Schonen van data een niet te onderschatten proces

Datakwaliteit kan nog beter

Nico Klaassen

Nadat de datakwaliteit is onderzocht, breekt paniek uit. Managers beweren dat zij hun informatiestroom en in sommige gevallen zelfs de eigen medewerkers, niet meer kunnen vertrouwen. Een vertrouwensbreuk onstaat tussen de eigenaren van data en degenen die er iets mee moet doen. Iedereen is in hoogste fase van paraatheid om met spoed alle problemen op te lossen.

Een moment van bezinning zou dan goed op zijn plaats zijn. Goed nadenken over wat de oorzaak van de problemen was (òf is) en hoe de organisatie dit structureel gaat oplossen heeft nu de hoogste prioriteit. Nadenken over een strategie voor de korte maar vooral voor de lange termijn is van belang, voor zowel de eigenaren van de data, als ook voor degene die de data en informatie gebruiken.

Op zoek naar de oorzaak

Er bestaan vele problemen ten aanzien van de kwaliteit van data; een serie oorzaken van deze problemen luidt:

- klanten of relaties die zelf meerdere accounts aanmaken via Internet;
- inconsistenties tussen NAW-gegevens binnen één organisatie, omdat data op meerdere plaatsen worden gebruikt èn beheerd;
- onervaren (nieuwe) gebruikers die data foutief invoeren;
- dubbele cliënten/relaties door samenvoegen van administraties;
- dubbele cliënten/relaties door ontbreken van eenduidige notatie ('Jansen en Co' versus 'Jansen & Co');
- onjuist invullen van gegevens omdat ze (nog) niet bekend zijn en het vergeten om het later aan te vullen of te controleren;
- inconsistenties door (deels) herstellen van databases;
- inconsistenties door fouten in programmatuur en bij het oplossen daarvan het nalaten van het herstellen van de data (feitelijk zijn de fouten in de data het resultaat van historische programmafouten);
- inconsistenties veroorzaakt door eerdere conversies;
- (functioneel) misbruik van velden.

Dit zijn allemaal heel herkenbare oorzaken, die in al dan niet ernstige mate kunnen zorgen voor een slechte datakwaliteit. Een analyse van de data en vergelijken van verschillende databronnen en onderzoek van programmatuur levert vaak het benodigde inzicht in de oorzaak van de problemen. Nadat de problemen met de data zijn geconstateerd en inzichtelijk gemaakt, wordt bepaald wat er geschoond kan worden. Dit wordt gedaan op basis van een business case waarin kosten/baten en risico's naast elkaar worden gezet. Zoals in het eerste artikel is aangegeven; de 'quick-wins' zijn belangrijk om commitment binnen de organisatie te krijgen. Maar alleen kijken naar kosten en baten is niet voldoende, ook een afweging van hoe geschoond wordt en op welk moment, kan de uiteindelijke doorslag geven in de bepaling wat te doen.

Wat te schonen

Het schonen van data is een complexe activiteit. De te maken keuzes worden onderbouwd vanuit een business case. Deze beschrijft de kosten van het schonen, de opbrengst van deze geschoonde data en de risico's die schonen met zich meebrengt. De kosten van de schoning zijn afhankelijk van het soort problemen dat is geconstateerd. Technische vervuiling is in veel situaties met query's of correctieprogrammatuur te herstellen. De kosten hiervan zijn relatief laag. Functionele vervuiling is complexer en er is een inspanning van de gebruikers nodig om waarden te corrigeren (vooral daar waar het gaat om bijvoorbeeld ontdubbelen van klantgegevens). De inspanning vanuit de organisatie is in dit geval erg hoog voor zowel uitvoering als controle. Bij de baten kan worden gekeken naar de invloed van productieverstoringen en de daadwerkelijke kostenbesparing, doordat efficiënter met de data kan worden gewerkt. Dit zijn baten die

Wat is datakwaliteit?

