Het kwantificeren van datakwaliteit aan de hand van meerdere dimensies

Walter Steven Beaujon

Master Stageverslag

Business Analytics

September 2012

Publieke Versie

Het kwantificeren van datakwaliteit aan de hand van meerdere dimensies

Walter Steven Beaujon

Business Analytics
Master Stageverslag
Vertrouwelijke versie

Vrije Universiteit Amsterdam
Faculteit der Exacte Wetenschappen
Studierichting Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam

Stagebedrijf:

Deloitte Risk Services B.V.
Data Analytics
Laan van Kronenburg 2
1183 AS Amsterdam
September 2012

Begeleider VU:
Prof. Dr. Chris Verhoef

Tweede Lezer:
Dr. J.F.M. Feldberg

Begeleider Deloitte:
H.E. Visser

Voorwoord

In het laatste jaar van mijn masteropleiding Business Analytics, voorheen Business Mathematics & Informatics, heb ik gedurende zes maanden stage gelopen bij Deloitte Risk Services B.V. binnen de afdeling Data Analytics. Het was een eer om een kans te krijgen om mezelf te kunnen verdiepen in verschillende aspecten van data bij een grote multinationaal bedrijf. De stage opdracht gaat over het vinden van de beste manier om datakwaliteit te kwantificeren aan de hand van meerdere criteria en om deze als service aan te kunnen bieden.

Ik zou graag mijn begeleider bij Deloitte, Hinke Visser, willen bedanken voor haar commentaar en medewerking tijdens de uitvoering van de stage en haar hulp en tips bij het schrijven van deze scriptie. Ik zou de senior manager van Data Analytics, Norbert van Haaften, ook willen bedanken voor zijn feedback en interesse in het onderwerp. Aan de VU zou ik mijn begeleider en tweede lezer ook willen bedanken voor hun hulp. Mijn begeleider Prof. Dr. Chris Verhoef heeft vanaf het begin een kritische blik geleverd die het hele project de juiste kant op heeft gestuurd. Mijn tweede lezer, Dr. Frans Feldberg, was altijd beschikbaar en stipt wanneer dit nodig was. Ik heb bij de medewerkers van Data Analytics een warm onthaal gevonden en zij waren ook altijd behulpzaam bij het programmeren en ondersteunend tijdens mijn stage.

Steven Beaujon

Amstelveen

25/09/2012

Samenvatting

Vertrouwelijke Informatie

Inhoudsopgave

1 Inleiding

Vertrouwelijke Informatie

2 Probleembeschrijving

Dit hoofdstuk de probleembeschrijving en de wat deze betekend voor Deloitte. Het eerste deel bevat het bedrijfscontext van Deloitte waarin de relevante afdelingen van Deloitte worden toegelicht. Het tweede deel bevat de probleembeschrijving en waarom dit onderzoek belangrijk is. Het laatste deel bevat een uitleg van het hele traject waarin Deloitte datakwaliteit monitort.

2.1 Bedrijfscontext Deloitte

2.1.1 Risk Services

De meest succesvolle bedrijven begrijpen dat risico's een belangrijk aspect van de bedrijfsvoering zijn en dat deze met het juiste beheer zal leiden tot gewenste resultaten. Deloitte Risk Services levert diensten gebundeld op het gebied van risicomanagement & control, inclusief IT-auditdiensten. Dat betekent signaleren, analyseren, beoordelen en managen van risico's. Deze aanpak helpt klanten met het volgende:

- In perspectief brengen van hun risico's.
- Organisatorische kloven overbruggen om risico te beheren.
- Niet alleen risico's verkleinen, maar ook intelligent risico's nemen om daar profijt van te krijgen.

De afdeling Risk Services bestaat uit de competentie teams: Controls, Data & Fraud, Security en Financial Risk Management.

2.1.2 Data Analytics

Het Data Analytics team, wat deel uitmaakt van Data & Fraud, is ontstaan uit de groeiende behoefte om steeds sneller en scherper beslissingen te nemen met bedrijfsgegevens. Dit eist diepe sectorkennis en beheersing van analytische technologie. Door deze effectief te combineren maakt het Data Analytics team het mogelijk voor een bedrijf om in staat te zijn de belangrijkste vragen te stellen en om de gepaste antwoorden te vinden. Het team biedt onder andere de volgende diensten aan:

- De Factuurontdubbelaar: de controle of bepaalde facturen niet per ongeluk meerdere malen zijn betaald.
- SAS99 Grootboek analyse: de controle op mogelijke frauduleuze boekingen in het grootboek.
- BCF radar: de controle of de klant te weinig BTW heeft teruggevraagd van de belastingdienst.
- Factuur vs. Contract analyse: de controle of facturen die opdrachtgevers hebben ontvangen niet te hoog zijn in vergelijking met het bijbehorende contract.

2.2 Probleembeschrijving

Elke organisatie is afhankelijk van data bij het nemen van beslissingen. De hoeveelheid data binnen een organisatie blijft ook steeds toenemen en hierdoor is het cruciaal dat deze data zo betrouwbaar mogelijk is opgeslagen. Datakwaliteit is de mate van betrouwbaarheid binnen een database en deze moet zo hoog mogelijk zijn om effectief de data te kunnen analyseren.

De effecten van betrouwbare data zijn als volgt:

- Kostenverlaging: door mogelijkheden te identificeren om kosten te besparen en uitgaven te beperken.
- Omzetverhoging: door mogelijkheden te identificeren om meer producten en diensten met meer waarde te verkopen en de klant correct te factureren.
- Vrijmaken van werkkapitaal: door mogelijkheden te identificeren om de balans te verbeteren, de voorraadefficiency te verbeteren en de inkomende kasstroom te versnellen.
- Bevordering van compliance: door mogelijkheden te identificeren om de kwaliteit en effectiviteit van interne controles te verhogen.

