Gaat de machine de grillige mens eindelijk snappen?

Leer de machine de taal

Leendert Paape en Bas Dudink

In datawarehouse- en datamigratieprojecten wordt de laatste jaren steeds vaker aandacht besteed aan het aspect datakwaliteit. Toch wordt het meer gezien als een noodzakelijke randvoorwaarde dan als een doel op zich. Hierdoor ontstaat het risico van onvoldoende prioriteit en te late reactie op incidenten, hetgeen nadelige gevolgen heeft voor de doorlooptijd van het project en de effectiviteit van de uiteindelijke oplossing.

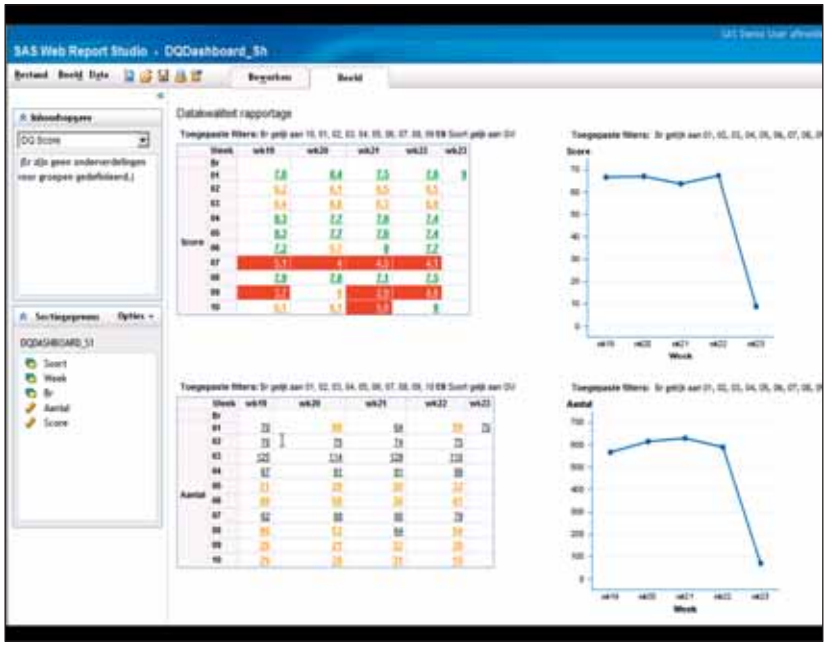
Door de opdrachtgevers en gebruikers in de business inzicht te bieden in de consequenties van de huidige status van de datakwaliteit, verschuift de prioriteit en is de weg open voor het realiseren van de juiste en tijdige maatregelen.
"Hoe komt het dat we in de dit jaar opgeleverde nieuwbouwwijk slechts 480 nieuwe klanten hebben terwijl we 493 aansluitingen hebben aangelegd en deze klanten in het eerste jaar nog niet van energieleverancier konden wisselen?" en "waarom is het fabricageproces gisteren ten onrechte stop gezet en zijn we 50.000 euro omzet misgelopen?" zijn twee voorbeelden van vragen die door het management worden gesteld en direct gerelateerd kunnen worden aan het ontbreken van voldoende aandacht voor het datakwaliteitsvraagstuk.
Op dit moment worstelen veel organisaties in datawarehouse- of datamigratieprojecten met datakwaliteitsproblemen. Toch is het niet zo dat deze problemen duidelijk zichtbaar gemaakt en openlijk geuit worden. De problemen worden in ieder geval binnenskamers gehouden. Er wordt verwezen naar zaken die in het verleden gebeurd zijn. Er is sprake van een zekere mate van schaamte om te bekennen dat men moeite heeft om systemen te migreren vanwege datakwaliteitsproblemen.

