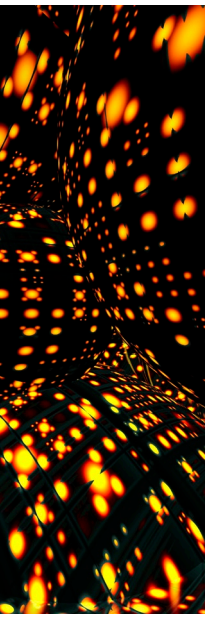


Implementatie datakwaliteit met Six Sigma

Van zandbak tot pretpark

Sandra Wennemers en Ruud Kuil

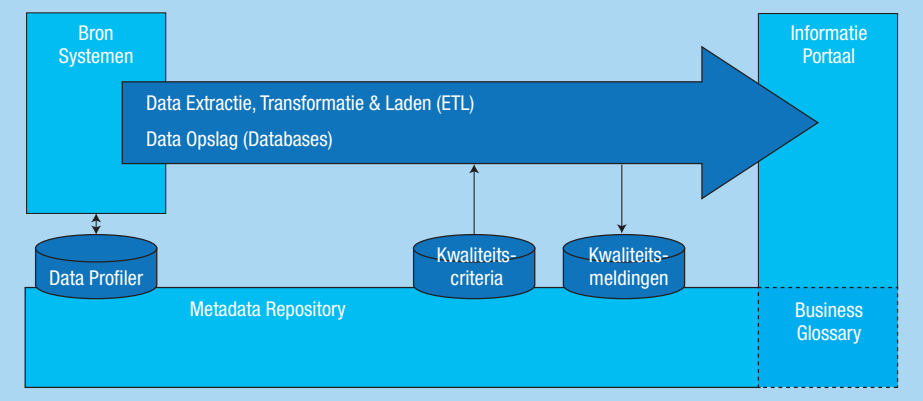
Het belang van datakwaliteit wordt door iedereen onderkend. In de praktijk zien we ook steeds vaker initiatieven tot het verbeteren en beheersen van datakwaliteit, die vanuit technisch oogpunt sterk van elkaar verschillen. In sommige gevallen wordt slechts wat geëxperimenteerd in een zogenaamde zandbak, waar andere bedrijven een volledig pretpark van datakwaliteit tooling implementeren.

In dit artikel wordt een overzicht gegeven van technische hulpmiddelen op het gebied van datakwaliteit. Deze informatie kan helpen bij het starten van uw eigen initiatief op het gebied van datakwaliteit of om een huidig initiatief uit te breiden.
Bedrijven hebben vaak weinig zicht op de mate van datakwaliteit totdat ze ermee geconfronteerd worden, bijvoorbeeld tijdens een datamigratie. Het inrichten van datakwaliteitsmetingen als een continu proces is nog vrij nieuw, maar is noodzakelijk in omgevingen waarin data steeds meer gebruikt worden voor centrale rapportage en/of uitwisseling over afdelings- en soms zelfs bedrijfsgrenzen heen. Vanuit dit oogpunt biedt Six Sigma een goede kapstok voor een datakwaliteitproces.
Binnen Six Sigma zijn enkele stappen gedefinieerd die samen een cyclisch proces vormen (zie afbeelding 1). Het proces wordt veelvuldig afgekort met DMAIC, dat is afgeleid van de eerste letters van de zich herhalende stappen waaruit dit continue proces bestaat.
Elk van deze stappen in een datakwaliteitproces wordt in dit artikel toegelicht. Daarnaast volgt een overzicht van de technische hulpmiddelen die momenteel op de markt zijn om deze stappen te ondersteunen. Tot slot wordt een voorbeeld architectuur van een mogelijke implementatie beschreven.