Lage datakwaliteit kan op meerdere manieren de bedrijfsvoering hinderen. L.P. English schat dat een bedrijf 10 tot 20 procent van haar omzet zal verliezen door lage datakwaliteit[1]. De effecten van slechte data kwaliteit zijn als volgt:

- Hoge onderhoudskosten van de data
- Verkeerde conclusies trekken
- Een grote meerderheid van de data conversies mislukt
- Meerdere betalingen aan dezelfde leverancier
- Het onderzoeken van fraude is zo goed als de onderliggende data toelaat
- Langzame systemen
- Minder vertrouwen van aandeelhouders, klanten en belanghebbenden

De volgende twaalf statistieken (Haug, 2010[2]) verwijzen naar het belang van datakwaliteit:

- 88 procent van alle data integratie projecten mislukken helemaal of overtreffen hun budget.
- 75 procent van alle organisaties herkennen dat ze extra kosten hebben wegens vervuilde data (vervuilde data is data waar de datakwaliteit met de tijd is verslechterd).
- 33 procent van alle organisaties hebben nieuwe IT systemen vertraagd of volledig geschrapt door vervuilde data.
- De VS verliest jaarlijks 611 miljard dollar aan slecht gerichte campagnes via de post en overheadkosten.
- Volgens Gartner[1] is slechte data de grootste oorzaak voor het mislukken van CRM systemen.
- Minder dan 50 procent van alle bedrijven beweren vertrouwen te hebben in de kwaliteit van hun data.
- Business Intelligence (BI) projecten mislukken vaak door slechte data.
- Alleen 15 procent van alle bedrijven heeft vertrouwen in de datakwaliteit van externe data dat ze ontvangen.
- Klantdata gaat gemiddeld met 2 procent per maand of met 25 procent per jaar achteruit.
- Organisaties overschatten meestal de kwaliteit van hun data en de onderschatten de kosten van fouten.

[1] De Gartner Group

- Bedrijfsprocessen, klantenverwachting, bronsystemen en compliance regels blijven constant veranderen en datakwaliteit management systemen moeten deze veranderingen meegaan.
- Veel tijd en geld wordt besteed aan het zoeken naar korte termijn oplossingen voor dringende crisissen in plaats van richten op lange termijn problemen.

Het onderzoeken van de kwaliteit van data gaat over het algemeen in verschillende fases. De twee belangrijkste fases bij het onderzoek naar datakwaliteit zijn de Data Assessment en Data Cleansing fases. Deze fases worden in de volgende paragraaf grondig toegelicht. Bij Data Assessment wordt er getest op bedrijfsregels, standaarden, plausibiliteit en validiteit. Als deze fase met succes wordt afgerond kan deze informatie gebruikt worden in de Data Cleansing fase. Data Cleansing is de fase waarin de data wordt opgeschoond (fouten repareren) om het geschikt te maken voor de algemene systemen en processen.

Het meten van datakwaliteit kan aan de hand van verschillende criteria. In de Data Assessment fase kan men ontdekken in hoeverre de database voldoet aan deze criteria en kan deze informatie aan de klant worden teruggekoppeld. Deze resultaten bestaan uit databases die meerdere malen groter zijn dan de oorspronkelijke database, door het feit dat alle waarden in de database gemeten moeten worden aan de hand van alle verschillende criteria. De resultaten die de klant ontvangt geven niet meteen een duidelijke indicatie van het niveau van datakwaliteit en zullen tot veel verwarring leiden. Als deze resultaten kunnen worden samengevat in een duidelijke score zou de klant meteen kunnen begrijpen waar deze aan toe is. Dit leidt dan tot de belangrijkste onderzoeksvraag:

Hoe kan men datakwaliteit met een enkele score kwantificeren?

Door een score van bijv. nul tot tien te koppelen aan de datakwaliteit van de klantendata kan de klant in één oogopslag zien hoe geschikt deze data is voor de doel waarvoor deze gebruikt wordt. De bedoeling is om een model op te stellen dat datakwaliteit kan kwantificeren aan de hand van tien verschillende criteria voor datakwaliteit. Voor elk van de tien dimensies dient een cijfer te worden berekend. Deze tien cijfers zullen aan de hand van een passende formule tot een totale score voor datakwaliteit leiden, genaamd de aggregaat score. Dit model zal gebruikt worden bij het ontwerpen van een tool die deze score kan berekenen voor een gegeven database. Deze tool zal met SAS geprogrammeerd worden en zal verder in deze scriptie de *SAS Data Quality Tool* genoemd worden. Belangrijke informatie over de database zal verkregen worden aan de hand van een vragenlijst die door de klant zal worden ingevuld. De informatie uit de vragenlijst zal worden omgezet in een bestand die de database zal ondersteunen. Dit ondersteunend bestand zal het *template* genoemd worden. Verder is het ook van belang dat deze tool een output kan leveren waar de klant kan zien waar er verbetering nodig is.

Als de SAS Data Quality Tool werkend is zal deze getest worden op verschillende test- en werkelijke databases. Deze tests zullen dienen om de tien dimensies te beoordelen om belangrijke dimensies te onderscheiden van minder belangrijke dimensies. De mogelijke formules zullen ook geëvalueerd worden om vervolgens de meest geschikte formule te kiezen om de uiteindelijke score te berekenen.

