Gestructureerde service of ad hoc proces?

SCHONEN VAN BESTANDEN

Dit artikel geeft aan de hand van een praktijkcase antwoord op de vraag hoe de kwaliteit van data kan worden vastgesteld en hoe op een gestructureerde wijze een schoningstraject kan worden vormgegeven door gebruik te maken van moderne technische hulpmiddelen.

Door Arjen de Graaf en Ortwin Verreck

Schoningsprojecten op data worden meestal gestart nadat zich in een bedrijf ernstige calamiteiten hebben voorgedaan, bijvoorbeeld met de facturatie. De ervaring leert dat schoningsprojecten in veel gevallen worden onderschat. De tijdsdruk en vooral de enorme hoeveelheid, veelal onverwachte, benodigde mutaties en de daaraan verbonden werkzaamheden zijn daar debet aan. In de praktijk worden dan veel uitzendkrachten ingehuurd om de hoeveelheid werk weg te werken. De vraag is dan ook: was dit op voorhand in te schatten, anders in te richten?

De casus: ontstaan van gegevensvervuiling

Bij de grootbank stond men aan de vooravond van een internationale implementatie van een nieuw Human Resource- of personeelssysteem. Een gestructureerde schoningsslag bleek relevant, ondanks dat alle medewerkerdossiers al in één centrale administratie waren vastgelegd. Er waren verschillende redenen voor de schoning.

Voorbereiding op de conversie. Aanvankelijk was de voorbereiding op de conversie de hoofdreden om te schonen. Het beeld bestond dat door een aantal technische tekortkomingen de gegevens soms niet ingelezen konden worden in het nieuwe personeelssysteem. Velden die verplicht waren in het doelsysteem, waren niet gevuld in het oude systeem – door het ontbreken van de benodigde invoercontroles.

Ontbreken van een vigerend beleid. De verschillende organisatieonderdelen maakten gebruik van hetzelfde systeem, waarbij de verschillende managers ieder een ander gedoogbeleid hanteerden. Er was feitelijk geen eenduidig beleid en vastgestelde beleidsregels werden dan ook met voeten getreden. Om nog maar niet te spreken van de vele beleidsveranderingen door de jaren heen. Ook de invoer van gegevens werd verschillend gedaan; zo werd het veld 'intern telefoonnummer' door de ene afdeling wel gevuld en hield een andere afdeling dit bij in Excel.

Opleiding van medewerkers. Er was de afgelopen jaren een groot verloop op de HR-afdeling geweest van administratieve medewerkers. Het waren juist deze medewerkers die met de benodigde materiekennis veel mutaties hadden doorgevoerd. Daarnaast speelde op de afdeling de werkdruk een grote (negatieve) rol. Door de werkdruk werden bijvoorbeeld vergoedingen toegekend aan bankmedewerkers, terwijl deze helemaal niet voor deze vergoeding in aanmerking hadden mogen komen.

Van verbetering 'in de drup'. Op het huidige systeem waren door de jaren heen diverse verbeteringen doorgevoerd, maar uit kwaliteitsmeting op de gegevens bleek dat tijdens het systeemonderhoud nogal eens gegevens werden verminkt. Sommige arbeidsovereenkomsten waren bijvoorbeeld niet meer aan een medewerker gekoppeld. Vrijwel op half-jaarlijkse basis waren schoningsprojecten geïnitieerd om de ergste nood te lenigen. Geautomatiseerd werden bijvoorbeeld periodiek de reiskosten woon-werkverkeer opnieuw berekend op basis van het bekende woonadres! Hierdoor ontstonden hardnekkige fouten.