Het draait allemaal om definities

Het definiëren van het probleem en het inrichten en afstemmen van de projectorganisatie is de eerste stap binnen de Six Sigma methodiek. Voordat je een probleem kunt oplossen, moet je een duidelijk doel definiëren. Het doel in dit artikel is het beheersen van datakwaliteit. Een goede definitie is meetbaar en om dit te bereiken moeten zowel data als kwaliteit nader worden gespecificeerd. Deze stap is overigens niet uit voeren zonder de betrokken stakeholders (data consumers, data owners enzovoort).[1]
Voor wat betreft het definiëren van data beschouwen wij data alleen in de gestructureerde variant, welke in databases of datawarehouses zijn ondergebracht. Tijdens de definitiefase wordt bepaald met welke dataobjecten we aan de slag zullen gaan. Naast een naam voor ieder dataobject is een eenduidige definitie onontbeerlijk. Denk aan beschrijvingen, oorsprong, formules, waardebereik enzovoort. Als voorbeeld in deze paragraaf gaan wij uit van het dataobject *Geboortedatum, de datum volgens de Juliaanse kalender waarop een natuurlijk persoon ter wereld kwam*.
Ook de kwaliteitscriteria dienen dusdanig te worden omschreven dat deze meetbaar gemaakt kunnen worden. Als we kijken naar de verschillende gradaties (*Technisch, Functioneel en Business Rule*[1]) zijn de *technische* kwaliteitscriteria het makkelijkst te kwantificeren. Voorbeelden van technische criteria zijn bijvoorbeeld een verplichte veldcontrole, domeincontrole en uniciteitcontrole.
De gedefinieerde kwaliteitscriteria dienen gekoppeld te worden aan de gedefinieerde objecten. Het is echter niet zo dat alle criteria gekoppeld hoeven te worden. Sommige dataobjecten kunnen het gebruik van bepaalde criteria per definitie uitsluiten. Zo zien we over het algemeen dat de uniciteitcontrole niet van toepassing is op datumvelden. Of een *geboortedatum* strikt noodzakelijk is, wordt bepaald door de *Data Consumer*.
De *Functionele kwaliteitscriteria* en *Business Rules* zijn veelal attribuutgebonden, bijvoorbeeld onze geboortedatum die altijd in het verleden dient te liggen. Over het algemeen zijn de criteria in deze categorieën regelmatig aan veranderingen onderhevig.
Kwaliteit is voor ons de mate waarin de (eind)gebruikers van de data deze kunnen gebruiken voor hun bedrijfsdoeleinden. Dat betekent dus dat zij de norm moeten zetten: Wat is goed? Waarbij wij er rekening mee moeten houden dat de norm per stakeholder kan verschillen.
In de praktijk zien we meestal dat deze fase ondersteund wordt door producten uit de Office collectie. Voor het verzamelen, opslaan en publiceren van datadefinities en kwaliteitscriteria zijn

Afbeelding 1: Het Six Sigma Proces.



Afbeelding 2: Technische inrichting van een datakwaliteitpretpark.

echter meer tools beschikbaar op de markt. Tabel 1 geeft een overzicht van de verschillende categorieën.

Meten met een duimstok of schuifmaat?

De tweede stap binnen de Six Sigma kwaliteitsmanagement-methode betreft het kwantificeren van de doelstelling en het uitvoeren van metingen. Om de vastgelegde definities en kwaliteitscriteria om te zetten naar tastbare metingen en resultaten zijn binnen Six Sigma enkele rekenmethoden veelgebruikt.
De Six Sigma rekenmethoden zijn gebaseerd op een tweetal invoerwaardes:
- Het aantal Opportunities (uitgevoerde metingen);
- Het aantal Defects (aantal gevonden afwijkingen).

