Nee zeggen is wel klantvriendelijk

Validatieprocessen in plaats van GIGO

Robert Allewijn, Bas Pruijn en Vincent Wylenzek

Herkent u de volgende situatie? Een opgeleverd Business Intelligence-systeem werkt technisch prima, maar de resultaten zijn door een gebrekkige datakwaliteit slechts beperkt bruikbaar. Of deze: tijdens projecten wordt veelvuldig gediscussieerd over wat kwaliteit is en wie ervoor verantwoordelijk is. Ja? Dan moet u misschien toch eens denken over een validatieproces voor uw inkomende data.

Datakwaliteit is een van de belangrijkste succesfactoren bij datawarehousetrajecten. Bij het ontwikkelen van een datawarehouse wordt vaak gemakshalve gekozen voor de zogenaamde garbage in, garbage out (GIGO) aanpak. GIGO is eenvoudig bij het ontwikkelen van een datawarehouse. Het legt, vanuit projectperspectief, de verantwoordelijkheid voor het probleem buiten scope. Echter, bij het beheer en het gebruik is een dergelijke projectaanpak feitelijk onwerkbaar en onacceptabel.

> Door het inzetten van de validatiemodule groeit het kwaliteitsbesef enorm

Alle datakwaliteitsproblemen komen toch bij de beheerorganisatie terecht. Zij zijn immers via rapportages de boodschapper van de gebrekkige datakwaliteit. Vervolgens zien we dat ad hoc business rules worden ontwikkeld om de kwaliteit van de gegevens in het datawarehouse op te poetsen. Het probleem buiten de projectscope plaatsen lost niets op. Het doorschuiven van problemen maakt de effecten ervan meestal juist groter.

De validatiemodule

In dit artikel wordt de validatiemodule gepresenteerd; een module die aan de voordeur van het datawarehouse data tegenhoudt die niet voldoen aan de kwaliteitseisen die de afnemers van het datawarehouse stellen. Deze aanpak werkt bewezen beter dan GIGO. In eerste instantie lijkt het op de GIGO-aanpak, maar nadere beschouwing leert dat 'garbage in' wordt voorkomen door partijen aan te spreken op hun verantwoordelijkheden. Geen 'garbage in' levert dus ook geen 'garbage out' op.
De validatiemodule bestaat uit een aantal (basis)componenten, die de validatie faciliteren. De betreffende componenten zijn: kwaliteitseisen (business rules); systeemeigenaren; eindgebruikers; gegevensleveringsovereenkomsten; validatieproces (ontvangst-, technische en inhoudelijke validatie); metadata (proces- en stuurmetadata). Elk van deze componenten wordt hierna besproken en de samenhang ervan wordt weergegeven.

Kwaliteitseisen, systeemeigenaren en eindgebruikers

In deze paragraaf wordt het validatieproces van begin tot eind in hoofdlijnen beschreven. In de daarop volgende paragrafen wordt dieper op de technische validatie ingegaan. Het inzetten van de validatiemodule begint bij het vaststellen van de kwaliteitseisen die de afnemers van het datawarehouse stellen. De gebruikers bepalen welke gegevens nodig zijn en de bijbehorende kwaliteitseisen. De systeemeigenaar stelt vast of levering van deze gegevens mogelijk is. Het is dus niet de beheerder van het datawarehouse die verantwoordelijk is voor de datakwaliteit, maar de gebruiker van de informatie. Het gebruiksdoel van de informatie bepaalt welke kwaliteitseisen nodig zijn. Deze eisen worden vastgelegd in gegevensleveringsovereenkomsten. Het DWH-projectteam kan assisteren bij het opstellen hiervan. Het team kan echter nooit de verantwoordelijkheid nemen voor de inhoud.

