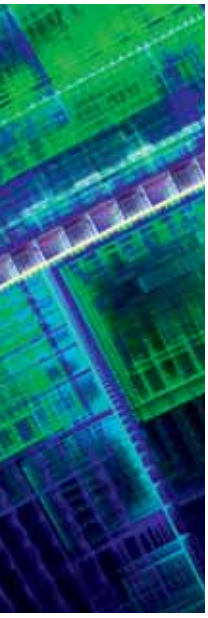


Continu proces van meten en verbeteren is noodzakelijk

# Altijd Datakwaliteit

Ruud Kuil en Sinbad Moors

**Het belang van datakwaliteit wordt breed onderkend, maar in de praktijk is vaak sprake van gesegmenteerde, vrijblijvende initiatieven. In dit artikel wordt een aantal handvatten geboden om een proces in te richten voor een (semi)continue meting van de datakwaliteit.**

In dit artikel wordt u aan de hand van één van onze geïmplementeerde datakwaliteitoplossingen (zie afbeelding 1) langs de belangrijkste onderdelen van een gedegen datakwaliteitoplossing geloodst. De eerste stap in de bouw van een proces voor datakwaliteitsmeting is het in kaart brengen van de bestaande stromen tussen data-elementen. In dit artikel wordt gesproken over data, in plaats van gegevens, vanwege de meer technisch georiënteerde opzet van het uiteindelijke proces.
De blauwdruk van datastromen vormt de basis voor het bepalen van de juiste plek (en het juiste moment) voor de continue meting. Redeneer bij de opbouw van deze plattegrond vanuit het eindpunt. Dit kan een enterprise datawarehouse (EDW) zijn, maar ook een geïmplementeerde masterdata management oplossing. Een kwaliteitsmeting kan ook een eenmalig karakter hebben, of slechts beperkt herhaald worden. Dit geldt bijvoorbeeld voor een datamigratie of periodieke rapportages en/of analyses. Vanuit het eindpunt wordt iedere inkomende stroom getraceerd terug naar de bron. Het is zaak om hierin grondig te werk te gaan. Ook ieder 'tussenstation' dient geregistreerd te worden, het kan tenslotte zomaar zijn dat op één van de 'tussenstations' de data één of meer transformaties ondergaan.

In het gunstigste geval zijn uitgebreide beschrijvingen van de metadata beschikbaar waarin is vastgelegd welke data, wanneer, worden getransporteerd en welke transformaties worden uitgevoerd. Dit bespaart veel tijd bij het definiëren van de juiste controles. In de gevallen waar deze informatie ontbreekt, is dit een uitgelezen moment om een volledige *up to date* blauwdruk te creëren. Wanneer de blauwdruk afgerond is, kan een keuze gemaakt worden:

