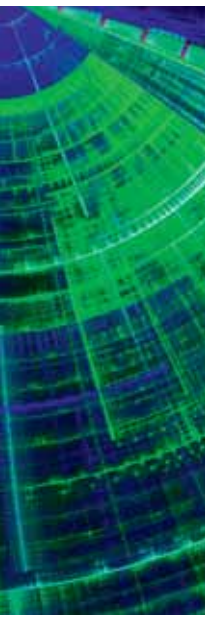


Verschil tussen primaire en secundaire processen speelt cruciale rol

Datakwaliteit en requirements – the missing link

Tom Breur

In menig BI-project klaagt men over datakwaliteit. Tegelijkertijd valt ook iets anders op: hoewel veel mensen zich bewust zijn van datakwaliteitsproblemen, blijft het toch vooral een 'motherhood and applepie' issue. Iedereen is het er over eens dat het een probleem vormt, en iedereen vindt dat er iets aan moet gebeuren. Maar als er vrijwilligers gezocht moeten worden om er ook echt iets aan te doen, blijft het meestal erg stil.

Upstream in het BI-proces is een matige datakwaliteit vaak de oorzaak van vertraging in datawarehousetrajecten. Met name de ETL-fase is hier het meest gevoelig voor, een fase die zelf al zo'n groot gedeelte van de doorlooptijd van het project voor zijn rekening neemt.
Maar ook downstream gebeurt het regelmatig dat op zich 'mogelijke' maar inhoudelijk ongeldige waarden moeten worden bewerkt. En daarnaast heb je ontbrekende waarden in heel wat smaken: n.v.t., onbekend, (nog) niet beschikbaar, enzovoort. Afhankelijk van het doel van rapportages of analyse moeten hier soms tijdrovende transformaties op worden uitgevoerd.
Om de context van datakwaliteit te verduidelijken maak ik een onderscheid tussen primaire en secundaire processen in een organisatie. Met primaire processen wordt direct waarde voor de klant gecreëerd; het boekingsproces bij een bank, reserveringen bij een reisorganisatie, enzovoort. Secundaire processen ondersteunen het primaire bedrijfsproces. Denk daarbij aan Finance, HR, Strategie, en dus ook (de meeste smaken van) BI. Secundaire processen worden doorgaans gekenmerkt door (beduidend) minder goede datakwaliteit. Waarom?

Primaire versus secundaire processen

De fundamentele reden waarom primaire processen gekenmerkt worden door hogere datakwaliteit is gelegen in het feit dat er 'automatisch' correctie optreedt bij voorkomende fouten. Een voorbeeld: als ik een boeking heb bij een luchtvaartmaatschappij, en bij de balie blijkt er voor mij geen stoel beschikbaar te zijn (laten we het structurele 'overboeken' even buiten beschouwing laten), is er *onmiddellijk* een probleem. En dit probleem zal ter plekke moeten worden opgelost, op kosten van de maatschappij. Als de bank een fout maakt op een rekeningoverzicht, zal de klant onmiddellijk reageren en het verschil reclameren. Als er een patroon optreedt in deze fouten, dan wordt dit doorgaans snel opgeschaald om herhaling te voorkomen. De reden dat primaire processen worden gekenmerkt door betere datakwaliteit is omdat fouten onmiddellijk leiden tot 'frictie' in het primaire proces. Dit zorgt ervoor dat 'automatisch' een zelfcorrigerend mechanisme in werking treedt. De principiële oorzaak hiervan is dat de 'pijn' die veroorzaakt wordt door eventuele fouten in de data direct gevoeld wordt door de business, en hen ook ogenblikkelijk geld en resources kost.
Secundaire processen daarentegen beschikken doorgaans over minder goede datakwaliteit, want bovenstaand 'zelfcorrigerend' mechanisme ontbreekt. Wanneer bijvoorbeeld de afdeling HR voor sommige werknemers niet kan beschikken over de datum waarop hun dienstverband werd beëindigd, geeft dit een probleem. Je kunt dan de pensioen benefits niet berekenen die aan deze ex-medewerker moeten worden toegekend. Echter dit 'blijkt' pas ruim nadat de werknemer uit dienst is. En de persoon die pensioenopbouw berekent is niet degene die verantwoordelijk was voor de (correcte) invoer van de datum 'einde dienstverband'. Gevolg hiervan is dat dergelijke problemen in secundaire informatiestromen de neiging hebben voort te duren.