In een serie van drie artikelen behandelt Nico Klaassen een aantal vragen over datakwaliteit. In het eerste artikel (DB/M 7, 2004) ging het om de meetbaarheid van datakwaliteit. In dit tweede artikel staat de vraag centraal of datakwaliteit nog beter kan en hoe om te gaan met vervuiling en schoning.

Wat is datakwaliteit waard? Wat kost een kwaliteitsmeting en wat levert het op? Die vragen worden in het derde en laatste artikel behandelt over datakwaliteit-verbetering, aan de hand van een business case.

omgezet kunnen worden in geld. Er zijn echter ook baten in de sfeer van imago te definiëren. Deze laatste categorie is niet zo zeer in geld uit te drukken, maar moet wel worden meegenomen als baten van de schoningactiviteiten.
Het schonen van data is niet altijd geheel zonder risico's. Deze risico's moeten in kaart worden gebracht. Zo zou het 'ontdubbelen' van data ertoe kunnen leiden dat gegevens onjuist worden samengevoegd, omdat er vanuit wordt gegaan werd dat het (onterecht) om twee keer dezelfde personen gaat. Dit kan verstrekkende gevolgen hebben, denk bijvoorbeeld aan iemand die voor iets moet betalen dat hij niet heeft gekocht, veroorzaakt omdat hij dezelfde naam en geboortedatum heeft als iemand anders.

De quick-wins zijn belangrijk om commitment te krijgen

Naast de business case-aspecten kosten, baten en risico's moet een duidelijk beeld ontstaan van het moment van uitvoering van een schoning. Het moment van schonen is met name bepalend voor de kosten en risico's. Zo kan bij een op handen zijnde dataconversie de technische schoning worden uitgevoerd als onderdeel van de op te stellen dataconversie-regels. Een foutieve code herstellen in het huidige systeem zou kunnen leiden tot het instabiel worden van de huidige applicatie. Het nalopen op dubbele vermelding van personen in bestanden leidt in de meeste gevallen tot het samenvoegen van allerlei gerelateerde records in diverse tabellen, iets wat erg complex is (de database moet tenslotte consistent blijven). Door de vermelding van personen in de conversie te 'ontdubbelen' met een oud-persoonsnummer en een nieuw-persoonsnummer is de consistentie eenvoudiger te borgen.

Naast het uitvoeren van schoningactiviteiten als onderdeel van een dataconversie, is het ook mogelijk om in de huidige productieomgeving data te schonen. Hiervoor kan gebruik gemaakt worden van reguliere (bedrijfs)processen, of ingrijpen in de database (direct muteren met query's) . Het ingrijpen in de database is altijd aan meer risico's verbonden en wordt dan ook afgeraden, tenzij men de database en vooral haar structuur volledig kent. Het alternatief is dan om gebruik te maken van reguliere bedrijfsprocessen om de data te corrigeren. Alle validaties die bij zo'n wijziging noodzakelijk zijn, worden door de programmatuur uitgevoerd en indien ook (audit) logging beschikbaar is, kan hier eenvoudig gebruik van worden gemaakt.
Als groot nadeel van het uitvoeren van schoning met reguliere bedrijfsprocessen is dat dit vaak zeer arbeidsintensief is en dat er relatief veel tijd moet worden gestoken in het voorbereiden van de schoning. De gebruikers moeten tenslotte weten welke data ze moeten corrigeren en wat de juiste waarde zou moeten zijn.

Hoe schonen

Bij het schonen zal met beleid moeten worden omgegaan met de data. Bij het uitvoeren van query's op de productieomgeving is het belangrijk dat eerst wordt nagegaan of een query het juiste effect heeft en geen ongewenste effecten heeft op andere data. Eerst beschrijven, daarna testen (bij voorkeur op een kopie van de productie-data) en vervolgens uitvoering en bewijzen dat het goed is gegaan.

