Een voorbeeld over definitieverschil in de Formule 1 Grandprix

Meten, verbeteren en voorkomen

Henk van Roekel

Iedereen zal het eens zijn dat datakwaliteit belangrijk is, maar als het aankomt op daadwerkelijk meten en verbeteren van datakwaliteit lopen meningen en inzichten behoorlijk uiteen. Dit artikel geeft inzicht in belang en definitie van datakwaliteit aan de hand van een concreet voorbeeld uit de Formule 1 Grandprix. En er wordt gekeken naar de wijze waarop datakwaliteit op een continue basis gemeten en verbeterd kan worden.

Om een beeld te krijgen van het belang van datakwaliteit en de aspecten die daarbij een rol spelen, gebruiken we een voorbeeld van de communicatie tussen een coureur en de pitsbox van zijn team tijdens een race. Hieronder is een deel van de communicatie tussen een Duitse coureur en een Amerikaanse pitsbox-medewerker weergegeven in de cruciale fase van een race. De coureur staat op de eerste positie en weet dat dit de beslissende race is voor het gehele kampioenschap van dit seizoen. Hij merkt dat de olietemperatuur oploopt en wil graag weten wat zijn voorsprong is op nummer twee. De data-uitwisseling leidt tot een fiasco.

Ronde 34. Coureur: "Olietemperatuur 200, moet ik binnenkomen?" Pitsbox-medewerker: "Binnen toelaatbare marge, doorrijden."
Ronde 35. Coureur: "Positie?" Pitsbox-medewerker: "1e plaats, halve ronde voorsprong."
Ronde 36. Coureur: "Wie zit er achter mij dan?" Pitsbox-medewerker: "Een achterblijver denk ik."
Ronde 37. Coureur: "Olietemperatuur 250!" Pitsbox-medewerker: "Binnen toelaatbare marge, doorrijden."
Ronde 38. Coureur: "Ik wordt ingehaald door nummer twee!"
Ronde 39. Coureur: "Motor staat in brand!"

De coureur eindigt de race op de laatste positie met een uitgebrande auto in de grindbak. Hiermee behaalt hij deze race geen punten en valt terug naar de derde plaats in het totale klassement van het kampioenschap. Wat is hier mis gegaan? Bij een analyse van de communicatie vanuit perspectief van datakwaliteit vallen enkele zaken op.
In de communicatie wordt gesproken over de olietemperatuur en het wel of niet binnen de toelaatbare marges zijn van deze temperatuur. Onduidelijk is echter in de communicatie of gesproken wordt over graden Celsius of Fahrenheit! Uit het vervolg blijkt dat de Duitse coureur de temperatuur in graden Celsius bedoelde, terwijl de Amerikaanse pitsbox-medewerker uitging van de olietemperatuur in graden Fahrenheit. De uitwisseling van data was in dit geval dus niet consistent, met alle gevolgen van dien.
Er wordt informatie uitgewisseld over de positie van de nummer twee op de baan. Hier valt op dat in eerste instantie wordt aangegeven dat de coureur een halve ronde voorsprong heeft op de nummer twee, vervolgens dat er vermoedelijk een achterblijver achter de coureur rijdt. Uiteindelijk blijkt de vermoedelijke achterblijver de nummer twee te zijn! Hier is duidelijk sprake van een probleem met de correctheid en tijdigheid van belangrijke informatie. In eerste instantie heeft de pitsbox slechts een vermoeden dat het een achterblijver is. Te laat ontdekt uiteindelijk de coureur zelf dat het de nummer twee in de race was, die nu de race heeft gewonnen.
Wanneer in de communicatie de datakwaliteit goed was geweest had de pitsbox de coureur kunnen aangeven om iets langzamer te gaan rijden om de olietemperatuur niet verder op te laten lopen. Daarmee had de coureur op de tweede plaats in de race kunnen eindigen en zijn leidende positie in het klassement behouden.