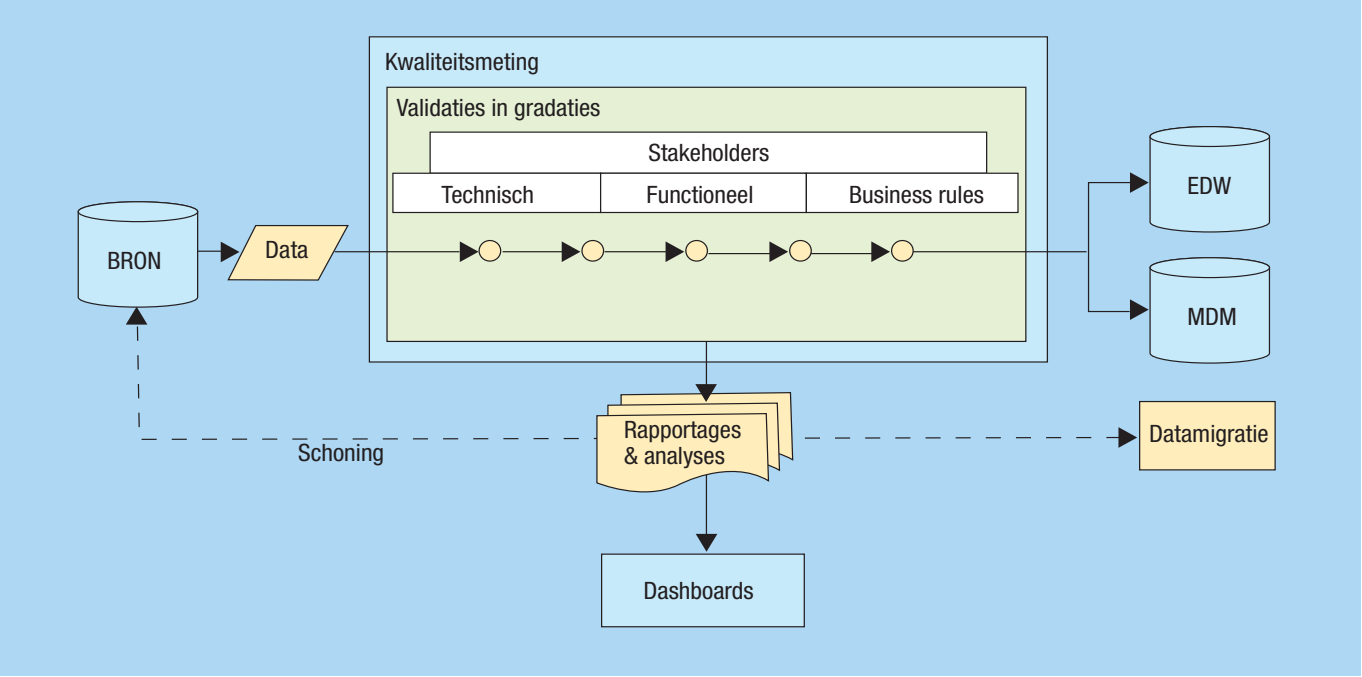
- of de gehele keten wordt meegenomen in de datakwaliteitsmetingen;
- of op basis van concrete kennis en ervaring wordt een aantal bekende knelpunten geselecteerd. Indien voor de laatste optie wordt gekozen, is het van belang het volgende in acht te nemen: hoe verder de metingen van de bron worden uitgevoerd, des te lastiger wordt het om de echte oorzaak van de problemen vast te stellen.

## Sleutelrol voor de stakeholders

Om de juiste controles en het eventuele proces van vervolgacties te kunnen definiëren, is het van belang om de juiste stakeholders erbij te betrekken. In veel gevallen zijn drie archetypen stakeholder te herkennen. Om te beginnen is de eigenaar van de data (vaak gedelegeerd aan een Data Steward) een belangrijke speler. De eindgebruiker (Data Consumer) van de data is een belangrijke tweede stakeholder.
Een stakeholder wiens rol niet altijd als even relevant wordt

## Six Sigma

Six Sigma is een wereldwijd veel gebruikte procesverbeteringsmethodiek. Bij Six Sigma staat het meetbaar maken van de bedrijfsprocessen en het herhaaldelijk meten over tijd centraal.
De resultaten van de metingen 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.
Waarschijnlijk is het bij het gebruik van de Six Sigma methodiek binnen datakwaliteitsmetingen overigens beter om te spreken van 'acceptabel' bij een score 6 dan van 'perfect'. Waarbij het laatste woord uiteraard is aan de betrokken stakeholders.




**Afbeelding 1**: Proces datakwaliteitsmetingen.

gezien, is de producent (Data Producer) van de data. De producent van de data is feitelijk de partij die aan het begin van de keten de data invoert of aanlevert. Indien de data transformaties ondergaan in een keten, neemt het belang van de producent af. De stakeholders of de door hen aangewezen personen kunnen de juiste input leveren voor het definiëren van de uit te voeren validaties.
*In dit artikel wordt de term validatieregel gehanteerd voor de functionele definitie. De controleregel beschrijft de implementatievorm van de validatieregel.*