Vervuiling

Gegevensvervuiling kan diverse oorzaken hebben; het schonen van bestanden is ook zelden alleen een technisch probleem. Zeker als men vervuiling ook in de toekomst wil voorkomen, dient proactief te worden omgegaan met de geconstateerde soorten van vervuiling. Om gegevens goed te kunnen schonen, is het van belang om de verschillende soorten vervuiling te onderkennen. Op basis daarvan kan een bijbehorend schoningsplan worden opgesteld. Er zijn grofweg drie soorten vervuiling te onderkennen.
Technische vervuiling is het eenvoudigst te detecteren. Het gaat daarbij altijd om vervuiling in de administratie waarvan eenvoudig is vast te stellen dat dit niet voor mag komen. Denk bijvoorbeeld aan een telefoonnummer met letters, een ongeldige datum (35-23-1001), een leeftijdsopgave van een medewerker van 140 jaar, en in deze casus ook aan een arbeidsovereenkomst zonder medewerker. Deze vervuiling kan worden gedetecteerd met technische controles, zoals domeincontroles, null-value controles, frequentiecontroles en controles op referentiële integriteit. Daarnaast zijn er ook specifieke oplossingen in de markt om technische vervuiling van specifieke NAW-bestanden te detecteren en op te lossen.
Functionele vervuiling heeft altijd betrekking op registraties die weliswaar kunnen voorkomen, maar niet mógen voorkomen. Voorbeelden hiervan zijn medewerkers die een leaseauto hebben en tegelijkertijd een OV-jaarkaart, medewerkers die een full-time jaarsalaris hebben onder het minimumloon, etcetera. Voor dit soort vervuiling worden bedrijfsregels opgesteld voor detectie. De bedrijfsregels zijn in de casus afgeleid uit de (CAO) reglementering, gesprekken met expert users en door 'rule mining' op het bestaande systeem.
Inhoudelijke vervuiling is veruit het lastigst vast te stellen. Het gaat hier om registraties die kunnen voorkomen en ook mogen voorkomen, maar die feitelijk onjuist zijn: 'uw (internet) bestelling voor deze flatscreen kost u 99 euro'. Dit soort vervuiling heeft invloed op de wijze waarop gemeten kan worden en op de mate waarin de vervuiling gecorrigeerd kan worden. Zo kan een onterechte declaratie worden teruggedraaid (functioneel), maar is het soms nog niet zo eenvoudig om de ontbrekende ingangsdatum (technisch) te achterhalen.

Drie typen metingen

Het meten van de bovengenoemde drie soorten vervuiling kan op drie manieren plaatsvinden.
De meest bekende aanpak is een *waarnemingsgerichte* aanpak. Daarbij krijgen medewerkers/controleurs een uitdraai uit het systeem en gaan ze proefondervindelijk, dus in de praktijk, kijken en vaststellen of de registratie wel klopt. Deze aanpak is arbeidsintensief, de kans op fouten is groot en de organisatie loopt een risico om haar eigen 'vuile was' bij haar klanten of medewerkers buiten te hangen. Dit laatste is veelal de reden waarom organisaties kiezen voor een referentiemeting.
Bij een *referentiemeting* wordt een ander, extern bestand gebruikt om de gegevensvervuiling vast te stellen. Hierbij is het vanzelfsprekend wel belangrijk dat men zich er vooraf van vergewist dat dit referentiebestand van een betere kwaliteit is. Zelfs met het raadplegen van authentieke bronnen blijft dit een belangrijk aandachtspunt.
Tot slot is er de mogelijkheid om een *intrinsieke* meting uit te voeren. Bij een dergelijke meting wordt de vervuiling binnen het systeem vastgesteld door bijvoorbeeld controles op verbanden (patroonherkenning) en kennisregels. Dit laatste is te vergelijken met een expert in de organisatie, die ook direct kan waarnemen dat iets inhoudelijk niet klopt. Zo zou je kunnen denken aan medewerkers die een lage salarisverhoging hebben gekregen en 'opeens' een aanmerkelijk hoger declaratiegedrag vertonen. Conclusie: technische vervuiling wordt met technische controles gedetecteerd, functionele fouten worden gedetecteerd met bedrijfsregels en een groot deel van de inhoudelijke fouten kan worden opgespoord met patroonherkenning en kennisregels.