Aspecten

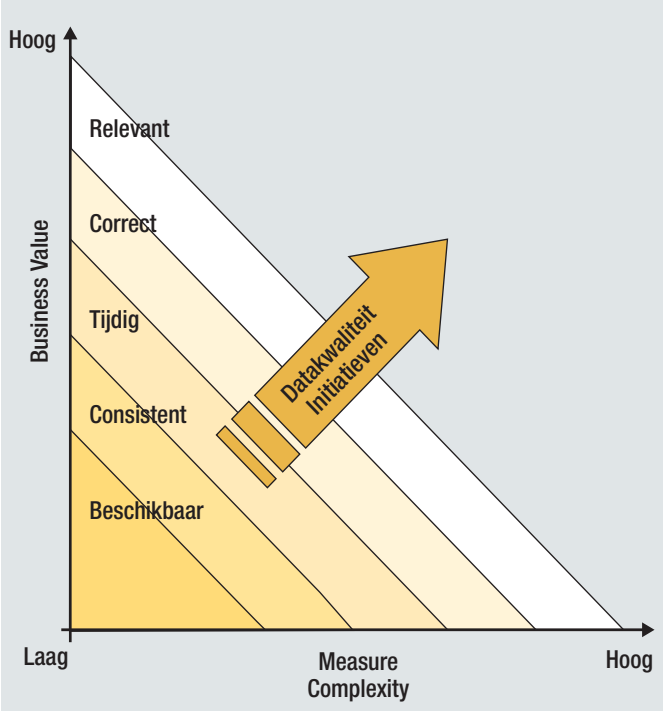
Uit dit voorbeeld blijkt dat de kwaliteit van data heel bepalend kan zijn in de prestaties van een persoon of organisatie. Tevens zien we dat een aantal aspecten een rol speelt.
Beschikbaarheid. Een eerste belangrijk aspect is het beschikbaar zijn van benodigde gegevens. Zonder beschikbare gegevens is het niet mogelijk om rationeel te sturen. In het voorbeeld was er slechts een vermoeden dat het een achterblijver betrof. Werkelijke data om dit te bevestigen waren kennelijk (nog) niet beschikbaar. Dit scenario lijkt veel op de praktijksituatie in organisaties waar

vaak op basis van een onderbuikgevoel besluiten genomen worden bij gebrek aan betrouwbare data.

Consistentie. In veel gevallen zijn binnen een organisatie data op meerdere plekken opgeslagen en aanwezig, vooral wanneer het gaat om afgeleide data voor besluitvorming. Niet zelden blijken gegevens die feitelijk gelijk zouden moeten zijn, in de praktijk dit niet te zijn. Vaak wordt dit veroorzaakt door interpretatieverschillen bij de verwerking van originele gegevens tot de afgeleide besturingsinformatie. Ook kunnen we hier te maken hebben met synchronisatieproblematiek; gaan de data wel over hetzelfde tijdstip of over dezelfde periode. Om deze definitieverschillen zichtbaar te maken en te voorkomen is een goede definitie van begrippen en data-elementen essentieel, ook wel masterdata management genoemd. In het voorbeeld is een definitieverschil over temperatuurmeting de coureur fataal.

Tijdigheid. Om effectief te kunnen handelen wanneer de situatie dat vereist is het noodzakelijk gegevens tijdig paraat te hebben. Volledigheid en consistentie van data bereiken kan een tijdrovend proces zijn. De balans vinden tussen perfecte data en tijdige data is hierbij essentieel. Na de race melden dat de achterblijver eigenlijk de nummer twee was levert niets meer op. Als door tijdsdruk volledigheid en consistentie van data niet gewaarborgd kan worden, moet dit altijd expliciet gemaakt worden. Deze onzekerheid kan dan meegewogen worden in eventuele besluiten op basis van deze gegevens. Wanneer in het Formule 1 Grandprix voorbeeld door de pitsbox aangegeven was dat de positie van de auto achter de coureur onbekend was, had de coureur wellicht nog een keer extra in zijn spiegel gekeken!