Gegevensleveringsovereenkomst

In een gegevensleveringsovereenkomst staan alle afspraken over de gegevenslevering. Naast naamgeving van de aanlevering, aanleverfrequentie en technisch formaat, staan hierin ook de functionele eisen en wensen beschreven waaraan de aanlevering moet voldoen. Ten slotte staan de inhoudelijke controles beschreven op basis waarvan een aanlevering als 'onjuist' of 'juist' kan worden bestempeld.
Met het opstellen en het accorderen van een gegevensleveringsovereenkomst is de aanleverende bron(eigenaar) verantwoordelijk voor het correct aanleveren van de gegevens. Onder correct wordt hier zowel technisch juist als functioneel juist verstaan. Het opbouwen van een datawarehouse op basis van het vertrouwen dat de bron inderdaad gegevens aanlevert

conform de afspraken, blijkt in de praktijk geen verstandige keuze. Het controleren van de aangeleverde brongegevens tegen de gemaakte afspraken is essentieel en het bestaansrecht van de validatiemodule.

Verschillen GIGO- en validatiemethode

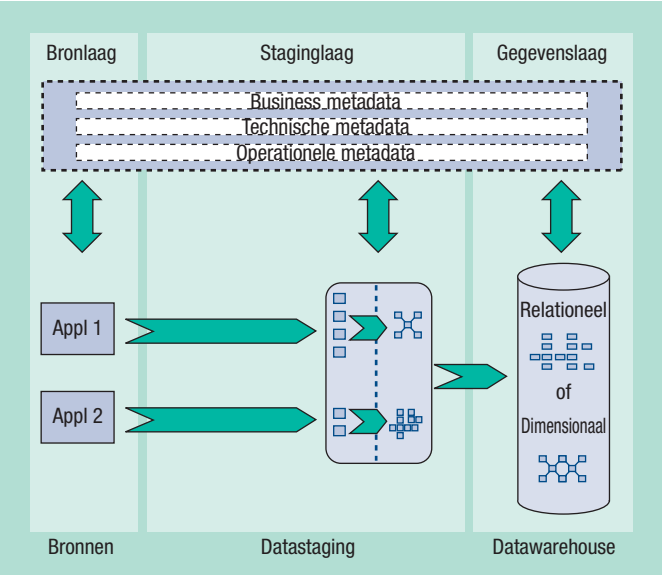
In de tabel staan de verschillen tussen de 'klassieke' GIGO-methode en de validatiemethode kort samengevat. Zoals daarin staat aangegeven blijft de verantwoordelijkheid voor datakwaliteit liggen waar deze hoort, bij de gegevens/proceseigenaar. Echter door de gegevenskwaliteit expliciet te maken en af te dwingen door gebruik te maken van de validatiemodule, ontstaat een juiste beleving van de verantwoordelijkheidsverdeling.

Binnen Business Intelligence-projecten dient expliciet aandacht te worden besteed aan de juiste verantwoordelijkheidsverdeling. In de standaard DWH-architecturen worden gegevens vanuit de bronsystemen via een staging-proces in het datawarehouse geladen. Op dit punt grijpt de validatiemodule in op het standaardplaatje. De plaats van de validatiemodule ten opzichte van een standaard architectuurplaatje, zie afbeelding 1, staat in afbeelding 2. Voordat de reguliere verwerking van de brongegevens plaatsvindt, start de validatiemodule. De validatiemodule kan brongegevens goedkeuren, waarna ze via de staging-processen verder worden verwerkt in het datawarehouse, of brongegevens afkeuren, waarna deze in een uitvalbak terechtkomen. In het vervolg van dit artikel gaan we nader in op het validatieproces binnen een DWH-project.

Het validatieproces

Nu de functionele specificatie is beschreven en de validatiecomponenten bekend zijn, worden de technische stappen van de validatiemodule toegelicht. Om het validatieproces op een adequate manier uit te voeren, is het opgedeeld in drie stappen: de ontvangstvalidatie; de technische validatie; de inhoudelijke validatie. Van elke processtap worden de procesmetadata natuurlijk opgeslagen.

Bij de ontvangstvalidatie wordt 'vanaf de buitenkant' naar de aanlevering gekeken. Hierbij valt te denken aan vragen als: voldoet de aanlevering aan de afgesproken naamgevingconventie; wordt hetzelfde bestand niet dubbel aangeleverd; is de pakbon meegeleverd bij het databestand; is de aanlevering wel door een vertrouwde partij gedaan? Wanneer de ontvangstvalidatie heeft vastgesteld dat het een legitieme aanlevering betreft, wordt de volgende validatiestap doorlopen. De technische validatie is de volgende stap in het validatieproces, waarin wordt gecontroleerd of het aangeleverde bestand



Afbeelding 1: Standaardarchitectuur.

