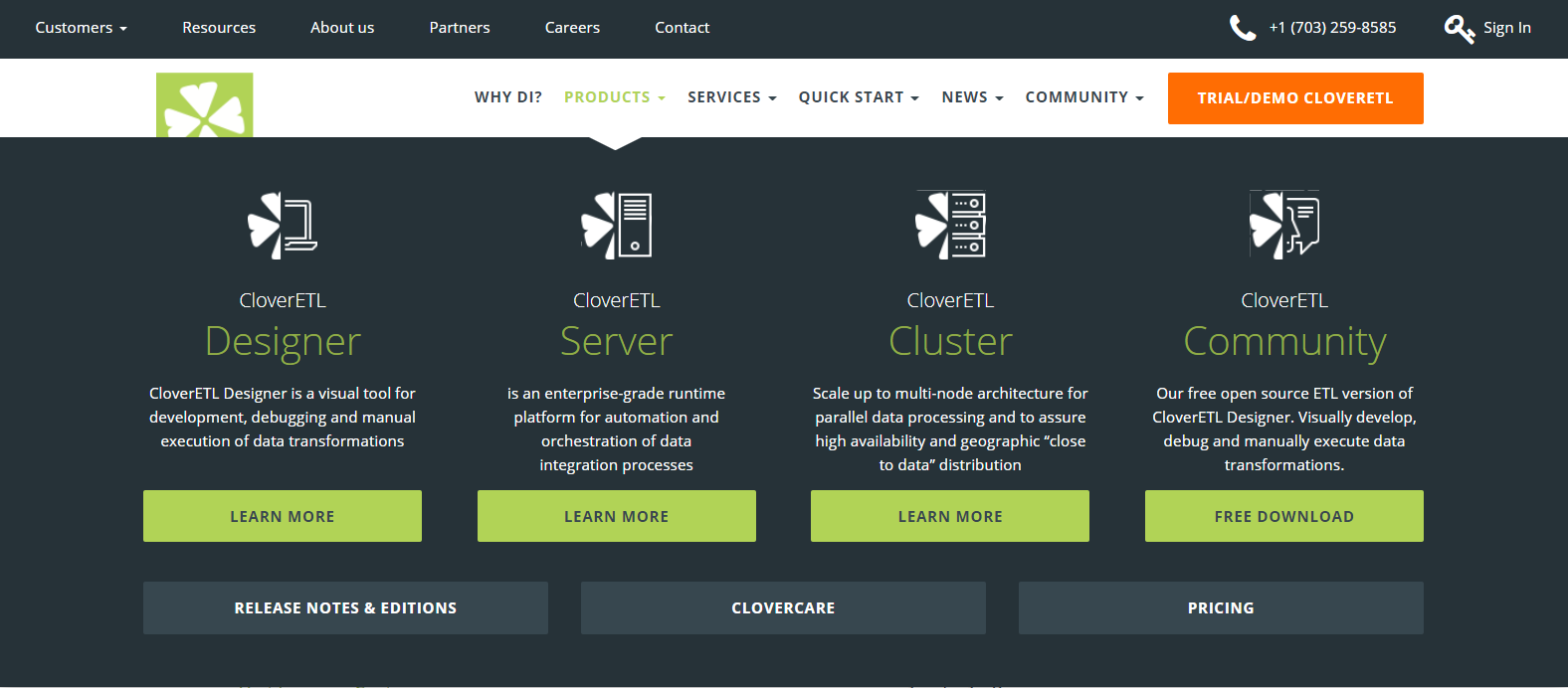
Homepagina: <https://www.cloveretl.com/>

Waarom gekozen om te onderzoeken?

CloverETL staat hoog beoordeeld als ETL tool. De software komt vaak voor in vergelijkingen met andere ETL tools en komt er zeker niet slecht uit.