2.3 Datakwaliteit framework

Vertrouwelijke Informatie

3 Literatuuronderzoek

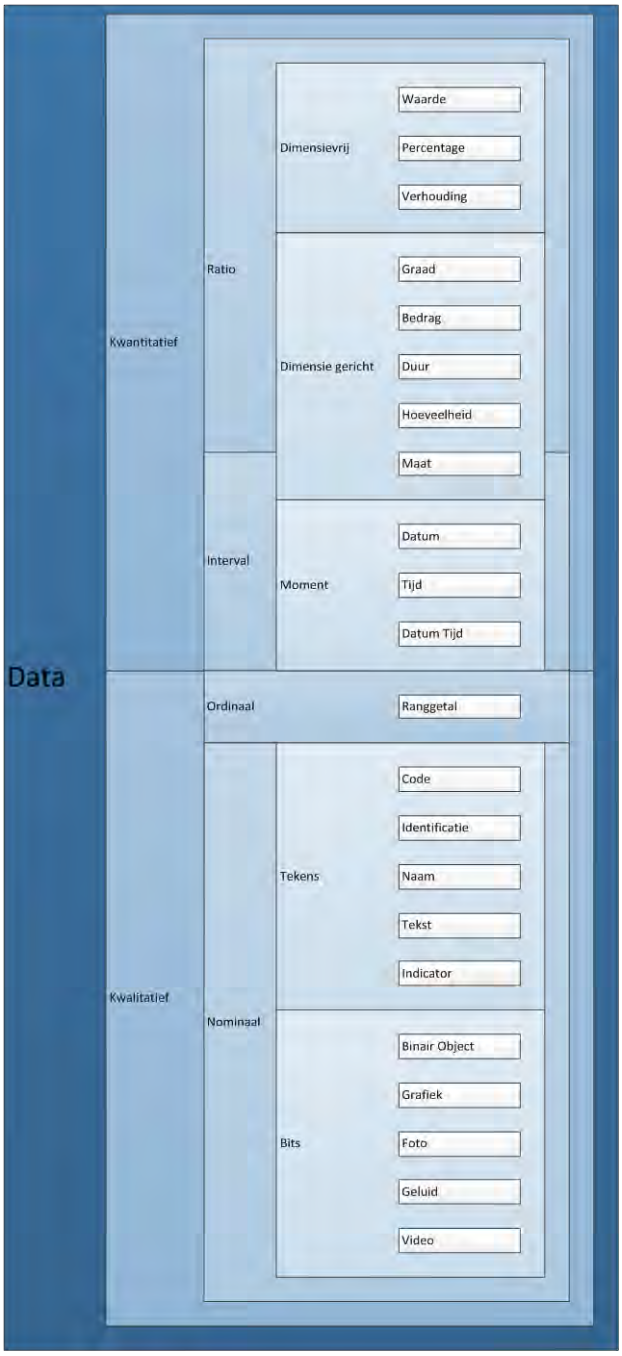
In dit hoofdstuk wordt de uitkomst van het literatuuronderzoek beschreven. Het literatuuronderzoek omvat de criteria voor datakwaliteit, een categorisatie van alle mogelijke datavelden, beschikbare datakwaliteit tooling, een uitleg van data quality scorecards en een verzameling van mogelijke scoring formules.

3.1 Criteria voor datakwaliteit

Vertrouwelijke Informatie

3.2 Categorisatie van mogelijke datavelden

Om klantdata beter te kunnen begrijpen is het belangrijk dat men begrijpt waar de data voor nodig is en op welke manieren deze kan worden weergegeven. Er zijn oneindig veel verschillende datavelden mogelijk en ieder behoort op een specifieke manier getest te worden voor de tien criteria. Door de mogelijke datavelden in categorieën op te delen kan men een beter idee krijgen van de mogelijke datakwaliteit operaties die per categorie van toepassing zullen zijn. De statistiek maakt onderscheid tussen de volgende vier niveaus: Nominaal en Ordinaal (kwalitatief) en Interval en Ratio (kwantitatief)[9]. De Verenigde Naties definieert 22 verschillende kern data types in een rapport over data componenten[10]. Deze kunnen in zes verschillende groepen worden verdeeld en deze zes groepen passen binnen de vier genoemde niveaus. In Figuur 3 staat een overzicht van de categorieën en subcategorieën. Deze worden in de volgende paragrafen verder toegelicht.



Figuur 1: De mogelijke categorieën van datavelden

3.2.1 Kwalitatieve datavelden

Kwalitatieve datavelden zijn alle velden die referenties naar niet-rekenkundige attributen bevatten. Deze velden bestaan uit ordinale velden en nominale velden. Kwalitatieve velden bevatten strings van getallen, letters en andere leestekens waar niet mee gerekend kan worden.

3.2.1.1 Nominale schaal

De nominale schaal is het simpelste meetniveau. Daarbij gaat het, zoals de naam al aangeeft, bij het meten slechts om de naamgeving. Dit niveau bestaat uit strings van tekens. Nominale velden bestaan uit twee groepen velden: Bits en Tekens. Bits zijn strings van alleen 0'en en 1'en. Tekens bestaan uit getallen, letters en andere leestekens. Alle 4 niveaus bevatten op zijn minst een nominale schaal. Nominale waarden kunnen gesorteerd worden in een alfabetische volgorde, maar deze volgorde heeft geen invloed op de werkelijke betekenis van waarden die naast elkaar zijn.

Bankrekeningnummers en Sofinummers zijn ook nominale velden, omdat de nummers bestaan als referentie naar een ander object.

3.2.1.2 Ordinale schaal

Metingen op ordinaal niveau kennen een natuurlijke ordening. Deze worden gebruikt om aan te geven dat bepaalde waarden volgens hun definitie boven andere waarden liggen. Een voorbeeld is de 5-puntsschaal bij enquêtes (zeer mee oneens - mee oneens - neutraal - mee eens - zeer mee eens). Bij een ordinale schaal is de volgorde duidelijk, maar zijn de verschillen niet interpreteerbaar: 'zeer mee eens' ligt niet noodzakelijk net zo ver boven 'mee eens' als dat 'mee eens' boven 'neutraal' ligt. De intervalschaal en ratioschaal zijn overigens ook ordinaal relevant.

3.2.2 Kwantitatieve datavelden

Kwantitatieve datavelden geven getallen weer. Deze kunnen bestaan uit ratio's of intervallen. Deze velden bestaan uit getallen. Deze getallen moeten allemaal wiskundig relevant zijn, waardoor bijvoorbeeld ordinale numerieke getallen niet onder deze categorie vallen.