Dit gedrag was ook zichtbaar in de beginperiode van Business Intelligence op het moment dat het management van organisaties inzicht kreeg in de prestaties van hun afdelingen. Het was natuurlijk prachtig dat er een rapport was dat de benodigde cijfers op een eenvoudige en flexibele manier boven water haalde, maar een lade of het bekende cilindrisch archief was toch vaak de uiteindelijke plaats van bestemming van het rapport. In het meest gunstige geval werden de resultaten binnen de afdeling zelf besproken.
Dergelijk gedrag zien we nu niet meer. Afdelingmanagers worden door Key Performance Indicators (KPI's) aangestuurd en vergeleken met andere afdelingen in hun organisatie. Op directieniveau wordt dit zelfs tussen organisaties gedaan (benchmarks). Iedereen rapporteert zijn eigen functioneren. Intern, maar tegenwoordig ook steeds meer extern, aangezien dit wordt opgelegd door formele wetgeving, branche-organisaties of mededingingsorganen.
Die volwassenheid hebben we nog niet bereikt ten aanzien van datakwaliteit. Hoe bereiken we dat en in hoeverre kan de huidige technologie ons hierin ondersteunen?

Verkrijgen van inzicht

Als je goed kijkt is er niets spectaculairs nodig. Wat voor het functioneren van een afdeling toepasbaar is, kan natuurlijk ook bij datakwaliteit uitgevoerd worden. Verkrijg Inzicht! Of zoals vrijwel wekelijks door een collega naar voren wordt gebracht: "meten is weten". En niets is minder waar.

De belangrijke eerste vraag is: wat is goede datakwaliteit? En als dat bekend is: hoe wordt gemeten of de kwaliteit van data goed is? Dat zijn terechte vragen en dit artikel heeft niet de intentie een wetenschappelijke beschouwing te starten. Het moet praktisch zijn, uitvoerbaar en het liefst snel resultaat opleveren. De druk is namelijk groot en als er onvoldoende waarde wordt geleverd, zijn we weer terug bij af en is datakwaliteit weer gewoon een randvoorwaarde in plaats van een harde doelstelling. De methodiek die wij hanteren bouwt op dezelfde leercurve, zoals die ook voor Business Intelligence is doorlopen. Stel eerst het kwaliteitsniveau vast. Wanneer is iets goed? En hoe bepaal je dat? Vervolgens maak je inzichtelijk wat de gevolgen zijn van het niet voldoen aan het vastgesteld datakwaliteitsniveau door dat te vertalen naar voor management identificeerbare en kwantitatieve gegevens. Enkele voorbeelden:
- bij een energiedistributieleverancier zorgt een discrepantie



Afbeelding 1: Dashboard SAS DQ tool.

tussen aansluitgegevens en contractgegevens voor een potentieel verlies in omzet.
- in de procesindustrie kan het ontbreken van data of het verkrijgen van incorrecte data leiden tot verkeerde conclusies ten aanzien van noodzakelijk onderhoud, met als gevolg het stilleggen van het productieproces en daarmee het missen van omzet of maken van onnodige kosten.
- Het niet kunnen identificeren van potentiële dubbele registraties bij landelijke registratiesystemen kan leiden tot fraude in de vorm van belastingontduiking of onterechte uitkeringen.
- het incorrect verwerken van dagelijkse batchprocessen naar het datawarehouse kan leiden tot verkeerde managementinformatie en daarmee verkeerde beleidsbeslissingen.

Business rule

Al deze zaken kunnen aan de ene kant worden geïdentificeerd door het definiëren en monitoren van datakwaliteit business rules en aan de andere kant visueel worden gepresenteerd.
'Business rule' is een veel mis- en gebruikte term. Laten we duidelijk maken wat wij er onder verstaan. In onze aanpak wordt een datakwaliteit business rule gedefinieerd met behulp van een aantal componenten. Hierbij komen de volgende begrippen aan bod: Field, Rule, Event, Task en Process, zie afbeelding 2.
Om deze begrippen nader toe te lichten zullen we een voorbeeld hanteren. We willen de datakwaliteit van het datawarehouse monitoren door te kijken naar twee key indicatoren: het aantal orders; de gemiddelde omzet per product.
Vervolgens worden de uitkomsten van deze key indicatoren uit een door het bronsysteem aangeleverd bestand met de verwerkte resultaten in het datawarehouse vergeleken. Daarnaast wordt gecontroleerd of er ook afwijkingen zijn boven een gedefinieerde marge ten opzichte van de afgelopen twee weken. Bij de eerste analyse dient er naast het vastleggen van het resultaat ook een vervolgproces gestart te worden. Bij de tweede analyse dient de betrokken *data steward* een e-mail te ontvangen.
De eerste check (input is output) komen we in de praktijk nog wel geregeld tegen. De tweede veel minder, terwijl die juist inzicht geeft in de trendmatige ontwikkelingen. Een ander voorbeeld van die laatste check is het monitoren van lege of incorrecte velden in aangeleverde bestanden in de loop van de tijd.
Wanneer moet ik aan de bel trekken bij mijn dataleverancier?
Terug naar de business rule componenten. We definiëren twee velden: Aantal_Orders en Gem_Omzet_Product. Gewoon logisch definiëren, beschrijving erbij, maar nog geen enkele relatie met een attribuut in een fysiek bestand. Daarna worden de regels gedefinieerd. Wanneer in het business rule monitoring proces een regel wordt uitgevoerd kan dat op drie verschillende niveaus:
- Record – voor elk record wordt de business rule uitgevoerd.
- Group – voor elke groep wordt een business rule uitgevoerd. In ons voorbeeld is het vaststellen van de gemiddelde omzet per product een dergelijke regel.
- Set – deze business rule wordt slechts uitgevoerd voor de hele set, bijvoorbeeld een bestand. Het bepalen van het totaal aantal orders is een voorbeeld van een dergelijke regel.

