Een kwestie van integrale definitie

Datakwaliteit is meetbaar

Nico Klaassen

In een recent onderzoek gaf Gartner aan dat de kwaliteit van data nogal te wensen over laat. De conclusie van de boodschap was dat door een lage kwaliteit en fouten in gegevens, niet de juiste en soms zelfs verkeerde beslissingen worden genomen.

De resultaten van het onderzoek stemmen menig manager waarschijnlijk niet erg gelukkig en roepen waarschijnlijk de vraag op: "Zijn de door mij genomen beslissingen wel juist en hoe kan ik ervoor zorgen dat het mij niet gaat overkomen?" Het is dus zaak dat functioneel beheerders en database administrators aantonen dat de datakwaliteit goed is.
Datakwaliteit is een soms ongrijpbaar begrip. Is de juiste waarde ingevuld in bepaalde velden? Is het gegeven conform de werkelijkheid? Welke werkelijkheid? Spreken we bijvoorbeeld over hetzelfde als we het hebben over een gegeven als 'klantrelatie'?

Definitie

Met een juiste definitie is het begrip datakwaliteit meetbaar te maken. Er kunnen (internationale) standaards worden gebruikt zoals een naamgevingstandaard, ISO-notatiewijze en referenties/ interfaces met specifieke instanties zoals de Gemeentelijke Basis Administratie (GBA), het straatnaam- en postcode-register en de Kamer van Koophandel. Hoe de gegevens er uitzien en welke keuzes men erbij maakt, is niet altijd van belang. Belangrijk is dat iedereen die waarde hecht aan de inhoud van de gegevens hetzelfde bedoelt en dat dit voor iedereen duidelijk is.

Wat is datakwaliteit?

In een aantal artikelen behandelt Nico Klaassen enkele vragen over datakwaliteit. In dit is eerste artikel gaat het om wat datakwaliteit is, en is het te meten? Hoe meet je datakwaliteit? In het tweede artikel staat de vraag centraal of datakwaliteit nog beter kan en hoe ga je om 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.

Een eenduidige definitie van datakwaliteit is niet te geven. Voor elke organisatie kan de kwaliteit van de gegevens maar ook het belang van een goede kwaliteit anders zijn. Indien een directe relatie wordt gelegd met klanten, is het van belang dat men alle gegevens up-to-date heeft en er géén fouten in namen en/of adressen voorkomen. De kwaliteit van de gegevens is dus een weergave van de accuratesse van de eindgebruikers en bij fouten beschadigt dit mogelijk het imago.

> Belangrijk is dat iedereen die waarde hecht aan de inhoud van de gegevens hetzelfde bedoelt

Hoe is datakwaliteit inzichtelijk te maken? In de definitie van datakwaliteit kan men een tweetal varianten onderkennen, de technische kwaliteit en de functionele kwaliteit. Beide soorten kwaliteit kunnen worden gemeten maar vragen een andere aanpak en ook inspanning. Hoe hiermee om te gaan wordt uitgewerkt.

Technische datakwaliteit

Met een technische data-analyse of een statistische data-analyse, worden alle opgeslagen data-elementen onderzocht op een aantal aspecten:
- formaatcontroles;
- vullingsgraad van velden;
- minimum- en maximumwaarden;
- domeinwaarden;
- domeinverdelingen;
- jaar- of jaar/maand-verdelingen.

Op basis van bovenstaande analyse-aspecten krijgt men snel het eerste beeld van de gegevens. Als velden niet consistent zijn ingevuld of als er extreme minimum- en maximumwaarden zijn, is dat een mogelijke indicatie voor een probleem met scherm-validaties. Maar ook kunnen extreme waarden of domeinen wijzen op mogelijke invoerfouten door eindgebruikers. Zo heeft het invoeren van 2072 in plaats van 1972 voor een factuurdatum al eens geleid tot het alsnog doorbelasten van verrichte diensten.

Om dit soort metingen te kunnen verrichten kan men uiteraard met op de markt verkrijgbare tools onderzoek doen. Regelmatig echter zijn de hoge kosten toch een nadeel van dit soort tools. Een alternatief is om deze controles via bijvoorbeeld een query generator uit te voeren. Een zéér eenvoudige wijze is mogelijk via het opslaan van resultaten in een scantabel en het op regelmatige tijdstippen uitvoeren van query's.

Hieronder volgt een voorbeeld van zo'n query.

```
Insert into dbo.scanresult
   (Tabelname, fieldname,
   scandatetime, scantype,
   scanresult)
Select
   'Employees' 'hiredate',
   timestamp, 'max',
   max([Hiredate])
   From dbo.employees
```

Niet alles is direct in een query op te lossen, maar stored functions kunnen daarbij erg helpen. Bij het samenstellen van de query's kan de DBMS Datadictionary veel waardevolle informatie geven, waarmee voorkomen wordt dat men elke query zelf moet samenstellen. Belangrijk is dat de resultaten in een afzonderlijke tabel of file geplaatst worden zodat men ze later nog kan gebruiken. Een melding als 'er is een rare datum gevonden' is waardevoller als ook aangegeven kan worden wat voor datum en bij voorkeur bij welke klant.