Impressie:

De eerste impressie van de website is redelijk goed. De site ziet er voor het grootste deel professioneel uit. Alleen van de genoemde “leading companies” komen ons er maar twee bekend voor. Ook de footer ziet er niet goed uit, en de kleine lettertjes helemaal onder aan de pagina geven nog meer slechte impressies. De FAQ is enorm uitgebreid en ze hebben zelfs een erg gedetailleerd gebruikers forum. Ook biedt CloverETL een quickstart guide bestaande uit meerdere filmpjes. Het pakket zelf is enorm duur, 5000 euro per license, maar biedt wel een demo.



Conclusie:

De tool ziet er geweldig uit, maar wij kiezen toch liever voor SSIS omdat deze van microsoft is. Ook is de tool erg duur, al biedt het wel een demo, maar SSIS is gratis. Ook willen wij liever niet met de limitaties van de demo werken.

## Apatar

Homepagina: <http://apatar.com/>

Waarom gekozen om te onderzoeken?

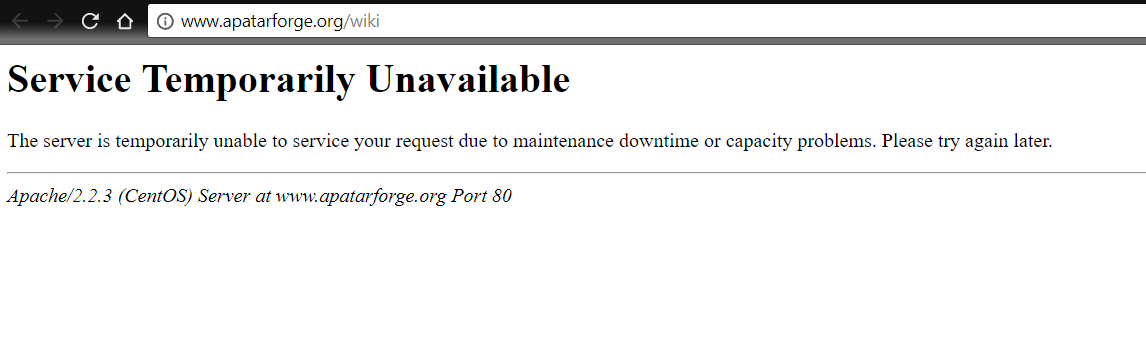
Deze ETL tool stond in de lijst van bekende en goede tools.

Impressie:

Toen wij de documentatie van deze ETL tool wilden doornemen, laadde de pagina niet. Dit is al een slechte eerste impressie. Verder bevat de FAQ alleen maar vragen die in het voordeel vallen voor Apatar. Het komt niet over alsof de vragen in deze FAQ ook echt gevraagd worden door gebruikers, onder door de vraag “Can Apatar server run in a multi-threaded fashion optimizing performance levels?”.

Ook de web demo op de site van Apatar werkt niet. Dit werkt zeker niet in het voordeel voor Apatar.

Voor het gebruik van de tool is een account nodig. Tegen de tijd dat de download klaar was, hadden wij al vele slechte impressies van Apatar. Omdat we liever geen account wouden aanmaken, maar alleen de software even wouden doorkijken, heeft dit de knoop doorgehakt voor het besluit deze software niet te gebruiken. Normaal is een account geen groot probleem, maar omdat de site al zo slecht was denken wij niet dat deze software het waard is.



Conclusie:

Dit wordt niet de ETL tool die wij gaan gebruiken voor ons project. De documentatie werkt niet, de FAQ geeft geen goede impressies, en zelfs de web demo werkt niet. Verder is er een account nodig voor gebruik en ook dit vinden wij niet erg fijn.