Validatie in de praktijk

Tijdens het vormgeven van een datawarehouse bij een klant werd gediscussieerd over de kwaliteit van de rapportages die uit het oude datawarehouse kwamen. De beleving was dat de opgeleverde informatie slecht was en dat de bronsystemen vol zaten met fouten. Aan de andere kant liepen de primaire klantprocessen goed. De kwaliteit van de data was dus blijkbaar toch niet zo slecht.

Na uitleg van het concept validatiemodule en het vastleggen van kwaliteitsafspraken werd gesteld dat het gebruik van een validatiemodule onwerkbaar was. Immers, gegevensafnemers stelden dat 100 procent kwalitatief goede data nodig waren. Het afdwingen hiervan zou leiden tot een situatie waarbij geen data konden worden opgenomen in het systeem. Deze constatering is volkomen juist. Als je namelijk data van 100 procent goede kwaliteit wilt ontvangen en deze is aantoonbaar niet beschikbaar, dan krijg je ook niets. Het is zelfs kwalijk te noemen als je gegevens ontvangt in de veronderstelling dat ze goed zijn, maar als vervolgens in de praktijk blijkt dat deze gegevens niet de werkelijkheid zijn.

De oplossing ligt niet in het accepteren van data die aantoonbaar niet voldoen aan de kwaliteitsverwachting van de eindgebruiker. Verlaging van het kwaliteitsambitieniveau of verhoging van de kwaliteit in de bron is wel de oplossing. Vervolgens ontstond binnen de organisatie een discussie over kwaliteit, doel van registratie van gegevens en gebruik van gegevens voor andere doelen. Daarnaast is veel aandacht besteed aan het minimaal noodzakelijke kwaliteitsniveau voor gebruik van data binnen het datawarehouse en het maximale kwaliteitsniveau dat de bronsystemen kunnen leveren.

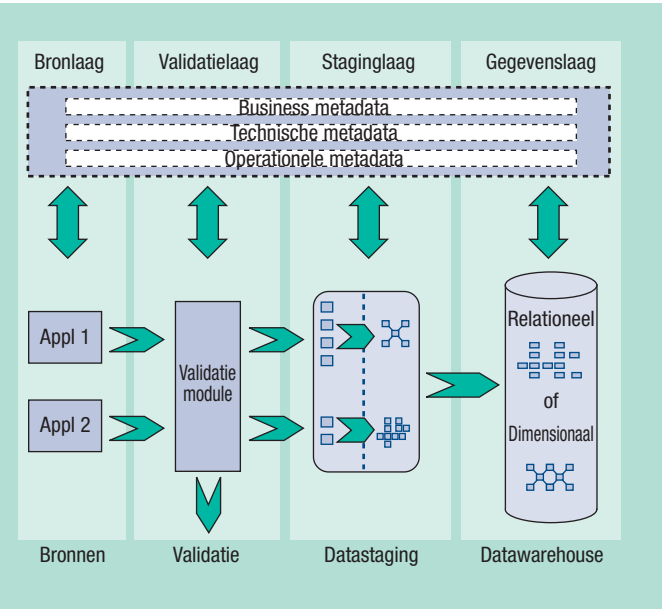
Na het vastleggen en het accorderen van datakwaliteitsafspraken in gegevensleveringsovereenkomsten is het datawarehouse gebouwd. De kwaliteit van de uiteindelijk geleverde data is niet 100 procent juist, maar de eindgebruiker weet waar hij/zij op kan rekenen. De data voldoen wel 100 procent aan de gestelde minimale kwaliteitseisen van de eindgebruiker. Het resultaat van het proces is echter wel dat de gegevens waarop de business haar beslissingen baseert een vooraf vastgesteld kwaliteitsniveau moeten hebben.

technisch correct is, conform de afspraken in de gegevensleveringsovereenkomst. Dit zijn onder andere: klopt de aanlevering met het afgesproken veldscheidingsteken; zijn aangeleverde datums correcte bestaande datums; verliezen we geen informatie bij het inlezen van de aanlevering? Wanneer ook deze stap correct is verlopen, vindt de inhoudelijke validatie plaats.

