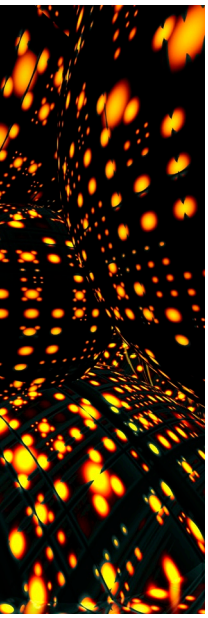


DataFlux pakt het probleem bij de bron aan

Hoe organiseer je datakwaliteit?

Petrick de Koning

Datakwaliteit is het fundament voor efficiënte bedrijfsprocessen en goede besluitvorming. Betrouwbare data geven een goed beeld van een organisatie, klanten en de markt en dit kan bijdragen aan omzetgroei, risicobeheersing en compliancy.

Volgens onderzoek van SearchDataManagement.com verliest een gemiddeld bedrijf ruim 8 miljoen dollar per jaar aan slechte datakwaliteit. De business case voor datakwaliteit behoeft dus geen betoog. Tot zo ver de theorie. Maar hoe implementeert een organisatie een programma dat de kwaliteit – en de bijbehorende voordelen – tijdens de gehele levenscyclus van data waarborgt? Datakwaliteit was in eerste instantie een *nice-to-have* met het oog op omzetverbetering of om te zorgen dat een klant niet zes keer dezelfde mailing ontving. Dit scheelt het bedrijf veel geld en vermindert de ergernis bij klanten. Toenemende economische druk, compliancy vereisten en andere wetgeving hebben ervoor gezorgd dat steeds meer organisaties investeren in datakwaliteit. In een aantal gevallen is de reden eenvoudig: als een bedrijf zijn datakwaliteit niet op orde heeft, is er een risico op hoge boetes. Bovendien is het belangrijk om te beseffen dat datakwaliteit een uniek probleem is dat de grenzen van databases overstijgt en begint bij correcte en eenduidige invoer. Het is dus zaak om datakwaliteit te integreren in de operationele systemen waar veel data ingevoerd worden en van waaruit veel datastromen de organisatie ingaan, bijvoorbeeld voor managementrapportages of business analytics-toepassingen.

Zelf services ontwikkelen

Deze integrale aanpak vanuit business-perspectief staat centraal in de datakwaliteitoplossingen van DataFlux. Dit bedrijf, onderdeel van SAS, bevindt zich de laatste jaren bovenaan in de leiderssectie van het Magic Quadrant voor datakwaliteittools van Gartner. Deze positie is gebaseerd op technologie die organisaties helpt om zelf specifieke services te ontwikkelen om data op te schonen, te valideren volgens eigen business-regels en deze consequent toe te passen en te toetsen. Op die manier is de kwaliteit tijdens het complete traject van invoer, integratie, opschonen en validatie tot verwijdering, volledig transparant en beheersbaar. Voor DataFlux gaat datakwaliteit veel verder dan een éénmalige opschoonactie. Als een bedrijf daarvoor kiest, is het onvermijdelijk dat snel weer een nieuwe 'datapuinhoop' ontstaat. Het hele proces kan dan weer opnieuw beginnen.

Inventarisatie

Datakwaliteit moet bij de bron worden aangepakt. Daarom begint een datakwaliteitprogramma altijd met een inventarisatie van de bronsystemen. Wat dat betreft is er zeker een parallel met de technologische ontwikkelingen rondom het inrichten van datawarehouses. Ook hierbij draait het om het integreren van verschillende bronnen dat moet leiden tot een eenduidige en gestandaardiseerde omgeving. Hoewel het in kaart brengen van de bronsystemen een logische stap lijkt, is het in veel gevallen niet zo eenvoudig. Als we uitgaan van klantgegevens blijkt dat data in veel verschillende, veelal legacy, systemen en op meerdere locaties zijn opgeslagen of in documentatie zijn vastgelegd. De primaire data zijn vaak wel inzichtelijk via case tools of ER-diagrammen, maar ook andere, gerelateerde data moeten inzichtelijk zijn om de kwaliteit te waarborgen.
Als deze stap wordt overgeslagen loopt men in dezelfde valkuil als bij datawarehouse-trajecten. Wanneer gedurende het proces blijk dat er relevante gegevens gemist zijn, kost het namelijk extra tijd en geld om deze data alsnog te integreren. Deze extra kosten worden veroorzaakt door het feit dat er extra werkzaamheden verricht moeten worden om de data te begrijpen en analyseren. De analyse gebeurt met exploratietechnologieën die in databases tabellen en velden doorzoeken op klant- of productgerelateerde data. Op basis van kwaliteitsregels voor namen en inhoud van velden 'herkent' deze software 'kandidaatdata' en verwerkt deze in rapportages. Deze inventarisatie is een noodzakelijk stap voor het plannen en de prioriteitstelling van het daadwerkelijke datakwaliteittraject.