# Mapping

## Order



## Training (cursus)



## Retailer\_Site



## Retailer



## Department & Branch



## Customer



## Bonus



## Product



## Order Item



## Product Type



## Product Line



## Employee



## Sales Target



## Position



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Target Table | Target Column | Data-type | Source Table | Source Column | Data-type | Expression | Default Value if Null | Data Issues/Quality/Comments |
| Name of target table | Column in target table | Data type for this target column | Name of the Source table | Column in source table where you would get this data from | Data type for this source column | Decodes, aggregates, conversions, if statements, lookup functions…… | Value to use in target field when source field is null | Used to document Not null, value if looked up, upper case, pk, fk, etc…comments, issues |
| Order\_item | Order\_Item\_Id | INT |  |  |  |  | IDENTITY(1,1) | NOT NULL, PK |
| order\_id | INT | Order | Order\_id | INT |  |  | Not null, FK |
| product\_number | INT | Product | Product\_id | INT |  |  | NOT NULL, FK |
| quantity | SMALLINT | Order\_Details | Quantity | SMALLINT |  |  | NOT NULL |
| Sales\_Order\_Item | Quantity | INT |
| ship\_date | DATE | Sales\_Order\_Item | Ship\_date | DATETIME |  |  | NOT NULL |
| Price\_Each | FLOAT |  |  |  |  |  | NOT NULL |
| Product\_type | product\_type\_code | INT | Product\_Type | Product\_type\_code | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| product\_line\_code | INT | Product\_Type | Product\_line\_code | INT |  |  | FK |
| product\_type\_en | VARCHAR(80) | Product\_Type | Product\_type\_en | NVARCHAR(50) |  |  | NOT NULL |
| Product\_line | product\_line\_code | INT | Product\_line | Product\_line\_code | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| product\_line\_en | NVARCHAR(40) | Product\_line | Product\_line\_en | VARCHAR(80) |  |  | NOT NULL |
| Employee | emp\_id | INT | Sales\_Staff | Sales\_staff\_code | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| Employee | Emp\_id | INT |
| emp\_firstname | VARCHAR(20) | Sales\_Staff | First\_Name | NVARCHAR(25) |  |  | NOT NULL |
| Employee | Emp\_fname | VARCHAR(255) |
| emp\_lastname | VARCHAR(20) | Sales\_Staff | Last\_name | NVARCHAR(30) |  |  | NOT NULL |
| Employee | Emp\_lname | VARCHAR(255) |
| street | VARCHAR(100) | Employee | Street | VARCHAR(255) |  |  |  |
| city | VARCHAR(30) | Employee | City | VARCHAR(255) |  |  |  |
| state | VARCHAR(2) | Employee | State | VARCHAR(255) |  |  |  |
| zip\_code | VARCHAR(10) | Employee | Zip\_Code | VARCHAR(255) |  |  |  |
| phone | VARCHAR(20) | Employee | Phone | VARCHAR(255) |  |  |  |
| status | VARCHAR(2) | Employee | Status | VARCHAR(255) |  |  |  |
| Employee | ss\_number | VARCHAR(10) | Employee | Ss\_number | VARCHAR(255) |  |  |  |
| start\_date | DATETIME | Sales\_Staff | Date\_Hired | DATETIME |  |  | NOT NULL |
| termination\_date | DATETIME | Employee | Termination\_date | VARCHAR(255) | Convert to datetime |  |  |
| birth\_date | DATETIME | Employee | Birth\_date | DATETIME |  |  | NOT NULL |
| bene\_health\_ins | CHAR | Employee | Bene\_Health\_ins | CHAR |  |  |  |
| bene\_life\_ins | CHAR | Employee | Bene\_Life\_Ins | CHAR |  |  |  |
| bene\_day\_care | CHAR | Employee | Bene\_day\_Care | CHAR |  |  |  |
| sex | CHAR | Employee | Sex | CHAR |  |  |  |
| salary | FLOAT | Employee | Salary | DOUBLE |  |  |  |
| phone\_work | VARCHAR(20) | Sales\_Staff | Work\_phone | NVARCHAR(20) |  |  |  |
| extension | VARCHAR(6) | Sales\_Staff | Extension | NVARCHAR(20) |  |  |  |