Een bewuste schoning bestaat uit een drietal componenten (de mutatiecyclus):

```
SELECT keycolumn, mutcolumn FROM table WHERE
                                              condition
SELECT keycolumn, mutcolumn FROM table WHERE NOT
                                              condition
```

De Situatie vooraf query; dit zijn twee query's die laten zien welke rijen worden geselecteerd met de muterende query en welke records *niet* worden aangepast. Niet alleen de waarden die gaan wijzigen maar ook de key-waarden.

```
UPDATE table, SET mutcolumn = "newvalue" WHERE
                                              condition
```

De muterende query; dit is de query die de data daadwerkelijk aanpast.

```
SELECT keycolumn, mutcolumn FROM table WHERE
                                              condition
SELECT keycolumn, mutcolumn FROM table WHERE NOT
                                              condition
```

Situatie achteraf query/bewijsvoering; dit zijn de query's die aantonen dat de muterende query de juiste records heeft gemuteerd. Door de correctie is het echter mogelijk dat de 'vooraf query's' en 'achteraf query's' niet direct met elkaar zijn te vergelijken.

Een voorbeeld maakt het duidelijk; het attribuut XCODE zou de waarde a, b of c zou moeten bevatten. Bij de data-analyse is ook een waarde 'e' gevonden. Als schoningmaatregel is vastgelegd dat de waarde 'e' wordt omgezet naar 'b'. Dit resulteert in de volgende overzichten.
Vooraf:

```
WHERE XCODE = 'E'              ; levert 15 rijen op
WHERE NOT (XCODE = 'E')        ; levert 1500 rijen op
```

Muterend:

```
SET XCODE = 'B' WHERE XCODE = 'E'
```

Achteraf:

```
WHERE XCODE = 'E'              ; levert 0 rijen op
WHERE NOT (XCODE = 'E')        ; levert 1515 rijen op
```

Een alternatief om toch dezelfde records te kunnen selecteren is door bijvoorbeeld de timestamp (èn of de muterende user gelijk te stellen aan bijvoorbeeld 'CORRECTIE') van de tabel ook te muteren. Als men zeker weet dat er géén andere (tussentijdse) wijzigingen zijn tijdens de uitvoering, kan dit worden gebruikt om aan te tonen dat er géén andere rijen zijn gemuteerd.

Vooraf:

```
WHERE XCODE = 'E'              ; levert 15 rijen op
WHERE NOT (XCODE = 'E')        ; levert 1500 rijen op
```

Muterend:

```
SET XCODE = 'B', TIMESTAMP = date() WHERE
                                        XCODE = 'E'
```

Achteraf:

```
WHERE XCODE = 'E'
AND TIMESTAMP
BETWEEN start AND end ; levert 15 rijen op
WHERE NOT (
XCODE = 'E'
AND TIMESTAMP
BETWEEN start AND end) ; levert 1500 rijen op
```

Een ander alternatief is het gebruik van een hulptabel met de key-waarden van de te muteren rijen. De 'vooraf query' bevat dan de 'INSERT' van key-waarden naar de hulptabel. Vervolgens worden (eventueel na controle door de gebruiker) tijdens de muterende query alleen die records aangepast waarvan de key's overeenkomen met de key's uit de hulptabel. In de 'achteraf query' wordt vervolgens ook de JOIN met de hulptabel gebruikt om aan te tonen dat er gericht is aangepast. Dit laatste alternatief is ook de beste werkwijze, echter de implementatie hiervan is ook het meest complex en het meest bewerkelijk voor de organisatie. Het voorbeeld waarbij de XCODE wordt aangepast met behulp van query's lijkt overdreven. Vaak geeft een DBA in een dergelijke situatie aan "dat wel even handmatig aan te passen". Maar dit is NIET de goede werkwijze; een fout is zo gemaakt en de problemen kunnen alleen maar groter worden. De boodschap is dan ook "doe de schoning gecontroleerd", waardoor achteraf veel problemen worden voorkomen zowel tijdens de uitvoering als ook in de bewijsvoering. Tijdens een gesprek met eigenaren van de data en auditors zal het gecontroleerde verloop van een schoning als belangrijkste punt naar voren komen, betrek deze rollen er dan ook tijdens het opzetten van de schoning direct bij. Bij het schonen met gebruikmaking van de reguliere processen is het van belang dat de gebruikers bij het doorvoeren van de wijzigingen een print maken van de situatie vooraf en achteraf en dit als bewijsvoering voorleggen aan de auditor. In de praktijk wordt dit gezien als véél werk en stelt de organisatie regelmatig voor om reguliere lijsten te gebruiken als bewijsvoering. Deze methode heeft echter als nadeel dat hierin het onderscheid tussen reguliere mutaties en mutaties als gevolg van schoning door elkaar kunnen lopen en de consequenties van de schoning niet goed kunnen worden aangetoond. Een auditor moet hierover een uitspraak doen om ervoor te zorgen dat problemen met bewijsvoering worden voorkomen.