Quality Knowledge Base

Om letterlijk dichtbij de bron te zitten, integreert DataFlux haar technologie met de bronsystemen van organisaties om op die manier de invoer van bijvoorbeeld klantgegevens te standaardiseren. Vanuit de invoervelden van systemen zoals SAP of Siebel worden automatische web services aangeroepen. Deze toetsen de ingevoerde data met de Quality Knowledge Base (QKB). Deze QKB bevat gangbare regels zoals standaardisatie van basisvelden naam, adres, postcode en telefoonnummer. Daarnaast worden ook aspecten als verkeerde spelling meegenomen. Dat geldt ook voor invoer van namen die potentieel dezelfde klant of contactpersoon kunnen zijn. Bijvoorbeeld Wim en Willem. Dit alles gebeurt op contextniveau. Deze standaardisatie houdt ook rekening met verschillende schrijfwijzen per taal. De spelling van namen in Nederland wijkt af van die in België. In Nederland is Van den Bosch in gebruik, terwijl in België Vandenbosch gangbaar is. Voor een bedrijf dat in de Benelux actief is, is het belangrijk om voor de klant geen dubbele gegevens te beheren.
De standaardisatie op basis van de QKB leidt tot validatie. Als data via het bronsysteem worden ingevoerd of gewijzigd, vindt er namelijk automatisch een toetsing aan de kwaliteitsregels plaats. Deze aanpak voorkomt foutieve invoer en ondersteunt ontdubbeling. Zo maakt een organisatie direct een fikse kwaliteitsslag. Het opschonen van historische gegevens is een ander traject en kan in batches. Ook in dit proces staat de toetsing aan de QKB centraal. Afhankelijk van de specifieke situatie zijn hiervoor diverse scenario's inzetbaar. Bijvoorbeeld voor versiebeheer als gecontroleerde records niet verwijderd mogen worden. In het geval van het opschonen van data uit operationele systemen, is het niet altijd wenselijk of zelfs toegestaan om dit tijdens de operatie te doen. Verder is ook mogelijk deze activiteiten in een schaduwomgeving te laten draaien.

Resultaten

De effecten van een integrale aanpak van datakwaliteit zijn in de praktijk bewezen. Zo had een retailer met honderden miljoenen klantcontacten per jaar problemen met de productdatabase. Deze bevatte meer dan 7 miljoen producten en onderdelen. Het was onduidelijke welke data actueel of verouderd waren. Bovendien ontbraken een uniform format en standaarden. De retailer worstelde met inconsistente data vol typefouten en onherkenbare afkortingen. Dit had een negatief effect op het serviceniveau en de klanttevredenheid, omdat producten of onderdelen slecht vindbaar waren en er regelmatig dingen misgingen bij bestellingen. Met DataFlux-technologie zijn alle productonderdelen gestandaardiseerd en ondergebracht in een gestandaardiseerde model- en productbeschrijving. Op die manier zijn alle fouten gecorrigeerd en vinden aanvullingen en wijzigingen gestandaardiseerd plaats. De datakwaliteit van de productdatabase is gewaarborgd. Klanten kunnen de juiste producten en of onderdelen veel eenvoudiger op de website vinden. Dit scheelt veel problemen met de levering en draagt bij aan een grotere klantloyaliteit.