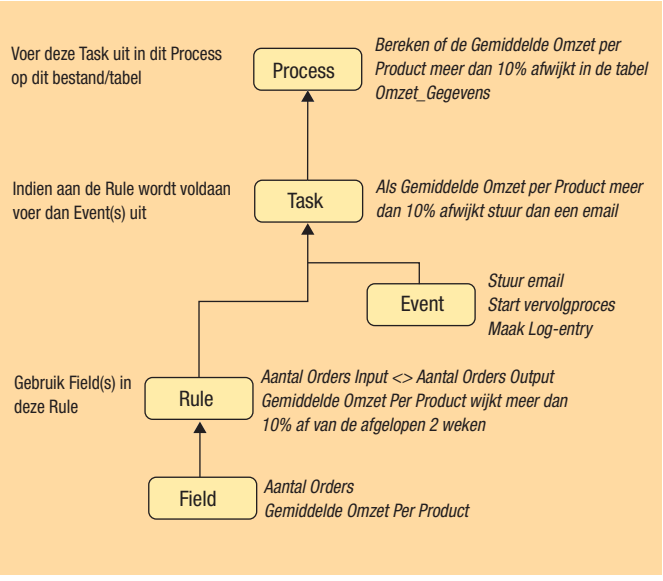
We definiëren in eerste instantie onze regels voor de gewenste check's. Hierbij kunnen we gebruik maken van de aanwezige functies die refereren aan historische, door de business rule monitor geproduceerde getallen.

De volgende stap is het definiëren van een Task, (afbeelding 2) waarbij aan het uitvoeren van een regel één of meerdere events worden gekoppeld. In ons voorbeeld zijn het maken van een log entry, het starten van een ander proces en het versturen van een e-mail voorbeelden van een event.

Uiteindelijk kunnen we een proces voor datamonitoring definiëren waarbij de vastgelegde taak wordt uitgevoerd op een bepaald bestand. Afbeelding 3 laat dit zien.
Op dit moment wordt ook een koppeling gemaakt tussen het logische veld, zoals we dat in de business rule hebben vastgesteld, en een fysiek attribuut in een file of tabel. Hiermee is het totale proces vastgelegd en kan het met behulp van een scheduler op regelmatige basis worden uitgevoerd. De resultaten kunnen door standaard dashboard-functionaliteit of via eigen rapportages worden gepresenteerd.
Zijn we er dan? Nee, we zijn weliswaar op de goede weg en hebben in ieder geval inzicht gekregen in wat de gevolgen zijn van de huidige status van datakwaliteit. Maar waar zit precies het probleem en hoe kunnen we het preventief oplossen?