3.2.2.1 Intervalschaal

De intervalschaal geeft een numerieke waarde aan samen met een dimensie. Het nulpunt is niet van speciaal belang, maar verschillen wel. Een voorbeeld is dat het verschil tussen het jaar 1950 en 1965 gelijk is aan het verschil tussen 1997 en 2012. De telling van jaren is weliswaar op een willekeurig moment begonnen, waardoor het duidelijk is dat jaar 0 niet het begin van tijd is. Momenten in de tijd vallen onder het intervalniveau, maar ook sommige maten, zoals temperatuur in Celsius graden.

3.2.2.2 Ratioschaal

Naast de kenmerken van een intervalschaal heeft de ratioschaal ook een absoluut nulpunt. Daarmee hebben ook verhoudingen van waarden op deze schaal betekenis. De groepen die onder dit niveau vallen zijn de dimensievrije datavelden en dimensiegerichte velden. Maten zoals lengte in meters of temperatuur in Kelvin zijn op de ratioschaal, omdat het quotiënt van twee waarden een zinvolle dimensieloze grootheid wordt.

3.2.3 De 22 kern data types

Hier volgen de 22 kern data types die de Verenigde Naties herkennen.

3.2.3.1 Dimensievrije waarden

Dimensievrije waarden zijn numerieke waarden die niet refereren naar een bepaalde dimensie, maar bestaan op zichzelf.

Waarde

Waarden zijn numerieke waarden die algebraïsch relevant zijn. De waarden 1, 2 en 3 in de som 1+2=3 zijn waarden zolang er niet wordt geïmpliceerd dat de waarden naar specifieke objecten refereren.

Percentage

Percentages zijn numerieke waarden die refereren naar fracties van honderd. Als 19% van je uitgaven aan BTW wordt uitgegeven is die 19% een percentage. Dit kan ook als 0.19 geschreven worden.

Verhouding

Verhouding refereert naar de verhouding tussen twee onafhankelijke waarden uit dezelfde dimensie. Bijvoorbeeld een lat van 2 meter is 2/3 zo lang als een lat van 3 meter. De verhouding zelf is in dit geval 2/3. Deze kan in de data staan als '2/3' of als 0.666.

3.2.3.2 Dimensiegerichte waarden

Dimensiegerichte waarden zijn numerieke waarden die refereren naar een hoeveelheid van een bepaalde maat of object. Het object zelf kan genoemd worden, maar mag ook geïmpliceerd zijn.

Bedrag

Een bedrag is een numerieke waarde van een bepaalde valuta.

Duur

Duur is een numerieke waarde van een bepaald tijdsinterval, zoals dag, maand, uur, seconde of een fractie daarvan. De leeftijd van een persoon valt ook onder duur.

Hoeveelheid

Een hoeveelheid is een niet monetaire numerieke waarde van een bepaald object of eenheid.

Maat

Een maat is een numerieke waarde wat refereert naar een unieke maat, zoals lengte of temperatuur.

Graad

Een graad is een numerieke waarde wat bestaat uit een verhouding tussen twee verschillende dimensies, zoals kilometer per uur (km/u) of kilogram per vierkante centimeter (kg/cm^2).

3.2.3.3 Momenten

Momenten zijn instanties in de tijd die geformatteerd kunnen zijn op verschillende manieren, zoals integers, reële waarden of met meer traditionele dimensies, zoals jaar, maand, week, dag, uur, minuut en seconde.

Datum

Data zijn momenten op de Gregoriaanse kalender die geformatteerd kunnen zijn in de vorm van een integer of meer traditionele dimensies, zoals jaar, maand, week en dag.

Tijd

Tijden zijn momenten in een dag die geformatteerd kunnen zijn in de vorm van een reëel getal of meer traditionele dimensies, zoals uur, minuut, seconde of een fractie van een seconde.

Datum Tijd

Datum Tijd waarden zijn momenten in de tijd die kalenderdagen zowel als de tijd meegeven.

3.2.3.4 Ordinale waarden

Ordinale waarden zijn numerieke waarden die een natuurlijke volgorde bevatten.

Ranggetal

Ranggetallen zijn ordinale waarden, die een orde in rang aangeven.

3.2.3.5 Tekens

Tekens zijn strings van getallen, letters en andere leestekens.

Code

Codes zijn strings die een waarde, methode of beschrijving bevatten die in een afgekorte of taalonafhankelijke stijl worden weergegeven. Een code kan maar op een beperkte hoeveelheid manieren worden weergegeven. Voorbeelden zijn Nederlandse postcodes (‘slechts’ 6.760.000 mogelijkheden), telefoonnummers, bankrekeningnummers, btw-nummers, landcodes, wereld coördinaten en vele anderen.

Identificatie

Een identificatie is een string die refereert naar een unieke instantie van een object zoals genoemd door een bedrijf. Identificaties zijn codes die onbeperkt veel mogelijkheden kunnen bevatten. Voorbeelden zijn klantcodes, product ID codes, factuurnummers of huisnummers.

Naam

Namen bestaan uit een string tekst zonder specifieke regels over lengte. Deze zijn benamingen voor personen, plaatsen, dingen of concepten. Namen behoren zinvol en leesbaar te zijn voor menselijke lezers en zijn afhankelijk van taal. Dit is dan ook de reden dat getallen zelden in deze categorie

voorkomen. Namen verschillen van codes door het feit dat de werkelijke formaatregels veel minder specifiek zijn en bestaan uit reeksen characters van willekeurige lengtes. Onder deze categorie vallen onder andere persoonsnamen, bedrijfsnamen, productnamen, straatnamen, landnamen en andere waarden die woorden vormen.

Tekst

Tekst zijn strings van characters die (één of meerdere) woorden vormen.

Indicator

Een indicator is een verbijzondering van de nominale schaal met maar twee mogelijkheden; een 2-puntsschaal. Hier onder vallen alle velden die maar twee mogelijke waarden bevatten, of deze getallen of strings characters zijn. Bijvoorbeeld: In het geval dat het veld aangeeft of een klant wel of geen telefoonnummer heeft kunnen de waarden op deze verschillende manieren worden weergegeven: {0,1}, {ja,nee} of {wel telefoonnummer, geen telefoonnummer}.