De rol van requirements

Er zijn twee redenen waarom primaire processen 'van nature' betere datakwaliteit hebben. Op de eerste plaats is er betere 'business alignment'. Op de tweede plaats is er een (veel) kortere feedback loop tussen data non-kwaliteit en problemen die hier het gevolg van zijn. En precies *daar* zit de enigszins 'verborgen' rol die requirements spelen bij het realiseren van datakwaliteit. Het doel van BI is om beslissingen te ondersteunen. Dat wil zeggen dat BI nooit business *eigenaar* wordt, maar een partner is die feiten en systemen aanlevert waarmee de business in staat wordt gesteld *zelf* betere beslissingen te nemen. Dit spreekt voor zich, maar wat hebben requirements hier mee te maken?

Wanneer BI-tools worden gebruikt om beslissingen te ondersteunen, zal de uitkomst van die beslissing de 'feedback' moeten genereren die datakwaliteit in de hand werkt, analoog aan het primaire proces. Dus wanneer een slechte beslissing achteraf het gevolg bleek te zijn van foutieve (of ontbrekende) informatie, treedt er een corrigerend mechanisme in werking dat vergelijkbaar is met het primaire proces.

Een verschil tussen primaire en secundaire processen is dat het zelfcorrigerende mechanisme in het laatste geval doorgaans met enige vertraging gebeurt. De crux is dat wanneer 'echte' beslissingen worden genomen op basis van BI, deze feedback het sterkst en snelst zal werken. Wanneer BI-rapporten voornamelijk een 'nice to know' karakter hebben dus niet!

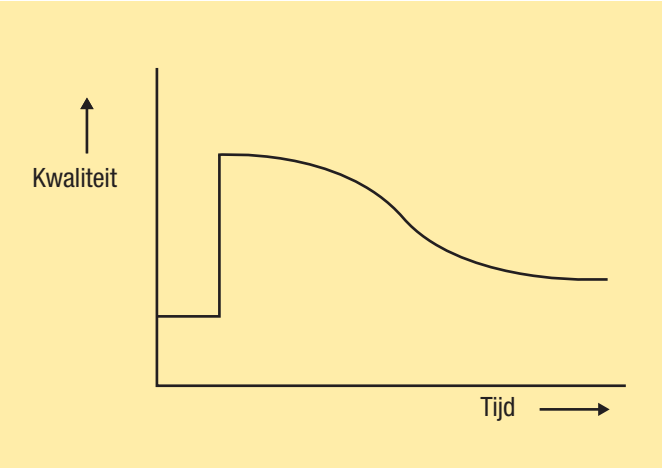
Hoe dichter probleemhouder en probleemeigenaar 'bij elkaar zitten' hoe beter