| fax | VARCHAR(20) |  |  |  |  |  |  |
| email | VARCHAR(50) | Sales\_Staff | Email | NVARCHAR(50) |  |  |  |
| cv | VARCHAR(50) |  |  |  |  |  |  |
| manager\_id | INT | Sales\_Staff | Manager\_code | INT |  |  | FK |
| Employee | Manager\_id | INT |
| pos\_numb | INT |  |  |  |  |  | NOT NULL, FK |
| branch\_code | INT |  |  |  |  |  | FK |
| Sales\_target | emp\_id | INT | Sales\_Target | Sales\_Staff | INT |  |  | NOT NULL, PK |
| sales\_year | SMALLINT | Sales\_Target | Sales\_Year | INT |  |  | NOT NULL, PK |
| sales\_period | INT | Sales\_Target | Sales\_Period | INT |  |  | NOT NULL, PK |
| retailer\_name | VARCHAR(50) | Sales\_Target | Retailer\_Name | NVARCHAR(50) |  |  | NOT NULL, FK |
| product\_number | INT | Sales\_Target | Product | INT |  |  | NOT NULL, PK |
| sales\_target | FLOAT | Sales\_Target | Sales\_Target | FLOAT |  |  | NOT NULL |
| retailer\_code | INT | Sales\_Target | Retailer\_Code | INT |  |  | NOT NULL, PK |
| Postion | pos\_numb | INT | JOB\_NUMBER | Jobs | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| pos\_abbr | VARCHAR(50) | JOB\_ID | Jobs | VARCHAR(255) |  |  |  |
| pos\_title | VARCHAR(50) | JOB\_TITLE | Jobs | VARCHAR(255) |  |  |  |
| min\_salary | FLOAT | MIN\_SALARY | Jobs | INT |  |  |  |
| max\_salary | FLOAT | MAX\_SALARY | Jobs | INT |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Target Table | Target Column | Data-type | Source Table | Source Column | Data-type | Expression | Default Value if Null | Data Issues/Quality/Comments |
| Name of target table | Column in target table | Data type for this target column | Name of the Source table | Column in source table where you would get this data from | Data type for this source column | Decodes, aggregates, conversions, if statements, lookup functions…… | Value to use in target field when source field is null | Used to document Not null, value if looked up, upper case, pk, fk, etc…comments, issues |
| Order | Order\_Id | INT | Order\_Header; | Order\_Number | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| Sales\_order | id | INT |
| Order\_Date | DATETIME | Order\_Header | Order\_Date | DATETIME |  |  | NOT NULL |
| Sales\_order | Order\_date | DATETIME |
| Emp\_Id | INT | Order\_Header | Sales\_Staff\_Code | INT |  |  | NOT NULL, FK |
| Sales\_order | Sales\_rep | INT |
| Finance\_Code | VARCHAR(10) | Sales\_order | Fin\_code\_id | VARCHAR(255) |  |  | FK |
| Order\_Method\_Code | INT | Order\_Header | Oder\_Method\_Code | ÌNT |  |  | NOT NULL, FK |
| Retailer\_Site\_Code | INT | Order\_Header | Retailer\_Site\_Code | INT |  |  | FK |
| Cust\_Id\_Id | INT | Sales\_order | Cust\_id | INT |  |  | FK |
| Region | VARCHAR(50) | Sales\_order | Region | VARCHAR(255) |  |  |  |
| Status | VARCHAR(10) |  |  |  |  |  | NOT NULL |
| Training | Training\_Id | INTINT |  |  |  |  |  | NOT NULL, PK |
| Course | VARCHAR(30) | Go\_training | Course | NVARCHAR(255) |  |  | NOT NULL |
| Pos\_Numb | INT | Go\_training | Sales\_staff | NVARCHAR(255) |  |  | FK |
| Retailer\_site | Retailer\_Site\_Code | INT | Retailer\_Site | Retailer\_Site\_Code | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| Retailer\_Code | INT | Retailer\_Site | Retailer\_Code | INT |  |  | NOT NULL, FK |
| Address1 | VARCHAR(50) | Retailer\_Site | Address1 | NVARCHAR(50) |  |  | NOT NULL |
| Address2 | VARCHAR(50) | Retailer\_Site | Address2 | NVARCHAR(50) |  |  |  |
| City | VARCHAR(40) | Retailer\_Site | City | NVARCHAR(40) |  |  |  |
| Region | VARCHAR(50) | Retailer\_Site | Region | NVARCHAR(50) |  |  |  |
| Postal\_Zone | VARCHAR(10) | Retailer\_Site | Postal\_Zone | NVARCHAR(10) |  |  |  |