Over het algemeen geldt dat iedere regel in een bestand of tabel meerdere dataobjecten bevat en er op één object meerdere kwaliteitscriteria van toepassing kunnen zijn.
Six Sigma gebruikt de volgende statistieken ter ondersteuning van de methodiek:
"Defects Per Million Oppertunities" oftewel DPMO:
*DPMO = (Aantal Defects/Aantal Opportunities) * 1.000.000.*
Defect percentage:
*Defect % = (Aantal Defects/Aantal Opportunities) * 100.*
Yield (aantal goed):
Yield % = 100 - (Aantal Defects/Aantal Opportunities).
De Sigma Score kan worden berekend aan de hand van de volgende formule: *Zscore = Inverse std. Normaalverdeling van (1 - (Aantal Defects/Aantal Opportunities)).*

De resultaten van de Sigma Score worden weergeven in een schaalverdeling van 0 tot 7 waarbij een 6 de 'perfecte' score is binnen de Six Sigma methodiek. Deze score staat voor 3,4 fouten op 1.000.000 metingen. Dit maakt deze meetmethode tot een gevoelige indicator van kleine wijzigingen die in een percentage niet altijd tot uiting komen. In de praktijk is het niet per definitie deze perfecte score die we nastreven, maar de norm zoals gezet door de stakeholders. De verhouding tussen de waardes toont aan dat met het hanteren van de Six Sigma scores een schuifmaat kan worden gehanteerd. Gebruik van enkel percentages is dan vergelijkbaar met een duimstok, zie tabel 2.
Voor het uitvoeren, opslaan en rapporteren van datakwaliteitmetingen zijn verschillende tools beschikbaar op de markt. Het is echter in sommige gevallen ook goed te doen met behulp van database query's en Excel.
De eerste eenvoudigste stap zou zijn om een query over de Persoon tabel te draaien om erachter te komen of iedere registratie een *Geboortedatum* heeft.
Tweede stap is om iedere geregistreerde *Geboortedatum* te controleren op validiteit. Een voorbeeld van een incorrecte datum is 29 februari 2011.
De derde stap in dit proces kan de vergelijking tussen de *Geboortedatum* en *Invoerdatum* of de *huidige datum* zijn.
In de laatste stap worden Opportunities en Defects in Excel overgenomen en gepresenteerd.
Bovenstaand voorbeeld is een situatie die prima door generieke

Categorie	Kenmerken
Business Glossary	HTML interface voor het publiceren van datadefinities voor business users en administratieve functionaliteit voor het onderhouden van definities.
Metadata repository	Opslag van definities, gericht op koppelingen naar modellen, systemen enzovoort. Voorziet in traceability mogelijkheden, interfaces naar repository modules van andere software.

Tabel 1.

Percent Defective	Meaning	Sigma Level	DPMO
69%	The values can be trusted 3 out of 10 times.	1	691,462
31%	The values can be trusted 7 out of 10 times.	2	308,538
6.7%	The values can be trusted 8 out of 10 times.	3	66,807
0.62%	The values can be trusted 9 out of 10 times.	4	6,210
0.023%	The values can be trusted almost 10 out of 10 times.	5	233
0.00034%	The values can be trusted nearly 10 out of 10 times.	6	3.4
0.0000019%	Achieving this level becomes prohibitively expensive.	7	0.019

Tabel 2.

Categorie	Kenmerken
Custom	Met behulp van database query's of open source filecheckers en eventueel Excel voor de presentatie.
Data Profiling	Gericht op het exploreren van data in een bestand en/of database. Inventariseert waarden bereik, maakt melding van uitzonderingen en gaat uit van een standaard set van controles. Wordt vaak aangeboden als module in een ETL-suite. Data Profiling tools zijn eventueel ook bruikbaar tijdens de definitiefase om vanuit technisch perspectief inzicht te krijgen in de meer technische datadefinities.
Data Quality Generiek	Bieden een groot scala aan standaard controles. Worden vaak aangeboden als module in een ETL-suite veelal ontstaan door overnames.
Data Quality Specialistisch	Bieden vaak zeer specifieke functionaliteit zoals bijvoorbeeld fonetische controles voor het ontdubbelen van een klantenbestand.
Rule Engines	Focus op flexibel implementeren van de controle van Business Rules en aandacht voor Business Rules administratie.