Dit is de reden waarom je altijd positief kritisch moet blijven bij het verzamelen van requirements. Waarom heb je dat rapport nodig? En wat ga je dan doen als je die informatie eenmaal hebt? Hoe verschilt dat van de huidige praktijk? Hoe wordt die keuze beïnvloed door kennis die je afleidt uit data? Het doel van dit soort 'kritische intakes' is uiteraard *niet* om een barrière op te werpen die BI minder toegankelijk of beschikbaar maakt. Op de eerste plaats probeer je het beslisproces helder te krijgen, en hoe dat het best ondersteund kan worden met data. Een bijkomend voordeel is dat de eerder beschreven feedback loop (veel) beter zijn werk kan doen, wat betere datakwaliteit in de hand zal blijven werken.
Oogmerk van deze 'positief kritische' aanpak is tweeledig. Op de eerste plaats leidt dit tot betere briefings, met scherpere requirements, wat betere datakwaliteit bevordert. Want alleen als een beslissing afhangt van een bepaald gegeven, is er een zelfcorrigerend mechanisme dat de structuur biedt om duurzame datakwaliteit op te leveren. Dit in tegenstelling tot 'nice to know' rapportages van hoegenaamd 'strategisch' belang (wie kent ze niet?). Op de tweede plaats leiden dergelijke intakes er toe dat BI prioriteit geeft aan ondersteuning van rapportages en systemen waarmee de meest waardevolle beslissingen voor de business worden genomen. Dit tweede punt leidt tot meer strategische invulling van data governance.

Business alignment

De praktijk is in de meeste gevallen dat we beginnen vanuit een situatie met imperfecte datakwaliteit. Als professional wordt van ons verwacht dat we het beste proberen te maken van een net niet ideale situatie. Een belangrijk deel van de complexiteit wordt veroorzaakt doordat er binnen organisaties van nature tegenstrijdige belangen heersen. Traditionele cost-accounting modellen meten resultaat in het verleden, maar investeren in de toekomst is ook nodig om competitief te blijven. De druk van aandeelhouders om constant te (blijven) presteren op de korte termijn kan de aandacht afleiden van de eeuwig durende zoektocht naar nieuwe groeimogelijkheden.
Wie kent niet de klassieke dilemma's als investeren in marketing of winstmaximalisatie voor het lopende jaar. Of een nieuw product zo snel mogelijk lanceren, versus doorontwikkelen (testen) om klachten te voorkomen. Spanning tussen het meest begeerlijke product (features) voor de consument, of kiezen voor een eenvoudiger maakbaar product. Goedkoop maar traag en onbetrouwbaar leveren versus investeren in logistiek, enzovoort. Het is aan BI om dergelijke business dilemma's helder bloot te leggen, de consequenties te onderbouwen met feitelijke gegevens, en belangenafwegingen te definiëren. Precies in die rol als onpartijdige bewaker van de 'single version of the truth' voeg je het meeste waarde toe aan de organisatie.

Business alignment impliceert dat inspanningen van uiteenlopende units binnen een organisatie efficiënt bijdragen aan het gezamenlijke doel. Interne frictie wordt tot een minimum beperkt, waardoor (bijna) alle inspanningen nuttig zijn ten gunste van het bedrijfsresultaat.
Business alignment wordt gerealiseerd door zo veel mogelijk probleemhouder en -eigenaar dezelfde persoon of unit te maken. Hoe dichter probleemhouder en -eigenaar 'bij elkaar zitten', hoe beter. Probleemhouder is degene die 'last' heeft van een probleem, de probleemeigenaar is degene die over resources beschikt om het probleem op te lossen. Als de probleemeigenaar meedeelt in het leed dat de probleemhouder voelt zal hij geprikkeld worden 'iets' te doen aan het probleem.
Laten we dit punt met een concreet voorbeeld illustreren. Bij het ontsluiten van een nieuwe bron voor een DWH komt een 'onverwacht' probleem aan het licht. Sommige datakwaliteitsproble-



Afbeelding 1: Lange termijn-effecten van een datakwaliteitsproject (alleen).

men doen zich bij uitstek voor in datawarehouses. De gegevens waren altijd van acceptabele kwaliteit voor de *afzonderlijke* systemen, maar bij consolidatie komen er inconsistenties aan het licht. Bronsystemen worden soms pas in het DWH voor het eerst met elkaar geconfronteerd wat tot onplezierige 'verrassingen' kan leiden.
De probleemhouder is hier de beheerder van het DWH, deze staat voor de lastige taak om fundamenteel onverenigbare gegevensstromen onder één noemer te brengen. Probleemeigenaar is de verantwoordelijke voor de bronsystemen, of mogelijk (nog) hoger in de organisatie waar integriteit van gegevens kan worden 'afgedwongen.'