| Country\_Code | INT | Retailer\_Site | Counrty\_Code | INT | Not null |  | NOT NULL, FK |
| Active\_Indicator | BOOL | Retailer\_Site | Active\_Indicator | BIT(1) | Not null |  | NOT NULL |
| Retailer | Retaielr\_Code | INT | Retailer | Retailer\_Code | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| Company\_Name | VARCHAR(255) | Retailer | Company\_Name | NVARCHAR(255) |  |  | NOT NULL |
| Retailer\_Type\_Code | INT | Retailer | Retailer\_Type\_Code | INT |  |  | NOT NULL, FK |
| Department | Dept\_Id | INT | Department | Dept\_id | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| Dept\_Name | VARCHAR(255) | Department | Dept\_name | VARCHAR(255) |  |  | NOT NULL |
| Dept\_Head\_Id | INT | Department | Dept\_head\_id | INT |  |  | NOT NULL, FK |
| Branch | Branch\_Code | INT | Sales\_branch | Sales\_branch\_code | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| Adress1 | VARCHAR(50) | Sales\_branch | Address1 | NVARCHAR(50) |  |  | NOT NULL |
| Adress2 | VARCHAR(50) | Sales\_branch | Address2 | NVARCHAR(50) |  |  |  |
| City | VARCHAR(40) | Sales\_branch | City | NVARCHAR(40) |  |  |  |
| Region | VARCHAR(50) | Sales\_branch | Region | NVARCHAR(50) |  |  |  |
| Postal\_Zone | VARCHAR(10) | Sales\_branch | Postal\_zone | NVARCHAR(10) |  |  |  |
| Dept\_Id | INT | Department | Dept\_id | INT |  |  | NOT NULL, FK |
| Country\_Code | INT | Sales\_branch | Country\_code | INT |  |  | FK |
| Customer | Id | INT | Customer | Id | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| Fname | VARCHAR(255) | Customer | Fname | VARCHAR(255) |  |  |  |
| Lname | VARCHAR(255) | Customer | Lname | VARCHAR(255) |  |  |  |
| Adress | VARCHAR(255) | Customer | Adress | VARCHAR(255) |  |  |  |
| City | VARCHAR(40) | Customer | City | VARCHAR(255) |  |  |  |
| State | VARCHAR(255) | Customer | State | VARCHAR(255) |  |  |  |
| Zip | VARCHAR(40) | Customer | Zip | VARCHAR(255) |  |  |  |
| Phone | VARCHAR(20) | Customer | Phone | VARCHAR(255) |  |  |  |
| Company\_Name | VARCHAR(255) | Customer | Company\_Name | VARCHAR(255) |  |  |  |
| Type | CHAR | Customer | Type | CHAR |  |  |  |
| Discount | FLOAT | Customer | Discount | INT |  |  |  |
| Max\_Quantity\_Order | INT | Customer | Max\_quantity\_order | INT |  |  |  |
| Bonus | Bonus\_Id | INT |  |  |  |  | IDENTITY(1,1) | NOT NULL, PK |
| Bonus\_Date | DATE | Bonus | Bonus\_date | DATETIME |  |  |  |
| Bonus\_Amo | FLOAT | Bonus | Bomus\_amo | NUM |  |  |  |
| Emp\_Id | INT | Bonus | Emp\_id | NUM |  |  | FK |
| Product | Product\_Number | INT | Product (AenC) | id | INT |  | IDENTITY(1,1) | NOT NULL, PK |
| Product (GO) | Product\_number | INT |
| Introduction\_Date | DATETIME | Product (GO) | Introduction\_date | DATETIME |  |  | NOT NULL |
| Product\_Type\_Id | INT | Product (AenC) | category | VARCHAR(1) |  |  | NOT NULL, FK |
| Product (GO) | Product\_type\_code | INT |
| Production\_Cost | FLOAT | Product (GO) | Production\_cost | FLOAT |  |  | NOT NULL |
| Margin | FLOAT | Product (GO) | Margin | FLOAT |  |  | NOT NULL |
| Product\_Image | VARCHAR(255) | Product (AenC) | Picture\_name | VARCHAR(255) |  |  | NOT NULL |
| Product (GO) | Product\_image | VARCHAR(150) |
| Product\_Name | VARCHAR(255 | Product (AenC) | name | VARCHAR(255) |  |  |  |
| Product (GO) | Product\_name | VARCHAR(255) |
| Product\_Description | VARCHAR(255) | Product (AenC) | description | VARCHAR(255) |  |  |  |
| Product\_Size | VARCHAR(20) | Product (AenC) | Prod\_size | VARCHAR(255) |  |  |  |
| Product\_Color | VARCHAR(255) | Product (AenC) | color | VARCHAR(255) |  |  |  |