3.2.3.6 Bits

Een bit is een symbool dat twee waarden kan aannemen. Deze waarden kunnen als een nul of één worden weergegeven. Deze categorie bestaat uit series van bits die samen objecten kunnen vormen als ze door het juiste systeem worden ingelezen. Binnen de context van de datakwaliteit tool zal deze groep zelden of nooit van toepassing zijn. Dit is dan ook de reden dat er geen tests gemaakt zullen worden die per dimensie fouten zoeken in deze velden. Alle digitale bestanden bestaan uit bytes, die op hun beurt bestaan uit bits.

Binair Object

Binaire Objecten zijn bestanden die geen grafieken, foto's, geluiden of video's zijn. Deze zijn Word documenten, Pdf's en andere documenten.

Grafiek

Grafieken zijn diagrammen, grafische representaties en andere wiskundige representaties in de vorm van een bestand.

Foto

Foto's zijn visuele beelden van personen, plaatsen of scènes in de vorm van een bestand.

Geluid

Geluiden zijn geluidbestanden zoals geluidopnames in de vorm van een bestand.

Video

Video's zijn beelden die opgenomen, geproduceerd of uitgezonden zijn als video's in de vorm van een bestand.

3.3 Data Quality tools

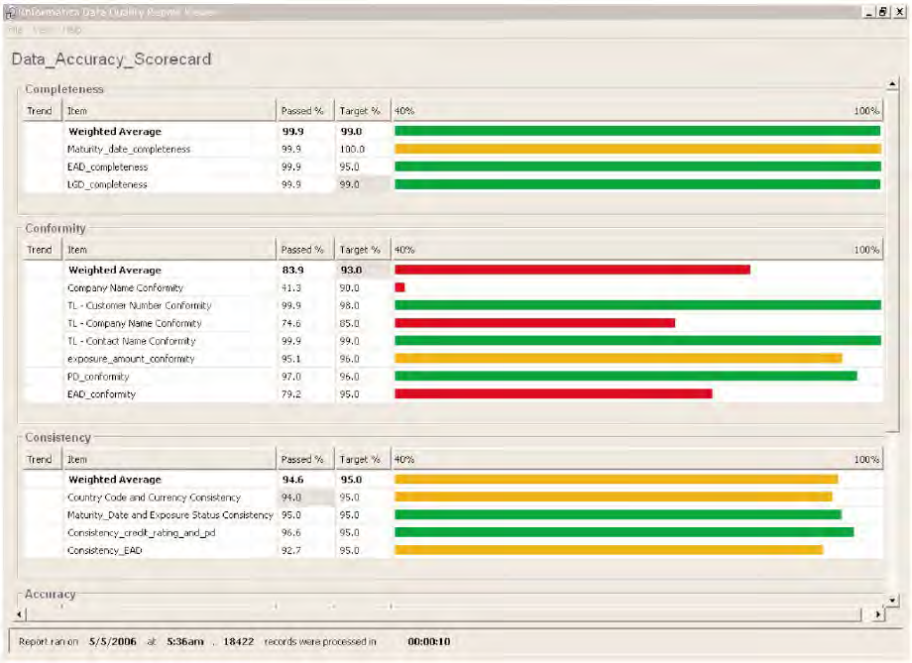
De SAS Data Quality Tool zal gebruikt worden om data assessment uit te voeren. Er bestaan ook andere tools die ondersteuning bieden bij data assessment en data cleansing. We gaan het alleen hebben over DataFlux, omdat deze de meeste functionaliteit aanbiedt en binnen SAS opgeroepen kan worden.

3.3.1 DataFlux

Vertrouwelijke Informatie

3.4 Data Quality Scorecard

De Data Quality Scorecard is een bekende manier om visueel datakwaliteit door middel van scores uit te drukken. Als een database door de data assessment fase heen is en men het aantal fouten per veld en per dimensie weet kunnen deze gevisualiseerd worden. Per dimensie kan er een overzicht komen van de relevante velden met de percentages van correcte waarden per veld. Per dimensie kan er ook een gecombineerde dimensiescore worden gevisualiseerd. Deze tien dimensiescores worden dan gecombineerd in een complete datakwaliteit score die bovenaan in de scorecard komt te staan. In Figuur 4 staat er een voorbeeld van een data quality scorecard[5]. Figuur 5 geeft een andere manier weer om datakwaliteit te visualiseren[11].



Figuur 2: Een data quality scorecard van de Informatica Data Quality Report Viewer

Data element	Relevance	Accuracy	Integrity	Consistency	Duplicate	Headline findings
1.1 Vendor Name	●	●	●	●	●	300 active names identified in SAP with no corresponding e-Procurement name. 10% Vendor names duplicated in e-Procurement (See R1.1)
1.2 Vendor Payment Terms	●	●	●	●	●	22% of vendor payment terms not in line with company policy in SAP (See R 1.2)
1.3 Vendor Bank Details	●	●	●	●	●	6% of vendors flagged as electronic settlement have no corresponding bank details (See R 1.3)
1.4 Vendor Active Flag	●	●	●	●	●	12% of vendors identified as inactive in SAP remain active in e-Procurement (See R 1.4)

Figuur 3: Een Data Quality Assessment Result Summary van Deloitte Zuid Afrika

3.5 Formules

De complete datakwaliteit score zal berekend worden aan de hand van een formule die de tien dimensiescores aggregeert in een enkele score. Deze tien scores zijn afhankelijk van de mate waarin ze zich voldoen aan de specifieke criteria voor datakwaliteit. Dit hoofdstuk bevat verschillende formules voor het berekenen van de dimensiescores en de totale score. Aan de hand van testdatabases zullen de meest geschikte formules gekozen worden.