Tabel 3.

Categorie	Kenmerken
Custom Reporting	Een interface en in beperkte mate ook reporting- en analysefunctionaliteiten.
Informatie Portaal	HTML interface die toegang geeft tot de overige componenten binnen de implementatie.
Reporting	Variërend van ad hoc reportingfunctionaliteit tot mobile dashboards.
Analyse	Mogelijkheden om cijfers op verschillende detailniveaus en vanuit verschillende invalshoeken te bekijken.
Statistiek	Voor een geavanceerde statistische aanpak.
Datamining	Het zoeken naar verbanden en patronen in de voorgevallen defect.

Tabel 4.

Data Quality tools kan worden ondersteund. Er zijn echter ook situaties waarbij een specialistisch tool nodig is. Een voorbeeld van een toepassing hiervan is: *Het controleren of een geregistreerde Persoon niet dubbel is geregistreerd door fonetische (mis)interpretatie ("Jansens" en "Janssens")*.
De kwaliteitscriteria hebben dus een grote invloed op de keuze voor tools uit één categorie of misschien meerdere, zie tabel 3.

Analyseren van symptomen

Na het meten is het van belang om de resultaten in het juiste perspectief te plaatsen en niet direct aan correcties te gaan werken. Indien de resultaten van de metingen niet goed worden geïnterpreteerd wordt snel een situatie bereikt van '*dweilen met de kraan open*'. De angel zit in het feit dat de oorzaak van een datakwaliteitsprobleem niet wordt weggenomen, waardoor problemen zich alleen maar zullen ophopen. De spreekwoordelijke kraan zal alleen maar harder gaan stromen. Op een bepaald moment zal de dweil verzadigd raken en wordt het stadium '*pompen of verzuipen*' bereikt.
De norm zal vaak op geaggregeerde uitkomsten betrekking hebben en de afwijking van de norm wordt bij voorkeur visueel gepresenteerd. In geval van overschrijding van de norm zal nadere analyse op detailniveau wenselijk zijn. Eventueel nog aangevuld met additionele statistische analyses.
Op zoek naar de oorzaak achter de problemen kunnen technische hulpmiddelen enorm helpen. Een voorbeeld hiervan is het gebruik van geautomatiseerde dashboards. Met de ondersteuning van de juiste grafieken wordt de interpretatie van de resultaten de goede kant opgestuurd.
Bij het opbouwen van een dashboard is het aan te raden om te beginnen bij de attributen welke worden gecontroleerd. Per attribuut worden alle gedefinieerde controles getoond om de bevindingen te kunnen interpreteren met alle relevante informatie voorhanden. Aggregeer de resultaten van hieruit naar de entiteiten en eventueel systemen of informatiedomeinen.
De eerder genoemde meetinstrumenten hebben ingebouwde, gestandaardiseerde rapportagetools die onder de categorie Custom reporting zijn geadresseerd. Daarnaast kan voor het analyseren van de meetgegevens gebruik gemaakt worden van de technologieën in tabel 4.