Voorbeeld van verkeerde alignment

Een organisatie bracht één maal per jaar een verkoopprognose uit. Deze werd gestratificeerd per regio en naar grootte van het filiaal. Het assortiment besloeg een paar dozijn productcategorieën. Het schatten en parametriseren van alle vraagcurves nam zo'n twee tot drie weken in beslag.

> Sommige datakwaliteitsproblemen doen zich bij uitstek voor in datawarehouses

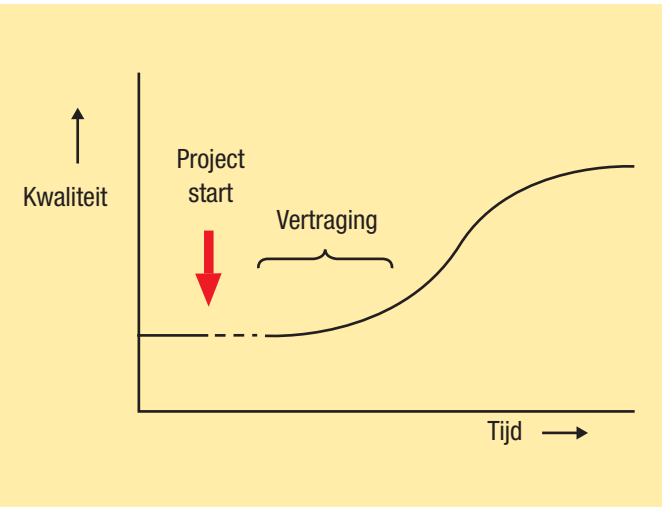
In de praktijk was de doorlooptijd van deze rapportages een veelvoud hiervan! Dat kwam omdat de aangeleverde gegevens voor ruim 14.000 filialen uiteenlopende ontbrekende of ongeldige waardes bevatten. Als dat onterechte *missings* waren moesten ontbrekende waarden worden geïmputeerd. Nadat er op geaggregeerd niveau forecasts waren berekend moesten deze worden gespecificeerd per regio, stratum en filiaal. En telkens als men detailrapportages produceerde kwamen nieuwe anomalieën aan het licht (*outliers* vallen op een hoger niveau van aggregatie veel minder op!).

Het meest 'verraderlijk' waren op zich mogelijke maar weinig plausibele waarden zoals een bedrag van 88.888,-, vooral als exact hetzelfde bedrag meerdere malen voorkwam. Let wel, speciale coderingen als 88.888 of 99.999 duiden op tekortkomingen in het datamodel van de data entry applicatie die door 'creatieve' front-office medewerkers worden omzeild.
Dergelijke waardes moesten worden vervangen door een meer plausibele schatting van het 'ware' bedrag. Maar elke keer dat een zogenaamde 'outlier' werd ontdekt moesten de missings opnieuw worden geïmputeerd en vraagcurves waarin die ongeldige waarneming was verwerkt opnieuw worden vastgesteld. Het lastige is dat eerst de geaggregeerde en dan pas gedetailleerde curves moeten worden bepaald (een artefact van de gebruikte statistische techniek). Het systeem werd uiterst 'creatief' omzeild; behalve 88.888,- bleken ook waardes als 88.887 enzovoort een 'speciale betekenis' te hebben. En hoe meer detail, hoe groter de kans dat 'nieuwe' outliers werden ontdekt, waarna het proces weer van voor af aan opnieuw moest beginnen.