De aanpak

Het te onderzoeken systeem in de praktijksituatie bij de grootbank werd voor de gegevenskwaliteitsmeting opgedeeld in een aantal logische deelgebieden, zoals medewerkersgegevens, contractgegevens en salarisgegevens. Per gegevensgebied werd vervolgens een volledige extractie gemaakt van het systeem ten behoeve van de technische meting. Onze praktijkervaring heeft uitgewezen dat het zinvol is een momentopname van het systeem te maken en de gegevens buiten het systeem te onderzoeken. Voor het meten gebruiken we gespecialiseerde tooling waarbij, bij voorkeur, met een volledige extractie wordt gewerkt omdat daarmee een goede patroonherkenning en ontdubbeling mogelijk wordt. De bevindingen van deze technische meting, die meestal binnen enkele dagen is afgerond, zijn nodig als voorbereiding op het vaststellen van de bedrijfsregels. In de gegevensgebieden waar veel technische uitdagingen worden geconstateerd, is het lonend om ook veel aandacht te besteden aan het opstellen van de bijhorende bedrijfsregels. De mate waarin bedrijfsregels en technische regels met elkaar verband houden c.q. correleren, bepaalt welke relevante patronen worden onderkend. Deze patronen dienen om de inhoudelijke

vervuiling op te sporen, maar ook om de oorzaak van de vervuiling te achterhalen die nodig is om de juiste schoningsaanpak op te stellen. Op de te nemen schoningsstappen wordt eveneens ingegaan.
De meting levert zo zicht op de vervuiling in termen van meer of minder verdachte gegevens. Na de meting is de omvang van de vervuiling vastgesteld en wordt de vervuiling gebruteerd: de bevindingen worden omgezet naar een bijbehorende geldwaarde en vertaald naar bedrijfsrisico's. Op basis daarvan kan worden bepaald welke vervuiling geschoond dient te worden en welke aanpak, handmatig of geautomatiseerd, het beste kan worden toegepast.
Daarnaast geeft de meting ook concrete informatie voor het (her)inrichten van de bedrijfsprocessen, de benodigde gegevenscontroles en opleidingstrajecten voor het nieuwe HR-personeelssysteem.

Schonen in zeven stappen

Nadat de gegevenskwaliteitsmeting is uitgevoerd, kan het schoningstraject worden opgestart. Belangrijk bij een gestructureerde schoningsaanpak is dat altijd zicht behouden blijft op de originele gegevens en de mate waarin en waarom deze gegevens zijn gecorrigeerd en geschoond.
Daarmee wordt niet alleen duidelijk wat het resultaat van de schoning is, maar wordt ook een juiste koppeling met andere systemen gewaarborgd: waar bijvoorbeeld een klant nog wel twee keer in geregistreerd staat. Een mogelijke aanpak hiervoor is dat het originele record wordt uitgebreid met ten minste een even groot aantal nieuwe velden met de toevoeging 'cleansed'. Daardoor krijg je dus naast het attribuut 'naam' het attribuut 'cleansed_naam'. Met de extractie, de uit de administratie gehaalde gegevens, worden de volgende zeven schoningsstappen uitgevoerd: 1. Standaardiseren; 2. Parsen; 3. Vergelijken; 4. Verrijken; 5. Matchen; 6. Bedrijfsregel cleansing; 7. Superrecord/verwijzingen.

Standaardiseren.
Het schoningstraject begint met het standaardiseren van de gegevens. Dit vereenvoudigt complexere schoningsacties, zoals bijvoorbeeld het ontdubbelen van gegevens, en zorgt tevens voor een grotere herkenbaarheid van de gegevens. Vormen van standaardisatie zijn bijvoorbeeld eenduidige opmaak van alle datumvelden (dd-mm-yyyy) en initialen van de naam die altijd worden weergegeven zonder puntjes. Maar ook dat voor een aanduiding als geslacht alleen de coderingen 'M' of 'V' worden gebruikt en niet ook nog eens '0' of '1'.

Parsen.
Onder parsen wordt hier verstaan dat een tekstveld wordt opgesplitst in meerdere velden, indien de tekst feitelijk meerdere gegevens bevat. Zo zou bijvoorbeeld een naamveld kunnen worden opgesplitst in de velden 'voornaam', 'tussenvoegsel' en 'achternaam' (en kan een aanhef 'Geachte heer' leiden tot een nieuw alternatief veld 'geslacht'). De bedrijfsmatig meest interessante vorm van parsing is gericht op bedrijfsgegevens parsing, zie ook het kader. Het goed parsen van met name tekstvelden levert een schat aan nieuwe informatie op, waarmee andere velden in een later stadium geschoond worden. In de markt zijn diverse tools beschikbaar met geavanceerde parsingmogelijkheden.