Verbeteren door wegnemen van de oorzaak

De vierde stap binnen Six Sigma betreft het ontwerpen en selecteren van oplossing(en) die oorzaken wegnemen die een grote invloed op de kwaliteit hebben. Op een bepaald moment zal de stap gemaakt moeten worden van het analyseren van de symptomen tot het bepalen van de feitelijke oorzaak. Dit kan alleen worden bereikt indien de verantwoordelijke stakeholders voor het systeem en eventueel informatiedomein meewerken. Gebeurt dit niet dan zal de oorzaak van de problemen bijna niet bepaald kunnen worden. Op het moment dat de oorzaak van de datakwaliteitsproblemen is vastgesteld kan worden begonnen met het definiëren van een oplossingsrichting. In veel gevallen zal een aanpassing in programmatuur noodzakelijk zijn. Daarnaast is niet alles technisch af te vangen en zal het wegnemen van de oorzaak in veel gevallen leiden tot het aanpassen van processen en de menselijke factor. Denk aan onduidelijke invulinstructies of te beperkte tijd die voor een handeling beschikbaar is.
Dit zal niet bij ieder gevonden probleem het geval zijn. Het is namelijk niet ondenkbaar dat een update query op regelmatige basis een goedkopere en snellere oplossing is dan het aanpassen van de software. In een aantal situaties kunnen eenmalige schoning- en/of verrijkingacties uitkomst bieden. Dit valt gedeeltelijk in de categorie symptoombestrijding. Denk hierbij aan de verrijking van persoonsregistraties voor een specifieke marketingactie. Buiten het feit dat het blijvend oplossen van datakwaliteitsproblemen vaak ingrijpt op processen en menselijk handelen, is er wel een aantal tools op de markt waarmee datakwaliteitbevindingen kunnen worden aangepakt, zie tabel 5. Six Sigma spreekt overigens zelf over process management tools als ondersteunend voor deze fase.

Beheren van het continue proces

Het implementeren van de maatregelen ter borging van ons doel gebeurt door de scope te handhaven dan wel uit te breiden. Dit laatste kan zowel door het aantal controles op gemeten

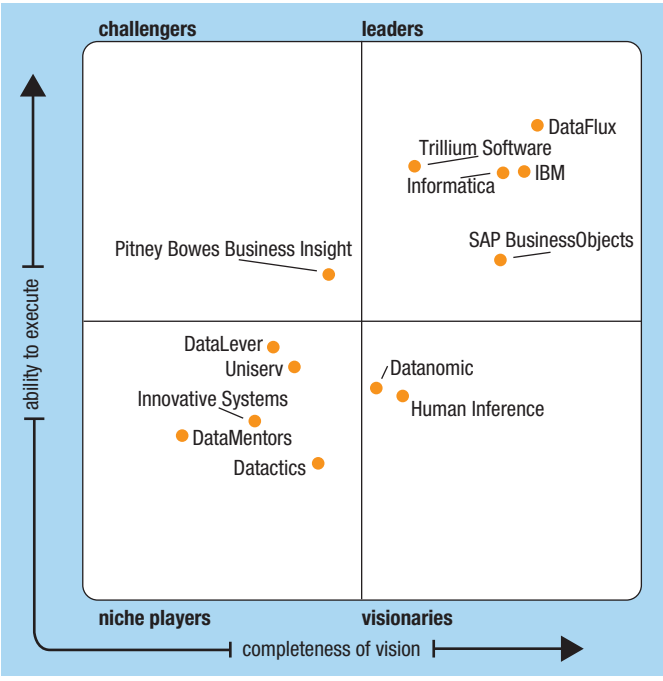
Categorie	Kenmerken
OLTP Aanpassingen	Bijvoorbeeld het toevoegen van picklists en reminders in administratieve systemen. Ook het draaien van een update of misschien het verhelpen van een fout in de programmatuur.
Data Quality Specifiek	Een aantal van de tools uit deze categorie biedt ook functionaliteit voor het opschonen, aanvullen of wijzigen van de aanwezige informatie.

Tabel 5.

data uit te breiden dan wel door meer data te gaan controleren. Omdat data een veranderend karakter hebben is het noodzakelijk om een proces in te richten om regelmatig de kwaliteit te meten en waar nodig te verbeteren. Begin klein en eenvoudig en bouw de omvang en complexiteit langzaam uit totdat alle definities en kwaliteitscriteria die van toepassing zijn aan bod zijn gekomen. Het is raadzaam de dataobjecten met de grootste impact op de business als eerst op het gewenste niveau te brengen en houden. Om uw datakwaliteitprocessen te kunnen beheersen raden wij aan om zaken als versiebeheer en scheduling gedegen in te richten. Deze ondersteunende processen kunnen worden gefaciliteerd met diverse tooling waar wij in dit artikel niet verder op in zullen gaan.