Een ander voorbeeld speelt zich af in de financiële sector. Een inkomensverzekeraar voor managers en ondernemers beheerde ruim 250.000 polissen en meer dan 4.000 actieve claims. Dit proces werd bemoeilijkt door inconsistentie en fouten in de datasystemen. Hoewel bekend was dat de datakwaliteit te laag was, waren de omvang en effecten van deze fouten onbekend. Juist omdat er onduidelijkheid bestond over welke gegevens in de datasystemen niet klopten en wat er fout ging in de processen, was het onmogelijk om de kwaliteit van nieuwe data te controleren. Samen met DataFlux pakte de verzekeraar het probleem bij de bron aan; bij de medewerkers die de data invoeren en er mee werken. Daardoor werden zij verantwoordelijk voor de kwaliteit van de data en in het verlengde daarvan de efficiency van de bedrijfsprocessen. Aan de hand van nieuwe business-regels zijn inconsistente data in de polissen aangepast. Tegelijkertijd is de invoer van nieuwe gegevens nu gebaseerd op een helder format dat bovendien real-time gemonitord wordt. De voortdurende toetsing aan de QKB bewaakt de correcte invoer van nieuwe gegevens. Inmiddels is datakwaliteit een integraal onderdeel van de bedrijfsvoering en kan het bedrijf claims sneller en efficiënter afhandelen.

Social media

De bestaande databronnen binnen de organisatie worden geregeld uitgebreid. Goede voorbeelden zijn het web of social media zoals LinkedIn, Facebook of Twitter. In sommige gevallen is het nuttig om deze ongecontroleerde data te koppelen aan een uniek klantenbestand. Deze datastroom leidt tot een continue noodzaak om de input te monitoren. Dit proces is grotendeels te automatiseren. Voor gereguleerde invoer via SAP en Siebel vereist een automatische check met de QKB weinig extra actie. Voor processen waar een klant zelf online gegevens invoert is het van belang om dit proces grondig te monitoren om datavervuiling en

Vijf succesfactoren

Datakwaliteit is uitgegroeid tot een volwassen discipline, om de voordelen van de beschikbare kennis en technologie optimaal te benutten is een gespecialiseerde partner onmisbaar.

Datakwaliteit is een business-onderwerp; definieer het doel van datakwaliteit met de business. IT faciliteert.

Stel prioriteiten en investeer genoeg tijd en capaciteit in een grondige inventarisatie. Als dit niet gebeurt, investeert een bedrijf alleen maar in het creëren van het volgende datakwaliteitsprobleem.

Blijf datakwaliteit voortdurend monitoren om de waarde en patronen inzichtelijk te houden en te kunnen anticiperen op interne en externe veranderingen.

Communicatie met gebruikers over het belang van datakwaliteit is ontzettend belangrijk. Niet alleen aan het begin van een project, maar ook tijdens en na de implementatie. De uitzonderingen die medewerkers, die data invoeren en beheren, tegenkomen bieden waardevolle input voor verbeteringen.

De componenten in een datakwaliteitprogramma

Een integraal datakwaliteitprogramma vereist enkele componenten die in onderstaande afbeelding zijn weergegeven.

Entity resolution. Meet de mate van gelijkheid tussen data-elementen uit verschillende bronsystemen. Door unieke instances te bepalen is informatie toe te wijzen aan een geconsolideerd record. Maar de data kunnen ook geoormerkt worden voor handmatige verwerking. Entity resolution is een lastige taak omdat er zelden een exacte overeenkomst met alle datavelden is. Intelligente vergelijkingstechnologie brengt deze data met elkaar in overeenstemming en ondersteunt de integratie in meerdere records.

Creëren en beheren van business-regels. Intuïtieve technologie helpt business-gebruikers snel regels te bepalen om afwijkende data te signaleren. Deze regels kunnen zich richten op de datakwaliteit zelf, bijvoorbeeld het standaardiseren van adresgegevens en een controle op dubbele data. Maar ook op de data-integriteit voor bedrijfsprocessen. Als iemand een salaris van tussen de 5 en 500 euro invoert, krijgt de beheerder standaard een waarschuwings e-mail.

Verificatie, normalisatie, standaardisatie en transformatie. De regels worden automatisch toegepast en aangepast, inclusief opschonen van bestaande data. Eventuele afwijkingen worden herkend en verwerkt tot nieuwe regels. Deze kernprocessen vinden grotendeels automatisch plaats in de Quality Knowledge Base.