Het is belangrijk om het onderscheid te onderkennen tussen goede datakwaliteit en acceptabele datakwaliteit. De stakeholders moeten aangeven welk niveau van datakwaliteit acceptabel is. Let hierbij wel op dat de partijen verschillende belangen kunnen hebben. In sommige gevallen wordt een bepaald niveau van datakwaliteit voorgeschreven uit regelgeving (compliance).
De eindgebruiker (EDW) zal data van goede kwaliteit willen ontvangen. Voor de eindgebruiker geldt dan dat bijvoorbeeld een gemiddelde score van 90 procent goed is. De eigenaar van de data heeft andere belangen en kan bijvoorbeeld aangeven dat een gemiddelde score van 80 procent goed genoeg is.
Voor alle partijen kan een ander acceptabel niveau van datakwaliteit bestaan. Zet de partijen daarom bij elkaar en laat gezamenlijk besluiten welk niveau van datakwaliteit als acceptabel beschouwd wordt. Deze gezamenlijke aanpak vergroot ook de betrokkenheid bij de vervolgstappen.

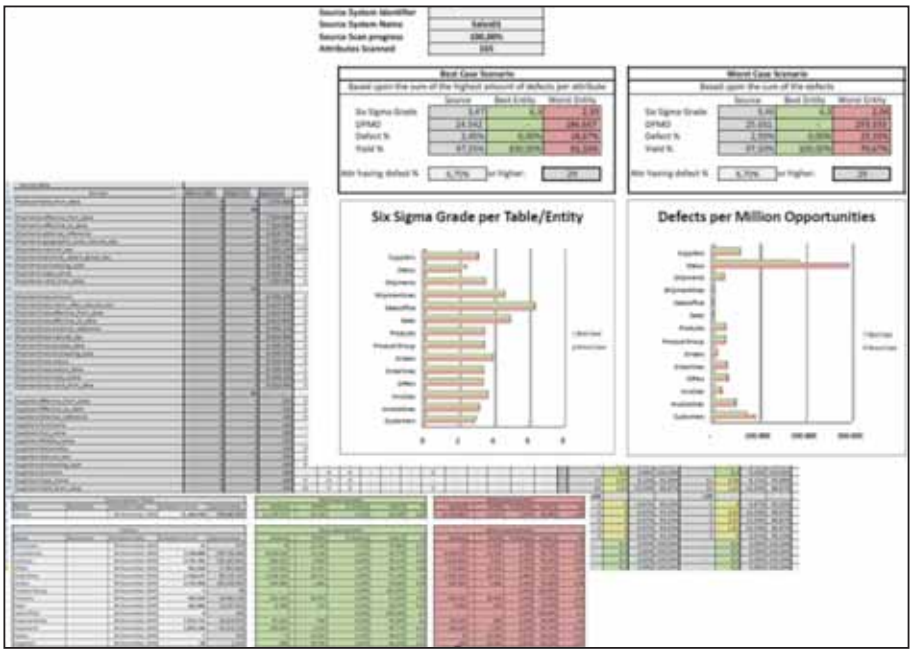
## Gradaties van datakwaliteit

Om het verdere proces te vergemakkelijken is het raadzaam om samen met de stakeholder(s) gradaties van datakwaliteit te definiëren. Een gradatie is feitelijk een verzameling van kwaliteitscriteria. In de beschreven oplossing is gekozen voor een drietal gradaties:
1. Technische gradatie;
2. Functionele gradatie;
3. Business rule gradatie.

In de technische gradatie worden validaties opgenomen die betrekking hebben op één specifiek veld in een database (één attribuut binnen één entiteit).
Onder deze gradatie vallen bijvoorbeeld validaties op:
- vulling van verplichte velden;
- correcte vulling van datatype (Numeriek, datum, boolean, enzovoort);
- vulling conform veldlengte (risico op truncations);
- aantal decimalen in numerieke velden.

De functionele gradatie bevat validaties waarbij de relatie van één veld met één of meer andere velden, al dan niet in dezelfde entiteit, wordt gevalideerd.
Onder deze gradatie vallen met name de volgende validaties:
- uniekheid van een sleutel;
- referentiële integriteit.

