# Conventiedocument Quadraam DWH

## Structuur van scripts

* In the visual studio solution staan twee scripts per tabel: een creatiescript om deze aan te maken, en een ETL-script om hem te vullen.
* Toelichting over de naamgeving van velden staat, waar nodig, in het creatiescript.
* Toelichting over de bron van tabellen en kolommen en eventuele filters, zuiveringen, en verrijkingen staat in het ETL-script. De standaardconventies in dit document worden niet nogmaals in de scripts benoemd.
* Waar een complex filter of verrijking op meerdere plaatsen gebruikt wordt, wordt deze in een functie gezet.
* Datamarts worden geïmplementeerd door views op de tabellen. Joins tussen tabellen worden geplaatst in deze views, niet in de eindrapporten.
* Om de structuur overzichtelijk te houden wordt zo veel mogelijk vermeden dat functies andere functies aanroepen, of dat ETL-scripts andere ETL-scripts aanroepen, of dat views op andere views gebaseerd worden.
* TODO-puntjes worden bijgehouden op het Trello bord, niet in de scripts.

## Naamgeving

* CamelCase, echter afkortingen wel in hoofdletters; voeg een hiervoor een ‘\_’ tussen, bijv: OP (onderwijs en personeel) OP\_vast, OP\_tijdelijk.
* De taal is bij voorkeur Nederlands.
* Namen zijn enkelvoudig.
* Het verschil tussen dimensie en fact tabellen wordt in de table schema gemaakt: dim.<tabelnaam> en fact.<tabelnaam>
* Gebruik geen afkortingen, behalve waar het veelgebruikte termen zijn. Uitgeschreven afkortingen kunnen weer toegevoegd worden op datamart, rapportageniveau.
* Geef eenheden aan; bijv. tijdsduur in uren of in minuten;
* Het achtervoegsel “ID” wordt gebruikt voor primary en foreign keys. IDs zijn altijd numeriek (dwz int/bigint/smallint/tinyint).
* Bij kolommen wordt de naam van de tabel toegevoegd waar het veelvoorkomende kolomnamen betreft, bijv “MedewerkerNaam” i.pv. “Naam”.

## Conventies voor kolomtypen

* De volgende conventies worden in principe op alle ETL scripts toegevoegd; uitzonderingen hierop worden in de scripts toegelicht.
* Datetime velden die puur over dagen gaan (bijv. geboortedatum) omzetten naar date.
* Int velden waar de inhoud dit toelaat omzetten naar smallint (waarden tot aan 32,767) of tinyint (255).
* Floats voor wetenschappelijke metingen en decimals worden overgenomen indien deze velden correct gebruikt worden.
* Binaire velden, bijv. een varchar die ‘Ja’ en ‘Nee’ bevat, omzetten naar bit.
* Varchar velden verkleinen tot het minimum nodig voor de data, met enige speling. Bijv een varchar(250) waarin de langste waarde 55 karakters is omzetten naar varchar(60).
* Varchar velden met een standaardlengte, bijvoorbeeld postcode, omzetten naar char.
* Nvarchar velden waar de inhoud dit toelaat omzetten naar varchar. De unicodering wordt dus verwijderd indien mogelijk.
* Voor kolommen van type (n)(var)char waarop in de ETL scripts JOINs plaatsvinden worden spaties aan het begin of eind verwijderd d.m.v. LTRIM en/of RTRIM.
* Met uitzondering van metingen, optelbare velden (bijv. aantal besteedde minuten) en datumvelden zijn de kolommen niet nullable, en null-waarden worden m.b.v. Coalesce vervangen door ‘’ of 0. Er kunnen een aantal velden zijn waarbij het relevant is of de waarde bewust niet is ingevuld, dan gebruiken we toch de NULL.
* Voor dimtabellen wordt de business key gebruikt zolang deze numeriek en uniek is. Waar dit niet het geval is wordt een surrogate key gebruikt, en is de business key te vinden in een veld met voorvoegsel “BK\_”.

## Conventies voor filtering

* Kolommen die in de bronsystemen overal leeg zijn, of overal dezelfde waarde bevatten als een andere kolom, worden niet meegenomen.
* Kolommen voor creatie- of wijzigingsdatum van die regel in het bronsysteem, worden op dit moment niet meegenomen. In een later stadium kan besloten worden hiervoor temporal tables te gaan gebruiken, dit is bijvoorbeeld het geval bij begin- en einddatum van de geldigheid van die regel.
* Op dit moment worden tabellen gevuld door ze eerst te truncaten, en daarna opnieuw te initialiseren vanuit de brondata. Waar er temporal tables gebruikt gaan worden zal dit anders gedaan worden, en voor de fact-tabellen zullen we in een later stadium zorgen dat alleen de nieuwe regels worden toegevoegd en de oude waarden blijven staan.

## Standaardfunctionaliteit

* De scripts beginnen standaard met Begin Try / Begin Transaction, en eindigen met Log / Commit / Catch / Log / Rollback. Er is een generiek logscript PROCEDURE [ETL].[Log] en deze schrijft in de tabel dbo.Logging. Dit wordt gecombineert met TRY CATCH constructie voor foutafhandeling.
* Bovenin de dimensietabellen wordt een regel toegevoegd met nulwaarden, en met het naamveld “Onbekend”. Deze heeft primary key 0 (als de PK van type tinyint is) of -1 anders. Deze regel is niet inferred.
* Parametrisatie, zoals “[$(SERVERNAAM)].[$(DATABASENAAM)]” is specifiek voor Visual Studio project om variabel de databases en servers aan te kunnen passen. Als je deze code binnen SQL wilt draaien moet je de $() even verwijderen.
* In Visual Studio staat code analysis standaard aan voor elke build, met uitzondering van de regels SR0007 en SR0014 (omdat deze false positives opleveren).

## Standaardfunctionaliteit later eventueel toevoegen

* De RecentList constructie wordt gebruikt om de meest recente gegevens op te vragen, bijvoorbeeld de meest recent toegevoegde geboortedatum van een patient. Dit wordt gedaan door te sorteren op patientnummer en wijzigingsdatum, deze per patient te nummeren, en vervolgens alleen de waarden te gebruiken waar RowNum = 1.
* Voor een aantal dimensies wordt een InferralList opgesteld; deze bestaat uit alle waarden uit een kolom uit de facttabel (bijvoorbeeld alle artsen genoemd in de DBC tabel). Deze wordt gebruikt om de dimensietabel aan te vullen waar de brondata niet alle gebruikte waarden blijkt te bevatten. Een nieuwe kolom, “Inferred” van type bit, is 1 op deze toegevoegde regels, anders 0. Je gebruikt dit als oplossing om onzuiverheden in de (bron)data via het DWH terug te koppelen. Mogelijke oplossingen zijn:
  + Zo mogelijk tikfouten oplossen in de bron
  + Ontbrekende gegevens toevoegen in bron (hulptabellen)
  + Indien bovenstaande stappen niet lukken, oplossen in ETL script