Oplossen

Na enige tijd te hebben gemeten wordt het duidelijk waar de problemen van het informatieraffinageproces zich bevinden. In de regel wordt een prioriteitenlijst opgesteld van de zaken die als eerste opgelost dienen te worden. Dan zijn er diverse mogelijkheden om de problemen op te ruimen. Ten eerste is er de 'zoek-het-lek-strategie'. Je zoekt naar de bron van de problemen en kijkt of daar iets aan te doen is. Nu kan het vinden van het lek al uitermate ingewikkeld zijn, maar met behulp van de business rule metingen moet duidelijk worden op welke plekken in de processen de problemen ontstaan.
De tweede strategie is de 'dweil-strategie'. Nu willen we liever niet dweilen met de kraan open, maar soms ontkom je er niet aan. Wijzigingsvoorstellen kunnen namelijk een gigantische implicatie hebben voor de operatie van een organisatie. Als er al tientallen jaren gewerkt is aan het automatiseren van de bedrijfsprocessen is het niet eenvoudig om deze processen beter op elkaar af te stemmen, en over deze bedrijfsprocessen betere en betrouwbare informatie te krijgen.



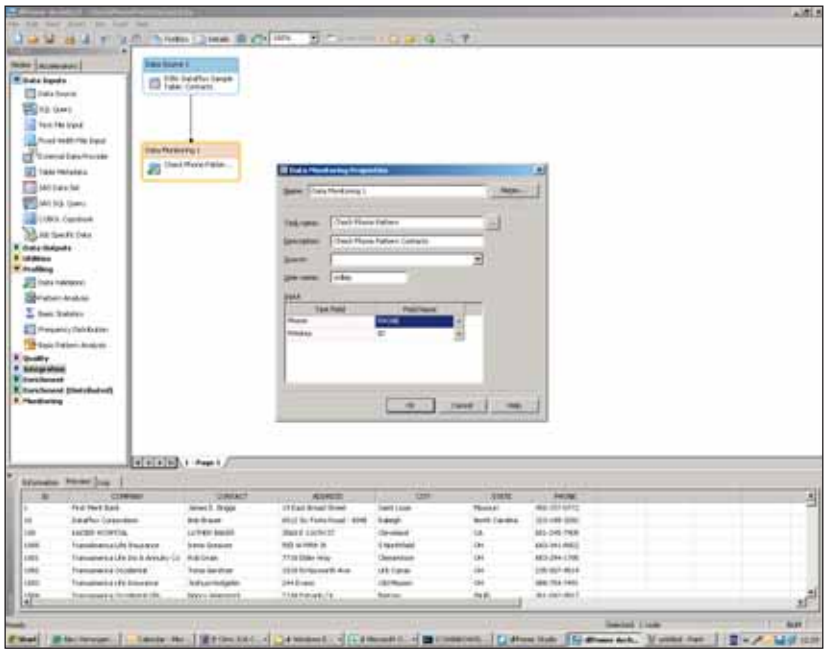
Afbeelding 2: Componenten.

Wat maakt de rol van gespecialiseerde software voor datakwaliteit nu noodzakelijk in deze gevallen? Waarom kan dit niet al met reguliere BI- of ETL-tools?
Het is fascinerend dat we met onze datakwaliteitsoplossing voor het eerst een poging doen om een machine werkelijk taal te laten begrijpen. Tijdens onze schooltijd moesten we om de Engelse taal te leren elke dag een bladzijde met woorden leren. Destijds absoluut niet leuk, maar nu is deze kennis onmisbaar in het dagelijks werk.
De enige manier om in de huidige wereld een computer beter te laten samenwerken met een mens is door deze onze taal bij te brengen. Dus het moet beschikken over woordenkennis.
Dit betekent duizenden en duizenden woorden als je de Nederlandse taal wilt herkennen en begrijpen. Maar met spelling alleen zijn we er niet, ook grammaticaregels worden aangeleerd, omdat de volgorde van woorden een tekst een volkomen andere betekenis kan geven. En aangezien de taal niet alleen wordt geschreven, maar ook auditief wordt overgebracht en ook weer wordt omgezet in geschreven taal, is het heel belangrijk om de machine fonetische regels te leren.

We moeten hiermee niet denken dat een machine de Nederlandse taal nu volkomen machtig is, maar het is verrassend wat met deze kennis door een programma gedaan kan worden. Laten we naar een aantal voorbeelden kijken:
London
Eur/UK/Lon
De hoofdstad van Groot-Brittannië.