De derde gradatie, de business rule gradatie, heeft betrekking op validaties op voorwaardelijke samenhang tussen meerdere velden over meerdere entiteiten. Hieronder vallen ook vergelijkingen op geaggregeerde selecties. Een kredietlimiet op een bepaald type klant welke dient te worden nageleefd is een voorbeeld van een dergelijke regel.
Omdat deze regels betrekking hebben op de regels die in de praktijk gelden voor de data bij de bedrijfsvoering wordt deze gradatie als business rule aangeduid.




**Afbeelding 2**: Detail gegevens momentopname datakwaliteitsmeting.

Het is raadzaam om in deze fase ook gekwantificeerde wenselijkheidgrenzen (goed versus acceptabel) aan de gradaties toe te kennen. Dit kan op meerdere manieren gebeuren en aangezien deze gradaties in een later stadium technisch geïmplementeerd gaan worden, is het van belang de wenselijkheidgrenzen weloverwogen te formuleren. Bijvoorbeeld:

1. Het gewogen gemiddelde van alle correcte voorkomens van de gedefinieerde business rules moet minimaal 80 procent bedragen;
2. Geen van de gedefinieerde business rules mag een correctheidspercentage hebben kleiner dan 80 procent.

Het verschil tussen de beide regels is dat in het eerste voorbeeld theoretisch individuele business rules kunnen voorkomen waarbij het percentage vele malen lager is dan de gestelde 80 procent, terwijl het overall percentage boven de gestelde grens van 80 procent uitkomt. Dit heeft dan tot gevolg dat fouten die misschien veel tijd en geld kosten niet worden opgemerkt en opgelost. Dit probleem speelt in de tweede variant niet.
Het voordeel van het definiëren van gradaties komt sterk naar voren in de volgende fase van het proces, het vaststellen van de feitelijke validatieregels.

## Vaststellen van de validatieregels

Op basis van de genoemde gradaties kunnen de validatieregels worden gedefinieerd. Het voordeel van het werken met de drie gradaties ligt in het feit dat per gradatie alleen die partijen betrokken hoeven te zijn die werkelijk iets kunnen bijdragen.
Bij het definiëren van de technische validaties is een deskundige op het gebied van de business bijvoorbeeld niet vereist.
Bij het opstellen van de validatieregels wordt doorgaans de 'GOED' situatie omschreven in logische bewoordingen.
Een voorbeeld van een validatieregel is:

*Een registratie van een persoon moet voorzien zijn van een geldige geboortedatum.*

Afhankelijk van het datamodel is hier uiteraard sprake van een technische of functionele gradatie van de validatie. Indien gedocumenteerde metadata beschikbaar zijn, kunnen op basis hiervan de technische en functionele validatieregels eenvoudig worden gevuld. Indien documentatie niet direct voorhanden is, moeten de validatieregels zorgvuldig worden opgesteld door hiertoe aangewezen partijen. Om de business rules te definiëren is input nodig van proces- en beleidsdeskundigen binnen het domein waarvoor de data zijn opgeslagen.

## Meten is weten

Aanbeland op dit punt is een overzicht beschikbaar van de te gebruiken validatieregels. Het is nu zaak om de logisch omschreven validatieregels om te zetten naar meetbare controleregels. Om performance van metingen te verbeteren is het raadzaam om de controleregels te definiëren naar de 'FOUT'-situatie. Het is tenslotte waarschijnlijker dat het minder tijd en inspanning kost om de regels op te halen waarbij aan een bepaald criterium niet wordt voldaan dan andersom.
Het eerder gegeven voorbeeld levert dan de volgende controleregel op:
*Geef alle rijen uit de entiteit Persoon waarvan het attribuut Geboortedatum een ongeldige datum of geen waarde bevat.*

De tweede stap is het bepalen van op welk punt in de keten de controle moet worden uitgevoerd. Het verdient de aanbeveling om de controles van een technische gradatie zo vroeg mogelijk in de keten uit te voeren. Indien de keten in kwestie onderdeel uitmaakt van een ETL-proces naar een EDW kan het overigens goed zijn dat gebruik gemaakt wordt van delta's voor de aanle-

vering. Dit kan tot gevolg hebben dat veel functionele en business rule controles pas na de Load in bijvoorbeeld het EDW kunnen worden uitgevoerd.
Dan is het de vraag hoe de metingen worden verricht. Hiervoor is een aantal standaard oplossingen beschikbaar op de markt. Met een goede programmeur is deze investering echter niet altijd noodzakelijk. Na het uitvoeren van de controles worden de resultaten verwerkt in dashboards. Aangezien het aantal controles (en momenten) snel kan oplopen, is het belangrijk om handmatige acties tot een minimum te beperken.