Correctheid. Consistente, tijdige en beschikbare data klinkt al heel goed. Maar wat nu als de data de werkelijkheid niet goed weergeven? Bijvoorbeeld door verkeerde invoer in een administratief systeem. Het probleem met de olietemperatuur kan ook veroorzaakt zijn door een invoerfout van de temperatuurmarge in de administratie van de pitsbox.



Afbeelding 1: Datakwaliteit-aspecten.

> Metingen die eenmaal ingesteld zijn kunnen als reguliere controleprocessen worden uitgevoerd

Relevantie. Het laatste aspect, maar zeker niet het minst belangrijke, in datakwaliteit is de relevantie van de data. Vaak worden we tegenwoordig overstelpt met informatie, maar hoeveel van deze informatie is nu werkelijk relevant voor de beslissingen die genomen moeten worden of de vragen die beantwoord moeten worden? Dit is een veel voorkomend probleem in datawarehouse-omgevingen, de gebruiker wordt overspoeld met beschikbare data zodat het selecteren van relevante data een grote uitdaging is. De werkelijke waarde van de aanwezige wel relevante data wordt hierdoor onvoldoende benut.

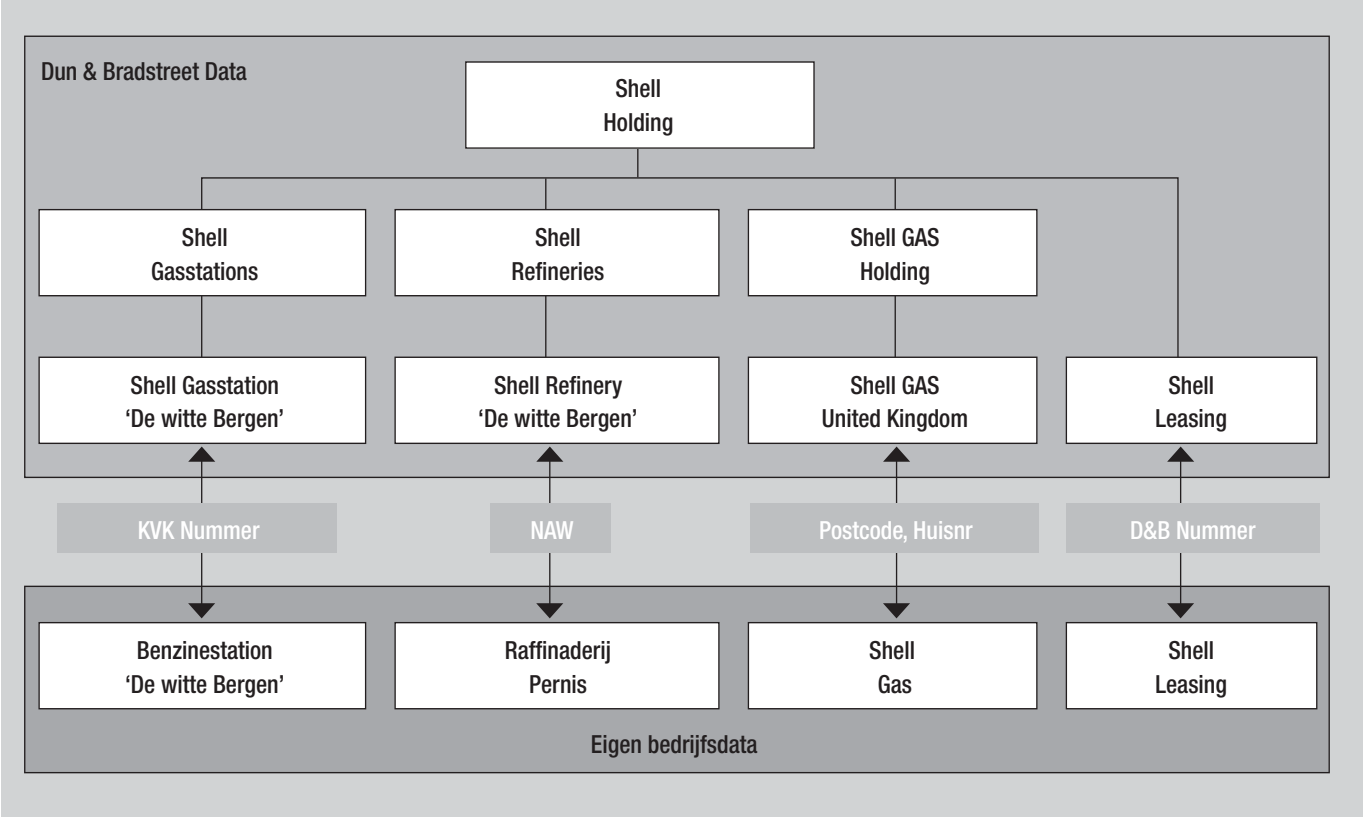
Meten

Nu we bekend zijn met de aspecten die van belang zijn voor de datakwaliteit, staan we voor de volgende uitdaging. Namelijk, hoe meten we datakwaliteit. Het is belangrijk om te beseffen dat datakwaliteit geen ICT-probleem is. Het zijn de gebruikers en business-eigenaren van informatiesystemen die verantwoordelijk zijn en baat hebben bij een goede datakwaliteit in hun systemen. Ook zijn zij de partijen die kunnen beoordelen wat datakwaliteit is en of deze aansluit bij de noodzakelijke kwaliteit in de bedrijfsprocessen waarvoor zij verantwoordelijk zijn. Wel kan een ICT-organisatie ondersteuning geven bij het meten van de bestaande datakwaliteit, het verbeteren van de datakwaliteit en het voorkomen van datakwaliteitproblemen.