3.5.1 Dimensiescores

Arkady Maydanchik[12] noemt in zijn boek Data Quality Assessment[13] drie verschillende soorten metingen bij het scoren van datakwaliteit: measurable records (de meetbare waarden), erroneous records (de foutieve waarden) en total records (alle waarden). De measurable records bestaan uit alle waarden die gemeten kunnen worden en niet leeg zijn. Erroneous records zijn alleen de meetbare waarden die niet voldoen aan de gezochte criteria. Total records zijn de meetbare waarden en lege waarden. De drie verschillende scores worden onderscheiden door het feit dat ze bij verschillende dimensies gebruikt moeten worden. De drie formules worden in Tabel 1 toegelicht. Maydanchik noemt één formule completeness . Deze wordt gebruikt om de dimensie volledigheid te meten. Met accuracy worden de dimensies gemeten waarbij lege waarden niet als fout worden beschouwen (in dit geval: de overige negen dimensies). Overall kan gebruikt worden wanneer men lege waarden wel als fout wil beschouwen. Als bijvoorbeeld een veld tien waarden bevat, waarvan één leeg is en van de negen gevulde waarden drie niet voldoen aan een specifieke dimensie, dan is completeness gelijk aan 9 (90%), accuracy gelijk aan 6,6 (66%) en overall gelijk aan 6 (60%).

Overall	$\frac{\text{Total records} - \text{Erroneous records}}{\text{Total records}}$
Completeness	$\frac{\text{Measurable records} - \text{Erroneous records}}{\text{Measurable records}}$
Accuracy	$\frac{\text{Total records} - \text{Erroneous records}}{\text{Total records}}$

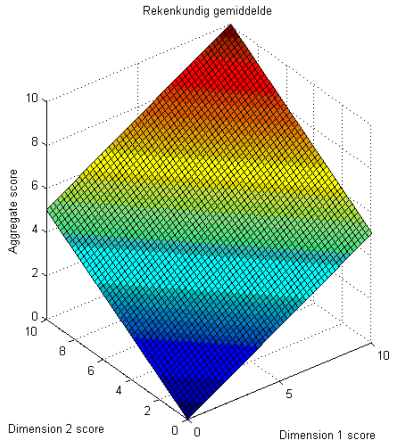
Tabel 1: De drie formules die Arkady Maydanchik noemt.

Voorbeeld: Als het model een dataveld moet testen op data formaat dan zijn er drie waarden mogelijk: waarden met het juiste formaat, waarden met een verkeerd formaat en lege waarden. Completeness refereert in dit geval naar de volledigheidscore, overall naar de data formaat score van het hele veld en accuracy naar de data formaat score van de volledige waarden.

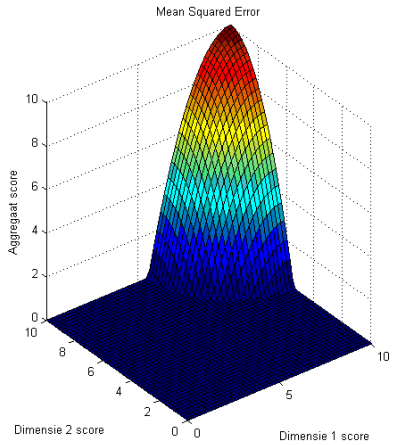
3.5.2 Complete datakwaliteit score

Dit deel verzamelt negen verschillende aggregaat scores die gebruikt kunnen worden om de tien dimensiescores te combineren. De formules worden weergegeven, samen met een toelichting van

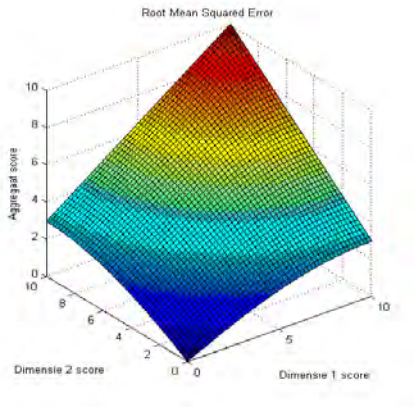
de formule en een grafiek die een indicatie geeft van de relaties tussen de dimensiescores en de resulterende aggregaat score. Deze formules worden in de grafieken uitgebeeld in een driedimensionale grafiek waarin de X en Y assen refereren naar de scores van twee verschillende dimensies en de Z as refereert naar de aggregaat score. Twee dimensies zijn gekozen in plaats van meer, zodat de relaties tussen de dimensiescores visueel kunnen worden weergegeven.



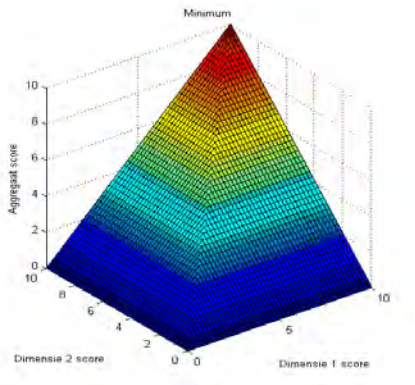
Rekenkundig gemiddelde – Dit is de meest eenvoudige formule. Deze formule berekent het gemiddelde van alle tien dimensiescores. Het kan de indruk geven dat de score hoog is terwijl het mogelijk is dat een aantal dimensiescores juist heel laag en de overige dimensiescores hoog genoeg zijn om hiervoor te compenseren. Als alle dimensies even belangrijk zijn geeft deze formule de meest duidelijke resultaten



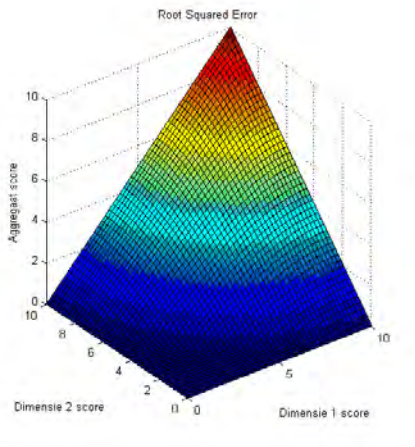
Mean Squared Error – De MSE kwadrateert de afwijking tot de maximum score en telt deze op. Doordat de afwijking wordt gekwadrateerd kan een enkele lage dimensiescore de complete score meer tegenwerken dan meerdere middelmatige dimensiescores. Deze formule kan gebruikt worden als het belangrijk genoeg is om de scores hoog te houden dat de aggregaat score exponentieel daalt bij een aantal fouten.