## Presentatie in dashboards

Op basis van de implementatie van de controleregels kan vervolgens gekeken worden naar de manier waarop het niveau van datakwaliteit wordt gepresenteerd. Omdat de belangen verschillend zijn per stakeholder is het aannemelijk dat één dashboard meerdere lagen bevat. Een aantal voorbeelden van rapportage zoals door Capgemini met succes in de praktijk toegepast is te zien in afbeelding 2.
Hoe de belangen van stakeholders met betrekking tot de presentatie van de datakwaliteit liggen, hangt geheel van de positie in de keten af. Iemand die aan het einde van de keten zit en op basis van de data beslissingen neemt, haalt hoogstwaarschijnlijk uit een grafiek of staafdiagram voldoende informatie om beslissingen gefundeerd te nemen.
Een stakeholder die data moet schonen heeft daar echter niets aan. Deze stakeholder heeft juist meer belang bij gedetailleerde overzichten van de aanwezige fouten. Concrete voorbeelden van records waar de fouten zijn opgetreden kunnen hierbij bijvoorbeeld behulpzaam zijn.
Neem als uitgangspunt voor de opbouw van de dashboards de eerder benoemde gradaties. Probeer in de rapportages de wenselijkheidgrenzen op een natuurlijke manier te integreren. Om tot een correct geaggregeerd niveau van weergave op gradatieniveau te komen is een aantal belangrijke stappen vereist. Zo is het van belang om het totaal aan uitgevoerde controles in beschouwing te nemen in de resultaatcalculaties en niet het aantal aangeleverde records.

Om de technische gradatie als voorbeeld te nemen:
In een datastroom zijn twee bestanden opgenomen. Bestand 1 bevat 100 records, Bestand 2 bevat 500 records.
Per bestand zijn twee attributen opgenomen welke beide te allen tijde gevuld dienen te zijn en een lengte mogen hebben van maximaal twintig karakters. De score wordt weergegeven in de vorm van een Defects per Million Opportunities (DPMO), formule hiervoor is:
((aantal geconstateerde fouten) / (aantal uitgevoerde controles)) x 1.000.000.