> Quick wins moeten worden voorzien van hoge prioriteit

Voordat het verzamelen van de gegevens start moet eerst bepaald worden welke informatie men wil achterhalen en wat men ermee wil gaan doen. Enkele voorbeelden:

- het zoeken naar gaten in reeksen;
- minimum- en maximumwaarden gebruiken om extremen te onderzoeken;
- het achterhalen waar bepaalde velden niet consequent zijn ingevuld, om problemen met programmatuur in kaart te brengen;
- de verdeling in domeinreeksen achterhalen
- plausibiliteitcontroles uitvoeren om typefouten te achterhalen;
- telefoonnummers in een specifiek formaat (10-cijfers, met of zonder streepjes) om gegevens beter te structureren;
- het juiste formaat van bankrekening- en gironummers (P9999999) om eenduidiger selectie van gegevens mogelijk te maken.

Vaak zal deze informatie al leiden tot *quick wins*, aangezien probleemgebieden snel inzichtelijk te maken zijn en relatief eenvoudig opgelost kunnen worden.

Functionele datakwaliteit

Bij deze controles is men erop gericht om specifieke gebieden, functionele of business-gebieden, inzichtelijk te maken. Lastig aan deze controles is dat ze vaak erg complex zijn en het moeilijk is om alle functionaliteiten te achterhalen waarop gecontroleerd moet gaan worden.
Op basis van de eventuele problemen of aandachtspunten die tijdens de technische analyse zijn geconstateerd en de inbreng van functioneel beheerders en gebruikers, kan een nadere definitie worden verkregen van welke gebieden inzichtelijk gemaakt moeten worden. Een workshop van een halve dag is vaak voldoende om de vragen vanuit functioneel oogpunt te achterhalen.

Bij veel functionele vragen die beantwoord moeten gaan worden is prioriteitstelling van belang. Ook hierbij geldt dat quick wins moeten worden voorzien van hoge prioriteit. Ook voor deze functionele vragen worden vervolgens controle-query's of -programma's ontwikkeld om de mogelijke problemen boven tafel te krijgen en ook hierbij geldt dat de beantwoording traceerbaar is naar de specifieke gevallen.
Enkele functionele vraagstukken zijn bijvoorbeeld: welke klanten hebben géén eigen account manager; welke artikelen hebben een voorraad onder de minimumvoorraad; welke coderingen worden niet meer gebruikt; welke klanten komen dubbel voor in ons klantenbestand?
Met name het laatst genoemde vraagstuk komt steeds vaak voor. Reden dat deze vraag in de Top 5 van de functionele vraagstukken staat, is dat door optimalisatie van de relaties beter inzicht kan worden verkregen in het klantenbestand. Maar bovenal zijn dit de complexere en moeilijkere problemen waar een eenvoudige query vaak géén uitkomst biedt en waar externe tools vaak hulp kunnen bieden.

Een voorbeeld om gegevens te kunnen ontdubbelen, toont aan dat het bij dit soort situaties erop neerkomt dat men eerst goed nadenkt, en dan pas moet gaan doen:

- bepalen welke gegevens van klanten op zich identificerend zijn (bijvoorbeeld het sofi-nummer: het eigen klantnummer is met grote regelmaat niet betrouwbaar om uniciteit te garanderen);
- bepalen welke combinaties van gegevens van klanten identificerend zijn (bijvoorbeeld naam, voorletters en geboortedatum, waarbij rekening moet worden gehouden met foutieve spellingen, puntjes en spaties tussen letters en andere tekens die geëlimineerd moeten worden om matchen mogelijk te maken);
- matchen van combinaties om dubbele gegevens te vinden en toekennen van scores om 'kans' op dubbel zijn aan te tonen (inclusief de referenties naar de waarden die dan overeen zouden komen);

- het omgaan met de dubbele waarden die men vindt: hoe weet men nu zeker dat ze ook inderdaad hetzelfde zijn en hoe dit aan de gebruiker te presenteren, die hierover een laatste beslissing moet nemen. (Men moet hierbij rekening houden met de kans dat een dubbele waarde voorkomt 5 procent is. Op een bestand met 1.000.000 waarden levert dit al snel 50.000 dubbele waarden op, het oplossen hiervan kost veel tijd.)

Is alles inzichtelijk?

Er zullen vele vragen beantwoord worden, maar de praktijk leert dat de vele mogelijke fouten die op kunnen treden *niet* allemaal inzichtelijk worden gemaakt. Waarom niet, zal men zich afvragen. Reden is dat het volledig onderzoeken van alle mogelijke problemen een erg kostbare en vaak tijdrovende aangelegenheid is. Men moet dus de aandacht richten op die aspecten die voor de diverse betrokken partijen van belang zijn. Weeg hierin een aantal aspecten tegen elkaar af; zoals afbreukrisico voor de organisatie, commerciële belangen, gebruikersgemak, de kans op vervuiling en uiteraard (de) kosten. Het is van belang dat deze partijen ook vroegtijdig in het project worden betrokken en hun belang voor kwalitatief goede gegevens kunnen inbrengen.
Uiteindelijk zijn de mogelijke problemen in kaart gebracht en zullen de acties uitgezet moeten worden om deze problemen op te lossen. Maar het oplossen alleen is niet voldoende, ook zullen aanpassingen noodzakelijk zijn om te voorkomen dat de problemen zich in de toekomst gaan herhalen. Het definiëren van een stap gericht op ‘corporate data modelling’ om eenduidige definities en de kwaliteitscriteria vast te leggen en een verbetervoorstel om tot een beter gegevensbeheer te komen, is dan ook een belangrijke vervolgstap. Voorkomen is ten slotte ook voor gegevensbeheer en kwaliteitswaardige gegevens dè beste oplossing. In het volgende artikel wordt dieper in gegaan op het schonen van gegevens en de aspecten die een rol spelen om problemen in de toekomst te verminderen.

Nico Klaassen

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