Zandbak of pretpark

Wij adviseren om datakwaliteitinitiatieven eerst klein op te zetten in een 'zandbak' en deze vervolgens beheerst te laten groeien. Deze groei kan eventueel worden ondersteund doordat bestaande datakwaliteitproblemen op waarde kunnen worden geschat. Gebruik de zandbak om onderbuikgevoelens te bevestigen of ontkrachten. Probeer regelmatig even stil te staan en om je heen te kijken of je zandbak nog wel een zandbak is. Het risico van te lang werken in de zandbak schuilt in het feit dat je heel snel het overzicht kwijt kunt raken.



Afbeelding 3: Kwadrant datakwaliteitsoftware. Bron: Gartner, juni 2010.

Houd rekening met het feit dat ondanks dat de ontwikkel- en beheerkosten (licenties, trainingen) in eerste instantie beperkt zijn, er onder water mogelijkerwijs een hoop oncontroleerbare en onbeheersbare kosten zijn. Zodra het proces een bepaald niveau van volwassenheid bereikt, is het zaak om langzaam maar zeker een beweging richting een pretpark in te gaan zetten. Dit kan bereikt worden door uitbreiding aan te schaffen op in de organisatie reeds beschikbare suites of door aanschaf van componenten van specialistische leveranciers. In afbeelding 2 is weergegeven hoe de verschillende softwarecategorieën met elkaar gebruikt worden voor een volledige pretparkinrichting.

Overigens moet worden benadrukt dat de volwassenheid tussen de verschillende toolcategorieën sterk verschilt. Dit geldt met name voor tools binnen de categorieën Metadata Repository en Business Rules Engines. De tools binnen deze categorieën verschillen sterk in functionaliteit. Afbeelding 3 geeft een aantal voorbeelden van zowel generieke als specialistische Data Quality tools, gerangschikt volgens Gartner. *Wel willen wij u er op wijzen dat Six Sigma calculaties en statistieken niet per definitie onderdeel van de tooling zijn, maar kunnen wel worden toegevoegd.*

Conclusie

Met het Six Sigma proces als geleide heeft dit artikel hopelijk een beeld kunnen schetsen van de kernaspecten van een datakwaliteittraject. Je kunt niet vroeg genoeg beginnen met het meten van een kleine hoeveelheid data. Maak je niet druk wanneer je nog niet alle criteria voor deze data in beeld hebt. Er zullen als het goed is voldoende iteraties komen waarin de scope kan worden uitgebreid. Vaak smaken de eerste defects en reparaties al snel naar meer. Blijf hierbij wel continu waken voor symptoombestrijding.

Wat de technische implementatie betreft is er een verscheidenheid aan technologieën in verschillende fasen van volwassenheid op de markt. Ook hier adviseren we klein te beginnen en niet meteen het pretpark in te duiken. Stap echter wel op tijd uit als je de zandbak ontgroeid bent, voordat de problemen je boven het hoofd groeien.

Tot slot willen wij nogmaals aangeven dat het verleidelijk is om problemen vanuit IT snel op te lossen. Een groot percentage van datakwaliteitbeheersing zit echter in het beschikken over de juiste mensen en processen. Dit geldt zowel voor de uitvoerders en stakeholders binnen de Six Sigma cycli als binnen de administratieve organisatie. Wanneer je het echt hebt getroffen zelfs van de organisatie in haar geheel.

Noot

1. Altijd Datakwaliteit, Database Magazine juni 2010, Ruud Kuil & Sinbad Moors.

Sandra Wennemers en Ruud Kuil

A.P. Wennemers is Principal Consultant en R.C. Kuil is Senior Consultant, beiden zijn werkzaam bij Capgemini op het gebied van Datamanagement.