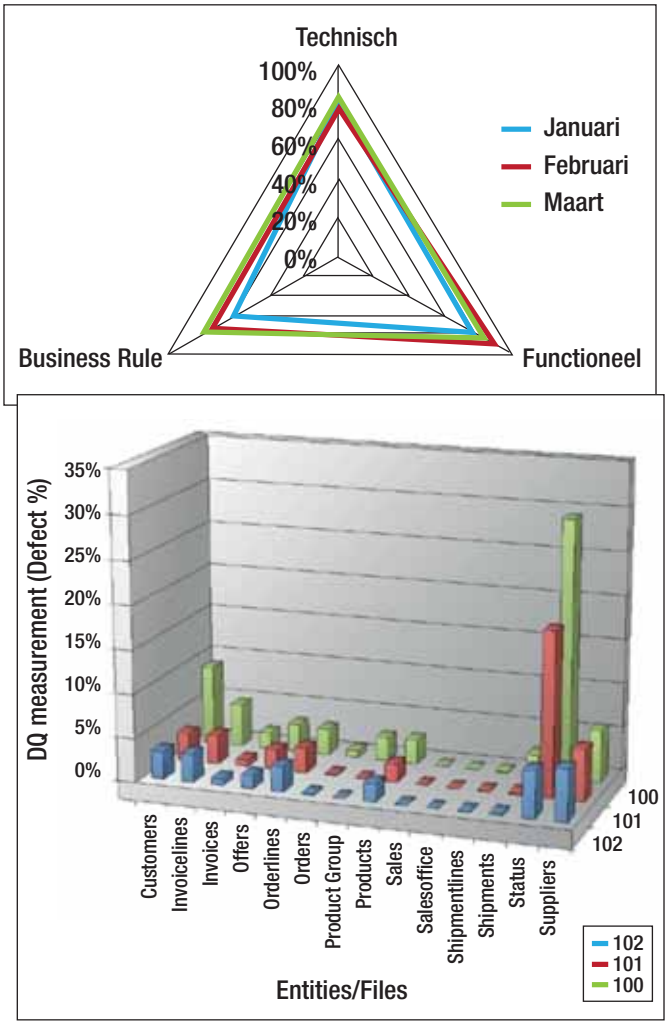
Omdat deze formule een calculatie op het aantal geconstateerde fouten is zal in de regel gelden dat; hoe lager de score, hoe beter de kwaliteit van de data. Op beide attributen worden twee controles uitgevoerd: controle op ingevulde waardes en controle op veldlengte.
Tijdens de controle komt naar voren dat:
Bestand 1.Attribuut A 25 records heeft welke leeg zijn en 30 waarvan de waarde maximale lengte overschrijdt;
Bestand 1.Attribuut B in 10 gevallen leeg is;
Bestand 2.Attribuut A in 100 gevallen leeg is;
Bestand 2.Attribuut B in 400 gevallen de lengte overschrijdt.

*1.A heeft een score van ((25 + 30) / 200) x 1.000.000 = 275.000.*
*1.B heeft een score van (10 / 200) x 1.000.000 = 50.000.*
*De score voor bestand 1 is dan ((25 + 30 + 10) / 400) x 1.000.000 = 162.500.*

*2.A heeft een score van (100 / 1000) x 1.000.000 = 100.000.*
*2.B heeft een score van (400 / 1.000) x 1.000.000 = 400.000.*
*De score voor bestand 2 is dan ((100 + 400) / 2000) x 1.000.000 = 250.000.*

## Dimensies

Een andere indeling dan de aangegeven gradaties is het indelen naar dimensies van datakwaliteit. Bij deze vorm van meting moeten de stakeholders aangeven welke dimensies van de data als belangrijkst worden beschouwd. Enkele voorbeelden van bruikbare dimensies zijn: Compleetheid, Accuratesse, Tijdigheid en Toepasbaarheid.
Bij datakwaliteitsmetingen op dimensies spelen twee uitdagingen. De eerste uitdaging ligt in de precieze definitie van een dimensie ter voorkoming van tegenstrijdigheden en/of overlap tussen dimensies. Men zou bijvoorbeeld kunnen stellen dat minimaal tien jaar historie in de data moet voorkomen alvorens de data Compleet genoeg zijn voor gebruik. Het aantonen van een dergelijke historie kan mogelijk eenvoudig indien er gebruik is gemaakt van timestamps, maar dan geldt vervolgens de vraag of de data van tien jaar geleden even Toepasbaar zijn als die van een jaar geleden.
De tweede uitdaging van deze methode is het meetbaar krijgen van een gekozen dimensie. Hoe kan worden aangetoond dat data bijvoorbeeld Compleet en Toepasbaar genoeg zijn voor gebruik?
Belangrijk bij het gebruik van dimensies is dat men klein begint. Iedere stakeholder benoemt een dimensie die van belang is vanuit zijn/haar perceptie. Vervolgens is het van belang dat alle partijen bij elkaar komen om te bepalen welke validaties moeten worden uitgevoerd om aan te kunnen tonen dat de benoemde dimensies voldoende meetbaar zijn.
Om de compleetheid van data te kunnen aantonen kan bijvoorbeeld initieel worden gesteld dat minimaal tien jaar historie aan orders wordt aangeleverd. Om de toepasbaarheid van de data te kunnen waarborgen kan bijvoorbeeld worden gesteld dat iedere order in de aangeleverde dataset een relatie heeft met minimaal één order-regel. Het is vervolgens zaak om het aantal dimensies stap voor stap uit te breiden totdat een complete set aan dimensies is gedefinieerd.