De schoning is klaar

De enige methode om aan te tonen of men klaar is met schonen, is door het uitvoeren van een aanvullende data-analyse. De werking hiervan komt overeen met de analyse zoals aangegeven in het vorige artikel. Door dezelfde analyse uit te voeren kan ook daadwerkelijk worden aangetoond dat de kwaliteit na de schoning beter is geworden en dat er niet 'appels en peren' met elkaar worden vergeleken. Het is mogelijk dat de kwaliteit van de data na de schoning weer achteruit loopt, tenzij er maatregelen worden getroffen. De investering die is gedaan voor de schoning zou na een paar maanden tot een jaar weer teniet gedaan kunnen zijn. Om nieuwe vervuiling te voorkomen is het van belang in een vroegtijdig stadium maatregelen te treffen om vervuiling van data tegen te gaan. Het beheren en beheersen van data is een aspect dat tijdens het definiëren van de schoningstrategie al moet worden uitgewerkt. Maatregelen die ingrijpen in de programmatuur, zoals aanvullende controles en het doorvoeren daarvan, kosten tijd en moeten zijn afgerond voordat de schoning wordt uitgevoerd of is afgerond.

Bij het definiëren van correctieve maatregelen in de schoning is het beschikbaar hebben van een (corporate) datamodel een belangrijk hulpmiddel. In een (corporate) datamodel worden de definities van de data en de verantwoordelijkheid voor de data eenduidig in kaart gebracht en vastgelegd. Indien data door verschillende afdelingen en of bedrijven worden beheerd, zal het 'corporate' aspect een belangrijke rol spelen. Items als 'bedoelt iedereen hetzelfde', 'hanteert men overal dezelfde definitie en dezelfde controles' zijn bij grote organisaties een belangrijk aandachtspunt. Als data verspreid worden beheerd in een organisatie en soms om diverse redenen dubbel zijn opgeslagen, wordt het mogelijk om de eigenaren aan te spreken op hun verantwoordelijkheid om data op goede en verantwoorde wijze te beheren.

Het opstellen van een (corporate) datamodel is een complexe klus maar maakt gericht analyseren en schonen een stuk eenvoudiger. Indien de stap naar een (corporate) datamodel te groot is, moet men er voor zorgen dat de controle-eisen die men heeft gebruikt bij de analyse en schoning goed worden gedocumenteerd. Het is een eerste stap in de richting van eenduidige definities.

Het schonen van data is een proces dat men niet moet onderschatten en waarbij de juiste beslissingen op het juiste moment moeten worden genomen. Het kan niet zo maar even tussendoor gedaan worden en vergt in veel gevallen een behoorlijke inspanning. Of het schonen van data uiteindelijk ook rendabel voor een organisatie is, wordt in het derde deel van deze serie over datakwaliteit in detail uitgewerkt.

Nico Klaassen

Ing. Nico Klaassen (nico.klaassen@capgemini.com) is Managing Consultant bij Capgemini.