# Aanpak

Op basis van de datadictionary en het RIM is er een database voor Outdoor Paradise gebouwd. Om de data naar deze database te migreren wordt er gebruik gemaakt van SQL Server Intergration Services als ETL Tool. De mapping tabellen worden gebruikt tijdens de migratie als controlestap, zodat de juiste data naar de juiste tabel wordt gemigreerd. In Excel wordt er een tabel met alle tabelnamen van de database aangemaakt. Zodra de datamigratie van een tabel is afgerond, wordt deze gecheckt in het Excelbestand. Op deze manier worden er geen tabellen overgeslagen of dubbel gemigreerd. Na het afronden van een migratie worden de eerste twintig waardes uit de tabel vergeleken met de brongegevens om een correcte datamigratie te waarborgen.

# Testen

We kunnen de migratie op een aantal verschillende manieren testen. Bijvoorbeeld door:

* De eisen aan het nieuwe systeem nog een keer doornemen en testen op het systeem door de eisen in praktijk uit te voeren voor zover dit mogelijk is.
* Test data te gebruiken om alle relaties, constraints en triggers te controleren
* Domeincontrolles: Met domeincontrolles controleren we of de waardes die we migreren wel geldig zijn.
* Bedrijfsregelcontrolles: door te controleren of er geen bedrijfsregels overtreden zijn door de te migreren data, kunnen we bepalen of de kwaliteit van de data goed is.
* Door voor de migratie en na de migratie de som van bepaalde kolommen bij te houden kunnen we controleren of er geen dataverlies/veranderingen plaats gevonden hebben tijdens de migratie.

# Evaluatie

Na de migratie hebben we geëvalueerd over wat er goed gegaan is en wat we beter kunnen doen de volgende keer. Het belangrijkste is dat je zonder een goede voorbereiding nergens bent. Op het moment dat er een goed plan van aanpak is, de juiste afspraken zijn gemaakt en de mappings correct geformuleerd zijn, weet je bijna zeker dat de migratie succesvol plaats gaat vinden. Tijdens de migratie merkte we wel dat er sommige aspecten beter achteraf gedaan konden worden. Zoals de auto incrementeel attributen achteraf te implementeren in plaats van vooraf, want tijdens de migratie zorgde dat voor enkele problemen. Het is ook zeer van belang dat alle data van te voren goed bestudeerd wordt en gechecked is of de juiste data klaar staat om gemigreerd te worden. Als je achteraf er achter komt dat het niet de juiste data was dan moet je het helemaal proces opnieuw doen en dan verlies je veel kostbare tijd. We hebben veel geleerd van de gemaakte fouten.