In afbeelding 1 is het belang van de verschillende te meten aspecten in relatie tot de inspanning van het meten van deze aspecten weergegeven. Hieruit blijkt dat het aspect relevantie bijvoorbeeld erg moeilijk te meten is, maar tegelijkertijd wel het belangrijkste aspect is in de uiteindelijke toegevoegde waarde van de betreffende dataverzameling.

Toch kan het meten van bestaande datakwaliteit beginnen met een relatief eenvoudige inventarisatie van de data in ICT-systemen. Zeker wanneer deze systemen gebruik maken van een relationele database kan met eenvoudige query's een eerste overzicht gemaakt worden. Op basis hiervan kan door de ICT-organisatie een lijst met verdachte situaties opgesteld worden die vervolgens door de systeemeigenaar en gebruikers verder geïnterpreteerd kan worden. Voorbeelden van eenvoudig te produceren lijsten zijn:

- Aantallen verschillende waarden per attribuut in de tabel;
- Per attribuut de minimale, maximale en gemiddelde waarde;



Afbeelding 2: Dataverrijking.

- Van de tien meest voorkomende en minst voorkomende waarden in een attribuut het volledige record tonen;
- Controle van integriteit van de data in het model door het volgen van alle primary key- en foreign key-relaties;
- Het tellen van het aantal voorkomens van null-values in attributen;
- Totaal tellingen en sommaties op numerieke waarden;
- Voldoen de waarden in code-attributen aan de gestelde domeinwaarden.