Vergelijken.
Na het parsen kan de vergelijking beginnen. Dit kan met het bronbestand zelf (intrinsiek), dat creëert namelijk de mogelijkheid om dubbele registraties te detecteren of met een betrouwbaar referentiebestand van bij voorkeur een authentieke bron. Het gebruik van een referentiebestand geeft naast de mogelijkheid voor ontdubbeling, ook de mogelijkheid om foutieve en ontbrekende gegevens aan te vullen. De vergelijking gebeurt altijd op basis van gestandaardiseerde en geparste gegevens van zowel de bron als het referentiebestand. Per attribuut worden verschillende vergelijkingsfuncties gehanteerd. Zo wordt er bijvoorbeeld voor gekozen om naast een exacte vergelijking met sofinummers, ook een fonetische vergelijking toe te passen op buitenlandse achternamen. Eventueel kan ook een waarschijnlijkheidsberekening worden toegevoegd. Bijvoorbeeld de namen 'De Graaf' en 'De Graaff' komen zo vaak voor dat een overeenkomst onwaarschijnlijk is, maar namen als 'Verreck' en 'Verrek' komen al veel meer in de richting.

Verrijken.
Tijdens het vergelijken kan worden vastgesteld of gegevens in de administratie ontbreken. Deze gegevens kunnen worden toegevoegd voor het vervolg van het schoningsproces, ofwel het gegevensbestand wordt verrijkt.

Matchen.
Het matchingsproces is erop gericht om dubbele registraties op te sporen. Daarbij worden diverse patronen gedefinieerd waarmee gedetecteerd wordt of er sprake is van een dubbele registratie. Een patroon bestaat daarbij uit een combinatie van vergelijkingen op attribuutniveau. Zo kan een persoonsgegeven worden ontdubbeld op basis van een overeenkomend fiscaalnummer of een overeenkomend mobielnummer, maar een nog beter patroon is een combinatie van beide. Het zal duidelijk zijn dat er ook patronen zijn die leiden tot 'verdachtsituaties' die alleen met een handmatige schoning kunnen worden opgelost.

Ontdubbelen.
Is eenmaal een dubbele registratie opgespoord, dan dient één van de registraties als *survivor* te worden aangemerkt, zodat vervolgens in de bronadministraties de implicaties en verwijzingen correct kunnen worden omgezet. Denk hierbij onder andere aan het feit dat de gemaakte afspraken met een medewerker wel mee moeten 'verhuizen' naar de 'survivor'-registratie.

Bedrijfsregel cleansing.
Voor de meting zijn diverse bedrijfsregels en patronen gedefinieerd; deze kunnen eveneens worden ingezet om vast te stellen welke correcties uitgevoerd dienen te worden. Zo kan bijvoorbeeld worden besloten dat bij een afwijking van het CAO-salaris, er gekeken wordt naar het dichtstbij liggende salaris dat wel aan de CAO voldoet en kan bijvoorbeeld binnen een vooraf vastgestelde tolerantiegrens alsnog het salaris worden aangepast aan de CAO.

Superrecord/verwijzingen.
Uiteindelijk wordt een superrecord gemaakt waarin de eigenschappen van diverse oorspronkelijke registraties en de diverse schoningsacties worden samengevoegd. Nadat het superrecord is gemaakt, is feitelijk de schoning afgerond (ofwel het transformatieproces). Er dient nu bepaald te worden welke aanvullende maatregelen nodig zijn om op gedegen wijze de geschoonde gegevens terug te krijgen, te 'laden' in de oorspronkelijke administratie. Dit wordt ook wel het load-proces genoemd. Naast dit load-proces zal bepaald worden welke handmatige schoningsactiviteiten vereist zijn.