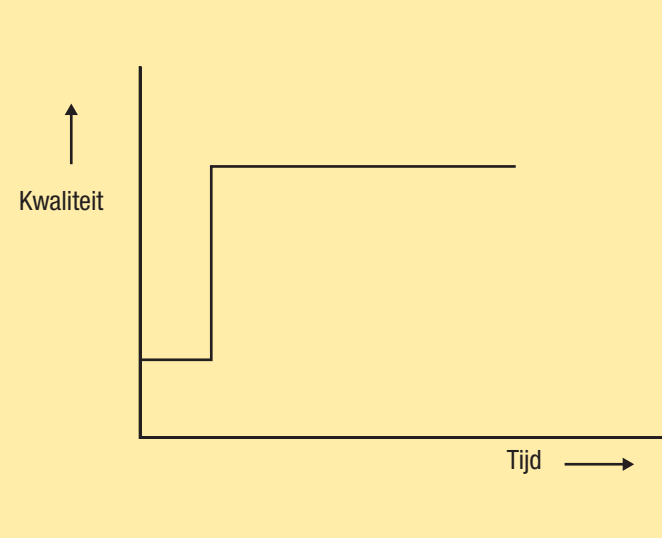
De afdeling BI was zeer bedreven geraakt in het 'interpreteren' van de data. Zij waren dan ook als enige in staat om betrouwbare rapportages en verkoopprognoses op te stellen, door de breedte van de gegevens waar zij over beschikten in het DWH. Sales management was dermate gecharmeerd van deze rapportages dat men hierop actief stuurde. Sales targets en beloningen werden hierop gebaseerd, wat de roep om betrouwbaarheid verhevigde. Gedreven door dit succes verlangde men in plaats van jaarlijks, in het vervolg rapportages op kwartaalbasis. Men was alleen niet tevreden dat de rapportages altijd zo lang op zich lieten wachten.
Hier hebben we een 'klassieke' BI-spagaat: wegens gebleken succes wordt er druk uitgeoefend om meer en/of sneller te leveren, en dan blijkt dat zelfs als er meer mankracht op dit forecasting proces wordt gezet, het slechts in zeer beperkte mate kan worden versneld. Probleemhouder van de onbetrouwbare data is BI, zij hebben de uitdaging om op zich *mogelijke* maar *ongeldige* waarden te imputeren, dat wil zeggen vervangen door de best mogelijke schatting. Wanneer zij dit niet goed (genoeg) doen, zullen hun rapportages niet worden geaccepteerd als zijnde onbetrouwbaar. Probleemeigenaren zijn de front-office medewerkers (of hun management) die het systeem vullen met de meest creatieve waarden ten behoeve van hun dagelijkse activiteiten.

Problematische datakwaliteit: wat te doen?

Aanleverende bronsystemen voor een DWH zijn (dagelijks) in gebruik en kunnen slechts in beperkte mate worden gewijzigd. Het is alsof je de winkel verbouwt maar de verkopen 'gewoon' door moeten gaan.



Afbeelding 2: Korte termijn-effecten van een datakwaliteitsprogramma (alleen).



Afbeelding 3: Lange termijn-effecten van een datakwaliteitsproject en -programma.

Sommige puristen stellen dat bronnen die data van slechte kwaliteit aanleveren deze simpelweg retour moeten krijgen. De gegevens worden dan pas ingelezen nadat de data zijn gecorrigeerd door de leverancier. Een andere aanpak is om de kwalitatief slechte data 'as is' in te lezen, en te laten gebeuren dat hiermee evident inaccurate rapportages worden geproduceerd. Het sterke punt van een dergelijke aanpak, die ik bestempel als de 'Verelendung strategie', is dat de producent van slechte data als 'boosdoener' ondubbelzinnig wordt geïdentificeerd. Deze aanpak rust op de aanname dat een senior projectsponsor vervolgens voldoende invloed zal uitoefenen (van boven af) om het voortbestaan van deze problemen te dwarsbomen. Een belangrijk nadeel van de 'Verelendung strategie' is dat er vooralsnog weinig waarde voor de business wordt gecreëerd met gegevens uit het DWH.