In dit voorbeeld willen we een aggregatie laten maken op hoofdsteden. Echter, er bestaat een dusdanige variatie dat simpel aggregeren niet de juiste getallen oplevert. Als mens zien we direct dat dit drie variaties zijn van Londen. Maar wat voor kennis gebruiken we hierbij?
Ten aanzien van het eerste woord kunnen we op twee manieren een conclusie trekken: namelijk dat het erg lijkt op Londen. Het klinkt erg gelijk en er is slechts 1 letter verschil met een bekende hoofdstad. Daarnaast is London een Engelstalige variatie op de hoofdstad die we zoeken.
In het tweede geval zien we als mens vrij eenvoudig dat het hier gaat om een hiërarchische beschrijving van de hoofdstad. In Europa ligt het Verenigd Koninkrijk en de hoofdstad is Londen. Ook de machine kan herkennen dat 'Lon' een afgekorte versie is van een bekende hoofdstad.
In het laatste geval is er geen afkorting of indicatie dat het om Londen gaat. Wij hebben geleerd dat Londen de hoofdstad is, de machine zal dit echter oplossen door een verrijking toe te passen. Immers met aanvullende kennis weten we dat Londen de hoofdstad van Groot-Brittannië is.

Een volgend voorbeeld:
F.G.
Freek Gerrit



Afbeelding 3: Datamonitoring.

FGM
JAN
Hier worden diverse variaties in het gebruik van voornamen weergegeven. Volledig uitgeschreven namen en afkortingen worden door elkaar gebruikt. Afkortingen zijn te herstellen door punten op de juiste plek toe te voegen. Dit gaat helaas niet goed voor Freek Gerrit. Deze namen moeten eerst afgekort worden. Met de vergaarde kennis worden de voornamen herkend en afgekort en daarna van de juiste punctuatie voorzien.

Het volgende voorbeeld:
Freek Gerrit
Reguslaan 34
Johan
De Wit
Schone Glazenwassers V.O.F.

Hier zien we een veel voorkomend probleem. Er bevindt zich informatie in een kolom waar het niet thuis hoort, namelijk een adres tussen namen. Het is belangrijk om dit te herkennen en op de juiste plaats terug te zetten. Elke regel informatie wordt herkend en van metadata voorzien waar onder meer uit blijkt dat 'Reguslaan 34' een adresaanduiding is en hier toch echt niet thuishoort.

Het volgende voorbeeld:
PeterRdeVries@rtl4.nl
VriesdePR123@hotmail.com

Dit is interessant, aangezien er hier niet alleen kennis nodig is van betekenis van spelling en grammatica, maar ook – om het in puzzeltermen te omschrijven – dat het een doorloper betreft. Doorlopers zijn iets voor de doorgewinterde puzzelaars, er is namelijk geen scheiding tussen de woorden. Zijn we dan toch in staat om te laten zien dat er een relatie is tussen e-mail adressen? Jazeker, ondanks dat scheidingstekens ontbreken is te herkennen dat in het eerste geval het e-mailadres begint met een voornaam, vervolgd wordt door een tweede voorletter, vervolgens een tussenvoegsel heeft en ten slotte een achternaam. In het tweede geval wordt eerst een achternaam herkend met het tussenvoegsel dat erop volgt. Daarna volgen twee voorletters. De getallen 123 worden in dit geval gebruikt om het e-mailadres uniek te maken, maar heeft voor de software geen betekenis en wordt dan ook genegeerd. Als we de herkende onderdelen op een rij plaatsen, dan valt direct op dat dit wel eens dezelfde persoon kan zijn.

Tot slot

De genoemde voorbeelden hebben voornamelijk betrekking op contactinformatie, maar de open structuur van het gebruikte DQ tool beperkt zich niet alleen hiertoe. Elke gewenste informatie, bijvoorbeeld product of document is toe te passen.

De hier toegelichte technieken vormen slechts een greep uit de mogelijkheden. Samen vormen ze een unieke manier om problemen te herkennen en de oorzaak te bepalen. Dit stelt je in staat om op basis van metingen en inzicht in het dashboard procesverbeteringen door te voeren. De verbeteringen zijn soms technisch van aard als data worden gecorrigeerd of processen anders worden geïmplementeerd. Maar wellicht zijn gedragswijzigingen en het beleggen van verantwoordelijkheden vaker de sleutel tot succes.

Leendert Paape en Bas Dudink
Leendert Paape is Principal Consultant en Bas Dusink is Senior Technology Solution Consultant bij SAS.