**Afbeelding 3**: Grafische weergaves resultaten van datakwaliteitsmetingen over tijd.

*Over de gehele set is een score van (565 / 2400) x 1.000.000 = 235.416,7.*

Om op geaggregeerde niveaus correcte resultaten van metingen te krijgen is het van belang dat altijd een gewogen gemiddelde wordt genomen.
Het is van belang om een goede balans te vinden tussen de verschillende niveaus, misschien is het zelfs nodig om meerdere dashboards te ontwikkelen. Niet alleen de visuele weergave is van belang, ook de gebruikte statistische vorm is van belang. Als bijvoorbeeld percentages worden gebruikt, hoeveel decimalen worden dan weergegeven?

Als de data een dusdanige omvang krijgen dat een relatief grote hoeveelheid fouten toch een gerapporteerde 100,00 procent oplevert, is het verstandig om naar een andere vorm te zoeken.
In de dashboards in afbeelding 2 is gebruik gemaakt van enkele ondersteunende statistieken. Naast percentages is ook gebruik gemaakt van een aantal Six Sigma statistieken, bestaande uit een Sigma Score (zie kader) en het aantal fouten per één miljoen kansen. Deze combinatie van statistieken geeft voldoende diepgang aan de rapportage om een adequate inschatting te maken van de datakwaliteit.

## Gemeten en dan?

Wanneer alle controles zijn uitgevoerd en de dashboards zijn opgemaakt moeten de resultaten worden vertaald naar daadwerkelijke acties. Er zijn situaties waarin schoning geen vereiste is, het registreren en documenteren is dan voldoende. Als data gebruikt worden voor analyse- en/of rapportagedoeleinden kan een kanttekening volstaan. Hierbij wordt bijvoorbeeld vermeld dat de getoonde resultaten voor xx procent betrouwbaar zijn voor de beoogde doeleinden.
In de meeste gevallen moeten de problemen echter structureel opgelost worden. Het is verstandig hierbij te beginnen met problemen van technische aard. Problemen van technische aard kunnen namelijk van invloed zijn op de uitkomst van functionele en business rule controles. Ditzelfde kan ook gelden voor de problemen in de functionele gradatie op de controles in de business rule gradatie.
De tweede reden om eerst de technische fouten op te lossen, is de inspanning die vereist is. Het oplossen van problemen in de functionele en business rule gradatie zal significant meer inspanning kosten in vergelijking met de technische gradatie. Het uitgangspunt is dat het systeem om datakwaliteit te meten en de (gerelateerde) dashboards zo zijn ingericht dat het hele proces eenvoudig volledig herhaalbaar is.

## Continu meten en verbeteren

Om de datakwaliteit door de tijd heen te kunnen borgen is een continu proces van meten en verbeteren noodzakelijk. Hierbij zijn de dashboards een handig instrument om de ontwikkeling van de datakwaliteit over tijd inzichtelijk te maken (zie afbeelding 3).
Het is van groot belang dat naarmate de tijd vordert regelmatig gekeken wordt of het bestaande proces nog afdoende is. Het kan zijn dat andere stakeholders benoemd zijn. De gedefinieerde validatieregels (en hun implementatie) moeten regelmatig getoetst worden op correctheid en toepasbaarheid. Zodra de verkeerde controles worden uitgevoerd op de juiste data heeft het hele proces weinig tot geen toegevoegde waarde.

## Conclusie

Vanuit de dagelijkse praktijk is een aantal handvatten aangereikt rondom het opzetten van een herhaalbare datakwaliteitsmeting. Indien u delen uit dit artikel kunt gebruiken in uw dagelijkse praktijk of als u informatie uit dit artikel als bevestiging van uw eigen doen en laten beschouwt dan zijn wij in onze optiek in onze missie geslaagd.

**Ruud Kuil en Sinbad Moors**
Ruud C. Kuil is Senior Consultant en Sinbad J. Moors is Managing Consultant, beiden werkzaam bij Capgemini op het gebied van Data Management.