> De balans vinden tussen perfecte data en tijdige data is essentieel

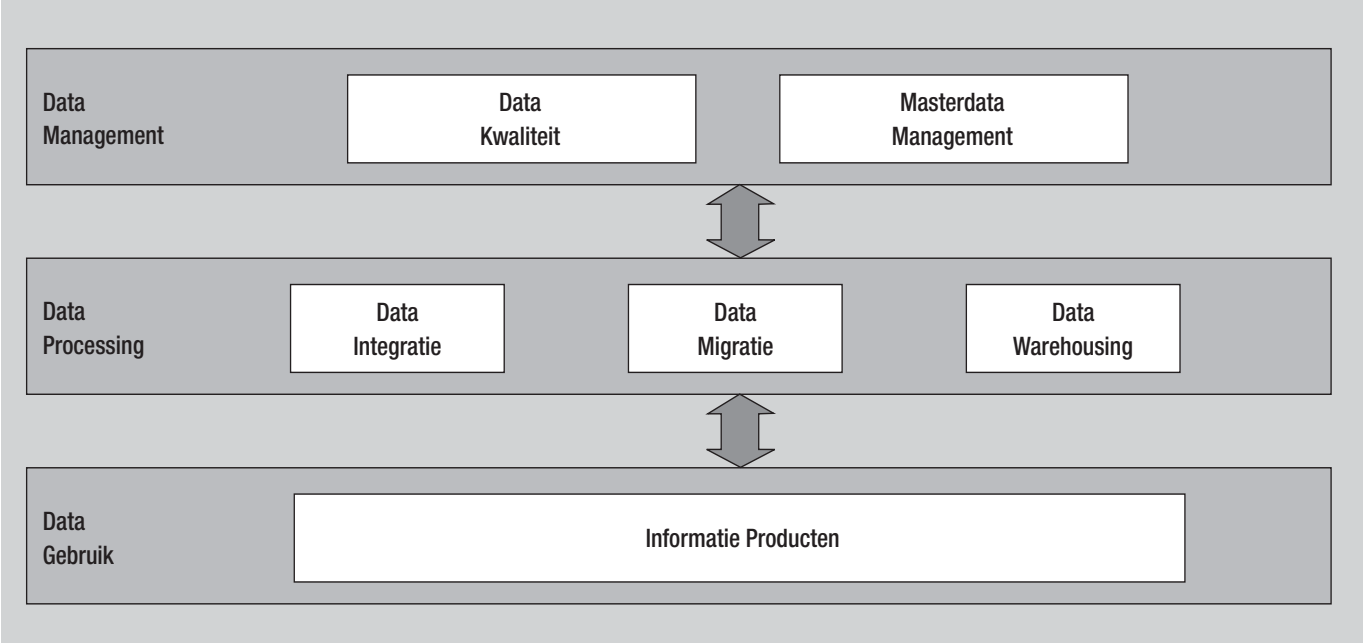
Op deze wijze kan met eenvoudige hulpmiddelen snel een groot deel van de problemen geïdentificeerd worden. Op basis van evaluatie van de resultaten uit deze eerste metingen door de systeemeigenaar en gebruikers kan verdere actie ondernomen worden.

Wanneer de behoefte bestaat om verdere analyses uit te voeren kan gebruik gemaakt worden van specifieke hulpmiddelen. Deze hulpmiddelen vallen onder de categorie 'data profiling tools' en zijn vaak onderdeel van een data-integratie of datawarehouse productportfolio. Zo bevatten bijvoorbeeld de data-integratie- en datawarehouse-producten van Oracle, IBM en Informatica alle een data profiling component. Met een data profiling tool kan op basis van complexe beslissingslogica de datakwaliteit diepgaander bepaald worden. Bijkomend voordeel van deze producten is dat metingen die eenmaal ingesteld zijn als reguliere meet- en controleprocessen kunnen worden uitgevoerd.

Verbeteren

Op basis van de meting kan besloten worden tot verbetering van datakwaliteit op die plaatsen waar er de grootste behoefte aan is voor de informatievoorziening in bedrijfsprocessen. De verbetering kan uitgevoerd worden door het toepassen van business rules op de bestaande data. Deze business rules worden bepaald door de systeemeigenaar en gebruikers en toegepast door de ICT-organisatie. Hierbij kan ervoor gekozen worden om deze business rules toe te passen op de data en deze daarmee werkelijk te wijzigen. Ook kan gekozen worden om business rules toe te passen bij de selectie van de data, hierbij blijven de originele data behouden en worden de data getoond na toepassing van de business rule. Het voordeel van de laatste methode is dat oorspronkelijk ingevoerde data altijd aanwezig blijven en herleidbaarheid van informatie op die manier maximaal gegarandeerd kan worden.