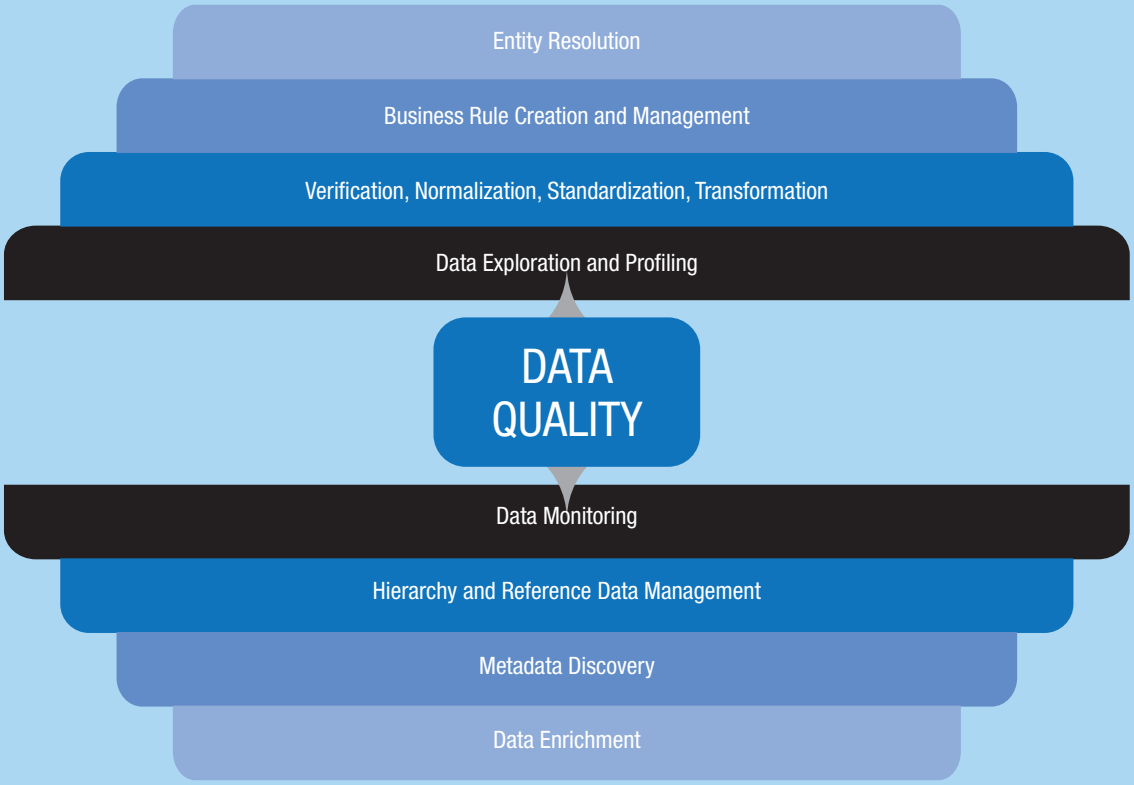
Data-exploratie en profiling. Profiling biedt statistisch inzicht in de data en helpt de oorzaak van kwaliteitsproblemen te achterhalen. Dit inzicht biedt handvatten om effectieve regels voor datakwaliteit te creëren. Data profiling biedt een complete analyse van alle data binnen een organisatie, onderzoekstructuren en bewaakt de volledigheid van alle databronnen.

Datamonitoring. Data monitoring voert periodieke controles van het databeheer uit en biedt het fundament voor goed gefundeerde besluitvorming via vroegtijdige waarschuwingen bij data-afwijkingen. Als afwijkingen van de gestelde regels snel ontdekt worden, is het mogelijk om direct actie te ondernemen en structurele data-issues te voorkomen.

Hiërarchie en referentiedatabeheer. Met classificatiecodes is het mogelijk om zogenaamde 'ouder-kind relaties' voor verschillende bronnen te vinden en in vast te leggen. Dit is mogelijk door relaties tussen klantgegevens te oormerken. Bijvoorbeeld om een arts te koppelen aan een maatschap of een ziekenhuis.

Metadata discovery. Metadata discovery legt bestaande trends en kenmerken van data bloot. Hierdoor ontstaat een goed beeld van de verschillende soorten bedrijfsinformatie, inclusief de bijbehorende bronnen. De combinatie van metadata-analyse en data discovery maakt de metadata inzichtelijk en ondersteunt goed beheer van de gehele levenscyclus van data.