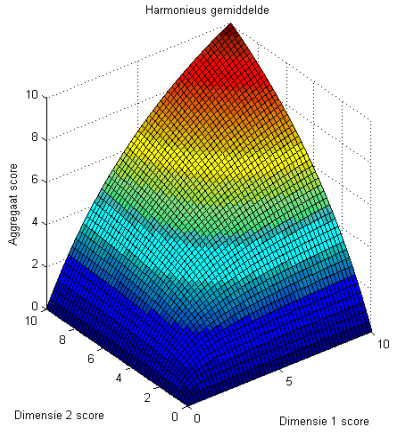
Root Mean Squared Error – Deze formule neemt de wortel van de MSE nadat de individuele scores zijn gecombineerd. Deze formule daalt sneller dan het gemiddelde, maar produceert pas een absolute nul als alle scores nul zijn.



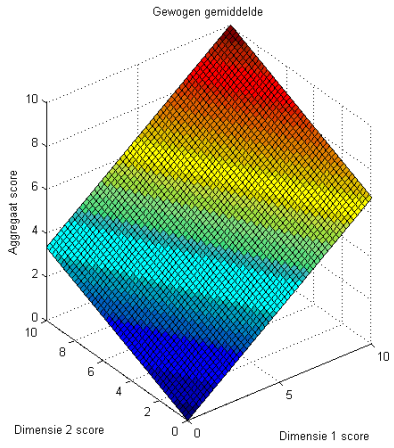
Minimum – Het minimum neemt altijd de laagste dimensiescore als datakwaliteit score. Dit zorgt ervoor dat alle aandacht wordt besteed aan de slechtste dimensie. Deze formule geeft meteen door waar het grootste probleem ligt.



Root Squared Error - Deze formule neemt net als de RMSE de wortel van de MSE nadat de individuele scores zijn gecombineerd, maar deelt de som niet door tien. Deze formule produceert lagere resultaten dan het minimum als er in andere dimensies ook fouten worden gevonden.



Harmonieus gemiddelde – Het harmonieus gemiddelde neemt de inverse van de dimensiescores en produceert hiermee een complete score die sneller daalt dan het gemiddelde. Deze formule is niet intuïtief te begrijpen en significant verschillend om aanbevolen te worden.



Gewogen gemiddelde – Deze formule kan gebruikt worden als het blijkt dat bepaalde dimensies belangrijker zijn dan anderen en zwaarder moeten meetellen in de totale datakwaliteit score. Het gewogen gemiddelde nemen is onnodig in het geval dat alle dimensies even belangrijk zijn, maar als er sprake is van een grotere risico bij bepaalde dimensies is deze de enige formule die aangeboden kan worden. Als gevolg van een uitgebreide risicoanalyse, zoals in 2.3.1 wordt toegelicht is dit het geval.

Hoogste 50% getrimd gemiddelde - Hier wordt de datakwaliteit score berekend aan de hand van de vijf laagste dimensiescores. Het gemiddelde wordt genomen van vijf minima. Deze formule heeft geen grafiek, omdat het weergeven met twee dimensies gelijk zou zijn aan het minimum.

Uit deze acht formules zal de beste gekozen moeten worden om te gebruiken bij het kwantificeren. Doordat de aggregaat score moet worden berekend aan de hand van tien verschillende dimensies en hoe deze met elkaar omgaan kunnen het minimum, MSE, RSE, RMSE en harmonieus gemiddelde niet worden aanbevolen. Deze vijf formules worden snel beïnvloed door individuele uitbijters. Het rekenkundig gemiddelde en gewogen gemiddelde zijn de meest duidelijke formules. Het gemiddelde geeft het verband tussen alle tien scores weer en deze is snel te begrijpen. Het gewogen gemiddelde wordt alleen significant beïnvloedt door een enkele dimensie als deze dimensie een hoog risico loopt. Op deze manier zijn het geen willekeurige individuele uitbijters die de aggregaat score beïnvloedden. Dit onderzoek gaat over het kwantificeren van data binnen de data cleansing fase, waarin er nog geen conclusies getrokken kunnen worden over de gewichten van de dimensies. Als de risico bepalen fase verder wordt onderzocht kunnen deze resultaten wel gebruikt worden om gewichten op te stellen. Voor deze reden zal dit onderzoek verder gaan met het rekenkundig gemiddelde als aggregaat score met de mogelijkheid om deze te vervangen met het gewogen gemiddelde nadat de risico's worden bepaald.

4 Model

Dit hoofdstuk bevat het model dat werd gebruikt tijdens dit onderzoek. Het eerste deel geeft een overzicht van de verschillende dimensies die zijn gevonden en hoe deze kunnen worden getest. Het tweede deel licht toe hoe het traject van datakwantificeren loopt en welke rol de tool hierin speelt. Het derde deel wijst hoe de template werkt en welke rol deze speelt. Het vierde deel laat zien welke stappen nodig zijn voordat de tool ingezet kan worden. Het laatste deel licht toe hoe de verschillende dimensiescores elkaar beïnvloeden.

4.1 Analyse van de dimensies en de datavelden

Voordat het model kan worden opgesteld zal er een overzicht komen van de relaties tussen de dimensies en mogelijke datavelden. In dit deel worden de dimensies toegelicht die in hoofdstuk 3 zijn opgenoemd. De eerste paragraaf bevat de dimensies die relevant zullen zijn voor het model en de tweede zal een overzicht geven van de dimensies die niet relevant zijn.