Een tweede mogelijkheid is het verrijken van bestaande data op basis van externe informatie. Hierbij is te denken aan de aankoop van persoons- of bedrijfsgegevens van partijen zoals de Kamer van Koophandel of D&B (Dun & Bradstreet). Deze partijen



Afbeelding 3: Informatie-architectuur.

garanderen de datakwaliteit van de gegevens die zij leveren. In afbeelding 2 is te zien hoe een koppeling van eigen gegevens aan de Dun & Bradstreet gegevens leidt tot een compleet inzicht in de bedrijfstructuur van een specifieke klant. Zonder deze verrijking waren slechts individuele klanten bekend.

> Eerste belangrijke component is het goed inrichten van masterdata management

Koppeling aan eigen bedrijfsdata levert verbetering van datakwaliteit en verrijking van de eigen data. Ook zijn er producten op de markt die een organisatie kunnen ondersteunen bij geautomatiseerde koppeling van eigen bedrijfsgegevens aan externe gegevensverzamelingen. Een voorbeeld hiervan is Human Inference; deze functionaliteit wordt combineerd met functies voor ontdubbelen en schonen van interne data.

Voorkomen van problemen

Het meten en verbeteren van datakwaliteit is een continu proces waarmee we het totale datakwaliteitniveau binnen een organisatie proberen te verhogen. Naast meten en verbeteren is het essentieel om problemen aan de bron te voorkomen, anders blijft het dweilen met de kraan open. Een eerste belangrijke component hierbij is het goed inrichten van masterdata management, daardoor wordt een organisatie in staat gesteld haar datadefinities en onderlinge verbanden tussen de definities te documenteren en te beheren. Belangrijk hierbij is dat dit niet alleen een omgeving is waarin gedocumenteerd wordt, zoals vele van de huidige metadata management-oplossingen, maar dat de omgeving ook geïntegreerd kan worden in de IT-systemen van een organisatie. Zodat bijvoorbeeld de domeindefinitie van een attribuut olietemperatuur actief gebruikt wordt om in alle operationele IT-systemen waarin dit attribuut voorkomt automatisch deze domeindefinitie toe te passen.

Conclusie

Het proces van meten, verbeteren en voorkomen moet worden gezien als een integraal onderdeel van informatie management binnen een organisatie. Essentieel hierbij is de betrokkenheid van systeemeigenaren en gebruikers als business-eigenaren van de gegevens. Zij zijn bepalend in het vaststellen van de bestaande datakwaliteit en het te bereiken niveau van datakwaliteit. Met een goede inbedding van datakwaliteit meet- en verbeterinstrumenten in de informatie-architectuur, zoals weergegeven in afbeelding 3, kan een ICT-organisatie haar klanten hierbij ondersteunen.
Er zijn drie lagen te onderscheiden. Ten eerste de data management-laag. Masterdata management en datakwaliteit zijn hierin de randvoorwaardelijke componenten voor een goede invulling van een informatie-architectuur. Op basis van een goede invulling van de datamanagement-laag wordt de data processing-laag opgebouwd. Hier zijn alle gegevensverwerkende processen gepositioneerd zoals die voorkomen in data-integratie, data-migratie en datawarehousing. Pas in de laatste laag moet de echte toegevoegde waarde blijken. Hier vindt het werkelijke gebruik van de data plaats in de informatieproducten voor een organisatie.

Henk van Roekel (henk.van.roekel@logicacmg.com) is Certified Business Intelligence Professional (CBIP) en werkt als principal consultant bij LogicaCMG.