Dataverrijking. Zodra data gestandaardiseerd en geïntegreerd zijn, kan de verrijking starten. Dat kan door data te vergelijken met geografische, demografische of andere gegevens. Verder is het mogelijk om data over producten, materialen of services te standaardiseren. Al deze gegevens worden getoetst binnen het bestaande datakwaliteitprogramma.



dubbele gegevens te voorkomen. Hiervoor kan een organisatie dan eigen regels opstellen en bijvoorbeeld specifieke velden voor postcode of telefoonnummer afdwingen. Als er dan toch afwijkingen zijn, kan het nodig zijn dit via e-mail kenbaar te maken aan een medewerker die dan een handmatige controle uitvoert.

Datadynamiek

Data zijn dynamisch en vragen voortdurend aandacht. Naast interne factoren zijn er ook externe invloeden die er voor zorgen dat bedrijven hun datakwaliteitprogramma toch moeten aanscherpen of aanpassen. Deze bedrijven hebben bijvoorbeeld de kwaliteit van hun marketing- of CRM-data goed op orde, maar lopen dan op een ander vlak toch tegen een nieuw probleem op.

> De effecten van een integrale aanpak van datakwaliteit zijn in de praktijk bewezen

Een goed voorbeeld is een autoverhuurbedrijf dat vanuit de overheid te maken krijgt met strikte eisen op het gebied van due diligence ten aanzien van klanten of prospects. Een dergelijk bedrijf heeft vanuit commercieel oogpunt flink geïnvesteerd in datakwaliteit van klantgegevens, maar moet dan ook dubieuze personen of organisaties identificeren. Dit gebeurt dan door nieuwe klanten of transacties te toetsen aan specifieke lijsten van nationale en internationale overheden. Ook deze toets, waarin aspecten zoals verschillende manieren van een naam spellen of gebruik van identieke adressen van belang zijn, moet in het datakwaliteitsproces geïntegreerd worden.

Business-analisten

Juist vanwege de onvoorspelbaarheid van deze veelal externe invloeden benadert DataFlux datakwaliteit vanuit een businessperspectief en heeft haar technologie daarop aangepast. Binnen klantorganisaties zijn business-analisten verantwoordelijk voor het ontwikkelen van de vereiste business-regels en de vertaalslag naar de bedrijfsvoering. Zij kennen de inhoudelijke eisen die aan de data gesteld worden en kunnen de vertaalslag naar de informatiebehoefte in de organisatie maken. Met intuïtieve tools worden business-analisten ondersteund in het opstellen van de regels. Deze kunnen uit eigen ervaring ontstaan of gebaseerd zijn op uitzonderingen die voortkomen uit periodieke audits. Alle ideeën zijn te toetsen met concrete voorbeelddata om de effecten en resultaten van de regels inzichtelijk te maken. Pas na het definiëren van de regels komt IT in beeld. Om de afstemming tussen beide te optimaliseren maken zowel de business-analisten als IT'ers gebruik van dezelfde tools. IT verzorgt de integratie in de bestaande omgeving en de implementatie van de web services in SAP of Siebel. Door deze werkwijze hoeft er niet opnieuw gespecifieerd of gecodeerd te worden en wordt voorkomen dat individuele interpretaties datakwaliteit negatief beïnvloeden.

Geen garanties

Omdat datakwaliteit de gehele levenscyclus van data omvat, kan het ook nodig zijn dat data verwijderd worden. Als een klant verhuist, moet het oude adres verdwijnen om verwarring en miscommunicatie te voorkomen. Dit kan op verschillende manieren. Met een query om incorrecte data uit tabellen te verwijderen of door oude data in een archief op te slaan. Dat is onder andere mogelijk door data te voorzien van een einddatum. Het feit dat data op een gegeven moment aan het einde van de levenscyclus zijn, doet niets af aan het belang van continue aandacht voor datakwaliteit. Zelfs als een bedrijf alle stappen – standaardisatie, validatie en monitoring – consequent doorloopt, is er geen garantie voor optimale datakwaliteit. Alleen een dynamische aanpak waarbij de business-regels auditresultaten voortdurend toetsen en aanpassen, biedt de zekerheid dat de kwaliteit in lijn blijft met de bedrijfsdoelstellingen. Een technologieplatform dat bedrijven in staat stelt eigen business-regels te ontwikkelen en toe te passen kan daaraan wel een structurele bijdrage leveren.

Petrick de Koning in gesprek met Wilbram Hazejager,
EMEA Solutions Architect and Product Management, DataFlux.