4.1.1 Het verband tussen de dimensies en de datavelden

Vertrouwelijke Informatie

4.1.2 Niet relevante dimensies

Vertrouwelijke Informatie

4.1.3 Dimensie tests

Vertrouwelijke Informatie

4.2 Het kwantificeringstraject

Vertrouwelijke Informatie

4.3 Model van template

Vertrouwelijke Informatie

4.4 Pre-processing

Vertrouwelijke Informatie

4.5 Scores

Vertrouwelijke Informatie

5 Template en vragenlijst

Vertrouwelijke Informatie

5.1 Template

Vertrouwelijke Informatie

5.2 Vragenlijst

Vertrouwelijke Informatie

6 Formule Aggregaat Score

Vertrouwelijke Informatie

6.1 De drie Maydanchik scores

Vertrouwelijke Informatie

6.2 Waarden die voldoen aan alle criteria

Vertrouwelijke Informatie

6.3 De beste formule per situatie

Vertrouwelijke Informatie

7 Gebruikte Data

Vertrouwelijke Informatie

8 SAS Data Quality Tool

In dit hoofdstuk wordt er toegelicht wat de tool, dat voor deze stage is gebouwd, kan doen en hoe deze is geprogrammeerd.

8.1 Programmeerstappen

Vertrouwelijke Informatie

8.2 Functionaliteit van de Tool

Vertrouwelijke Informatie

9 Data Quality Scorecard

Vertrouwelijke Informatie

10 Resultaten

Vertrouwelijke Informatie

11 Conclusies en aanbevelingen

Dit hoofdstuk bevat de conclusies die getrokken kunnen geworden als resultaat van dit onderzoek. De eerste paragraaf bevat een samenvatting van het hele datakwaliteit traject en een indicatie van de beste manier om de eindscores weer te geven. De tweede paragraaf bevat aanbevelingen voor vervolgonderzoeken en een stappenplan om deze methodes te implementeren in een realistische bedrijfssituatie.

11.1 Conclusies

Vertrouwelijke Informatie

11.2 Aanbevelingen

Vertrouwelijke Informatie

12 Bijlagen

Syntax voor de reguliere expressies voor data formaat***Vertrouwelijke Informatie***

SAS Code

Vertrouwelijke Informatie

13 Literatuurlijst

Data Quality algemeen

[1] The High Cost of Low-Quality Data – Larry P. English
http://www.information-management.com/issues/19980101/771-1.html

[2] The costs of poor data quality – Anders Haug e.a., University of Southern Denmark, Journal of Industrial Engineering and Management (2010)

[8] Data Quality: Making data fit for purpose – Deloitte UK (2007)

Tien criteria

[3] Data Quality Assessment & Cleaning – Nicolaas Nobel, Deloitte NL (2011)

[4] 10 Criteria voor datakwalitiet (presentatie) – Deloitte NL (2010)

Data Quality scorecard

[5] Building a Data Quality Scorecard for operational data governance – David Loshin, DataFlux (?)

[11] Deloitte's Data Quality assessment: Better data, better business – Deloitte ZA (2009)

[12] How to Create a Data Quality Scorecard – Arkady Maydanchik

Data Quality Scoring

[6] How to Measure and Monitor the Quality of Master Data – Thomas Ravn
http://www.information-management.com/issues/2007_58/master_data_management_mdm_quality-10015358-1.html

[7] How to Measure Data Quality: Metrics And Scorecards – Tom Breur (07-2010)
http://www.xlntconsulting.com/resources/How%20to%20measure%20data%20quality%20metrics%20and%20scorecards.html

[13] Data Quality Assessment (2007) – Arkady Maydanchik

Datavelden

[9] KWALITATIEF EN KWANTITATIEF ONDERZOEK
http://foldoc.org/data+type
http://en.wikipedia.org/wiki/Data_type
http://en.wikipedia.org/wiki/Code
http://nl.wikipedia.org/wiki/Nominale_schaal

[10] UN/CEFACT Core Components Data Type Catalogue Version 3.0 (29-09-2009)
www.unece.org/cefact/codesfortrade/CCTS-CatalogueVersion3.pdf

Monitoring Data Quality Performance Using Data Quality Metrics – David Loshin (11-2006)

Simple Data Quality Scoring with SSDQS & SSIS

http://www.bimonkey.com/2011/10/simple-data-quality-scoring-with-ssdqs-ssis/

Data Quality Measurement and Assessment – Howard Veregin (1998)

Field Code Changed

Field Code Changed

Field Code Changed

Field Code Changed

http://www.ncgia.ucsb.edu/giscc/units/u100/u100_f.html

Andere datakwaliteit dimensies

AN EVALUATION FRAMEWORK FOR DATA QUALITY TOOLS (Practice Oriented) - Virginie Goasdoué

Total Quality Management Blueprint - Dale B, Bunney H (1999)

Assessment Methods for Information Quality Criteria - Felix Naumann1 Claudia Rolker

Datakwaliteit vaak struikelblok voor klantgericht communiceren - Frans Plat (2009)

Datakwaliteit en gegevensbeheer - Leendert Hinds (2011)

Bestuurlijke informatiesystemen en automatisering - Prof Dr. T.M.A. Bemelmans

Datakwaliteit en klantgerichte marketing: De basis voor gericht klantcontact - Gert-Jan Bruinsma / John Oosting (T-Systems Nederland B.V.)

Deloitte

Data Quality services – Deloitte (2000)

Data Quality and integrity solutions – Enterprise Risk Services, Deloitte BE (2006)

Data Quality and integrity solutions – Deloitte Enterprise Risk Services NL

Our detailed approach: Data Quality assessment – Stages overview

Data Quality management: Our perspective – Deloitte Consulting LLP (07-2009)

Data kwaliteit (van bijvoorbeeld: stamgegevens) – Ating Temmar, Deloitte NL

Regulatory capital efficiency: RWA optimisation – Deloitte NL (09-2010)

Data Quality framework – Justice sector information strategy, NZ Ministry of Justice (06-2008)

Powerpoints

Analyse technieken voor Assesment

Nummers en formaten

Tools (DataFlux info)

Datakwaliteit in uw financieel systeem: Assessment & cleansing

Problemen met Datakwaliteit: Voorbeelden

Dataconversie en Datakwaliteit