Bedrijfsgegevens parsing

Stel, er zijn drie verschillende artikelen in het magazijn met elk hun eigen voorraad en elk hun eigen omschrijving.
Omschrijving artikel 1:
10 stuks metaal schroef 2" met oog.
Omschrijving artikel 2:
10 stuks zink schroef 3" met oog.
Omschrijving artikel 3:
metalen oogschroef aantal 10 en lengte 2 inch.
Deze gegevens moeten we parsen en opdelen in de volgende velden:

Aantal	Materiaal	Lengte (inch)	Type
10	metaal	2	Schroef met oog
10	zink	3	Schroef met oog
10	metaal	2	Schroef met oog

Dit parsen werkt ook als er verschillende talen in de administratie zijn opgenomen. Hierdoor wordt het niet alleen mogelijk om vervolgens tot een uniforme schrijfwijze te komen, maar tevens om vast te stellen dat artikelen wellicht ook daadwerkelijk dubbel op voorraad worden gehouden. Het schonen van de gegevens levert daarmee een besparingsmogelijkheid op.

Casus: aanpak van schoning

Het schoningsproject van het HR-personeelssysteem bij de grootbank werd opgepakt door middel van twee sporen: handmatige schoning (door de administratieve medewerkers van de bank) en geautomatiseerde schoning (met behulp van ETL-tooling).
Er waren grofweg drie redenen om zo veel als mogelijk voor geautomatiseerde schoning te kiezen. Naast de 'wens' van de grootbank dat het schoningstraject de staande organisatie niet teveel moest belasten, was er ook een beperking in het huidige systeem. Door de ingebouwde controles was het namelijk soms simpelweg niet mogelijk om de vervuiling uit het systeem te verwijderen. Deze controles waren niet te omzeilen tijdens de schoning. De derde en wellicht belangrijkste reden was dat soms een complex aan regels moest worden geïnterpreteerd om de juiste schoningsbeslissing te kunnen nemen. Uiteindelijk werden in de casus bijna 90.000 records geschoond. Slechts 6 procent daarvan werd handmatig geschoond. In totaal werd net iets meer dan de helft van de medewerkerdossiers geraakt. Doordat de schoning vooraf ging aan de uitrol van een nieuw systeem, maakte dat systeem daarmee een frisse start.
Voor de schoning is vooral gebruik gemaakt van geïntegreerde tooling. De gegevens werden uit het bestaande systeem gehaald met dezelfde techniek als tijdens de meting van de vervuiling en na een complex aan technische controles, bedrijfsregels, kennisregels en patroonherkenningen, werden de vervuilde gegevens geschoond en als geschoonde gegevens in de juiste volgorde en vorm weer aangeboden aan de basisregistratie. Een beperkte set van gegevens welke als verdacht werd aangemerkt werd vervolgens gereed gezet voor handmatige schoning. Dit gehele proces werd voorzien van diverse controlelijsten en loggings zodat ook de interne controledienst kon vaststellen of de schoning zorgvuldig was uitgevoerd.

Schonen of voorkomen?

Deze gestructureerde aanpak met behulp van geïntegreerde tooling zorgde bij de casus ook voor een aantal nevenvoordelen in het traject na de schoning. Het bleek namelijk mogelijk om, tegen zeer lage kosten, op reguliere basis een gegevenskwaliteitsmeting uit te voeren. Daarnaast heeft de grootbank 'on the fly' haar eigen kennisregels daadwerkelijk vastgelegd en hebben enkele HR-medewerkers een aanvullende opleiding gekregen. Ook is een aantal processen en werkinstructies aangepast. Doordat vervuiling in een steeds vroeger stadium, door de periodieke metingen, wordt gedetecteerd kan er sneller worden ingegrepen.
Het schoningsproces met de bijbehorende tooling is dus de trigger geweest voor het inrichten van periodieke metingen en die zijn nu een standaard service geworden, waarvan basisadministraties binnen de grootbank gebruik kunnen maken en waarmee managementinformatie wordt gegenereerd over de kwaliteit van de gegevens (en mogelijke oorzaken van vervuilingen). Mogelijk dat in de toekomst het meetsysteem voor de gegevenskwaliteit zelfs gebruikt gaat worden als service tijdens het opvoeren van nieuwe gegevens.

Arjen de Graaf en Ortwin Verreck

Arjen de Graaf (arjen.de.graaf@arvix.nl) is medeoprichter en algemeen directeur van ARVIX. Ortwin Verreck (ortwin.verreck@arvix.nl) is medeoprichter en technisch directeur van ARVIX.