Korte termijn- versus lange termijnoplossing

Er zijn globaal twee manieren om datakwaliteit aan te pakken, een voor de korte, en een voor de lange termijn. Een datakwaliteitsproject is gericht op de korte termijn, een datakwaliteitsprogramma op de lange termijn. Als originele bronnen nog beschikbaar zijn (papier of elektronisch, gescand) is manuele herbevestiging een optie. Dit beschouwen we als de 'Koninklijke weg.' Datakwaliteit is nooit 100 procent, maar door gecorrigeerde data zelf te controleren krijg je een empirische schatting van de kwaliteit na afloop van een opschoningsproject.

Als er voldoende redundantie aanwezig is om bronnen met elkaar te vergelijken kunnen hieruit soms regels worden afgeleid waarmee datakwaliteit kan worden opgewaardeerd. Dit gebeurt dan door inconsistenties tussen meerdere bronnen te vergelijken. Als drie systemen dezelfde informatie bevatten, en een vierde wijkt hier vanaf, dan kies je bijvoorbeeld voor de 'meerderheidsstem.' Als je gebruik maakt van gespecialiseerde datakwaliteit software kun je ook meer geavanceerde heuristieken of algoritmes gebruiken (bijvoorbeeld Fuzzy matching).
Als organisaties snel verbetering willen realiseren in datakwaliteit kan een project opportuun zijn. Eén enkele verbeterslag is waarschijnlijk echter niet voldoende om kwaliteit op de lange termijn zeker te stellen. De resultaten volgen waarschijnlijk de grafiek in afbeelding 1.

Als de oude processen die resulteerden in non-kwaliteit onveranderd blijven, zal de kwaliteit op termijn weer afglijden naar het oude niveau. Nieuwe data stromen het systeem binnen na het project, waardoor de kwaliteit langzaam maar zeker weer afglijdt. Een lange termijnoplossing vereist een datakwaliteitsprogramma om er voor te zorgen dat nieuwe gegevens die worden ingevoerd het vereiste kwaliteitsniveau hebben.
De reden voor vertraging (zie afbeelding 2) in de verbetering is dat oude gegevens van matige kwaliteit de bronnen blijven vervuilen tot in lengte der dagen. Een combinatie van deze twee benaderingen zorgt voor snelle verbetering, en tegelijkertijd wordt kwaliteit naar de toekomst toe geborgd. In afbeelding 3 is te zien hoe zowel een snelle verbetering als lange termijn kwaliteit worden geborgd.

Conclusie

Het verschil tussen primaire en secundaire processen is een wezenlijk onderscheid dat een cruciale rol speelt bij het zoeken naar duurzame oplossingen om betere datakwaliteit te verkrijgen. BI heeft 'van nature' meer last van matige datakwaliteit omdat we gegevens uit een secundair proces betrekken. Het meest krachtige en effectieve mechanisme om datakwaliteit te beïnvloeden is de inspanningen te concentreren op het ondersteunen van de belangrijkste en meest waardevolle beslissingen voor de business. Door 'echte' beslissingen effectief te ondersteunen wordt goede datakwaliteit 'als vanzelf' in de hand gewerkt. De organisatiestructuur en toewijzing van verantwoordelijkheden speelt daarnaast ook een rol bij het realiseren van datakwaliteit. Telkens als probleemhouder en -eigenaar (te) ver van elkaar verwijderd zijn is er een verhoogd risico dat doelen niet (goed) op elkaar aansluiten. We laten deze dilemma's het best naar de oppervlakte komen door conflicterende belangen expliciet te formuleren. Ontstaat er frictie omdat individuele targets niet op één lijn liggen? Is er een conflict tussen korte en lange termijndoelstellingen? Op een punt in de hiërarchische lijn waar verantwoording (weer) samen komt kun je doelstellingen heroverwegen, of voor een andere organisatie-inrichting kiezen.

De mate waarin men beslissingen ondersteunt wordt niet bepaald *door* datakwaliteit, maar bepaalt *zelf* de datakwaliteit. En dat terwijl BI professionals je soms het tegenovergestelde willen doen geloven.

Tom Breur is eigenaar van XLNT Consulting.