> In de metadata is vastgelegd op welke locatie bronbestanden aangeleverd mogen worden

Tijdens de inhoudelijke validatie worden alle functionele controles uitgevoerd waaraan de aanlevering moet voldoen. Deze controles kunnen zeer divers zijn. Hierbij valt te denken aan: de inhoud van de kolom 'geslacht' mag alleen 'M' of 'V' zijn; het percentage marktaandeel mag niet boven 100 uitkomen; het opgegeven tijdstip mag tussen 00:00 en 23:59 liggen, of gelijk zijn aan 9999 (onbekend); het totaal van de kolom 'omzet' moet gelijk zijn aan de gegevens op de pakbon; de kolom 'reden korting' moet zijn gevuld als het kortingbedrag niet 0 is. Het resultaat van de inhoudelijke validatie kan drieledig zijn:

- De gegevens zijn conform de gegevensleveringsovereenkomst, de aanlevering wordt goedgekeurd;
- De gegevens zijn niet correct, de aanlevering wordt afgekeurd;
- De gegevens zijn niet correct, maar mogen toch worden verwerkt in het datawarehouse. Dat deze gegevens niet correct zijn, is niet essentieel voor de verdere verwerking en kan later met een nieuwe aanlevering worden gecorrigeerd.



Afbeelding 2: Architectuur met validatiemodule.

```
Select          rownum, <1>
From            <2>
Where           <3> like '<4>'
```

Afbeelding 3: Voorbeeld van standaard validaties.

Metadata

Metadata zijn een essentieel onderdeel van de validatiemodule, want zonder metadata wordt de onderhoudbaarheid van de tool erg arbeidsintensief en kan deze niet universeel worden ingezet. Het gaat om twee typen metadata, waarvan we in de validatietool gebruik maken.

Procesmetadata. Opslag van alle validatieresultaten vindt plaats in de metadata. Hierin worden niet alleen technische metadata over verwerkingstijden, aantallen records opgeslagen, maar ook business metadata. Alle resultaten van de processtappen zijn na de validatie beschikbaar in de metadata. Bij het constateren van validatiefouten wordt zowel een technische foutmelding geregistreerd voor de beheerders, als een functionele omschrijving van de foutmelding voor de gebruikers.

Op deze manier is er een sluitende administratie over wat is aangeleverd, welke controles zijn uitgevoerd en waarom een aanlevering eventueel is afgekeurd. Op basis van deze procesmetadata worden rapportages verzorgd. Er zijn rapportages voor de eindgebruikers van het datawarehouse, zodat zij weten welke gegevens zijn verwerkt in het datawarehouse. Ook zijn er rapportages voor de aanleverende bronsystemen, zodat zij weten welke gegevens nog gecorrigeerd en opnieuw aangeleverd dienen te worden.

Stuurmetadata. De validatiemodule dient te zijn opgezet als een flexibele, breed inzetbare en goed beheerbare oplossing. Aan de hand van een voorbeeld simuleren we hieronder de werking van de validatiemodule.

In de metadata van de ontvangstvalidatie is vastgelegd op welke locatie bronbestanden aangeleverd mogen worden. Er loopt voortdurend een proces om te bepalen of er aanleveringen zijn binnengekomen. Wanneer er een nieuwe aanlevering is, start de validatie en wordt de aanlevering vergeleken met naamgevingconventies, zoals die in de metadata zijn vastgelegd. Wanneer een aanlevering voldoet aan de naamgevingconventies vindt het proces van technisch valideren plaats. Voor de technische validatie is een bewuste afweging gemaakt om deze niet volledig metadatagestuurd te ontwikkelen. Hier wordt gebruik gemaakt van een ETL-tool die ook voor het datawarehouse wordt gebruikt. Door deze aanpak kan een ultieme flexibiliteit worden geboden aan in te lezen bronbestandformaten. Het resultaat van deze technische validatie is een ingelezen bronaanlevering die is opgeslagen in de database van de validatiemodule.

	GIGO	Validatiemodule
Wie is verantwoordelijk voor datakwaliteit?	Gegevens/proceseigenaar. In de praktijk is echter niemand verantwoordelijk. De datakwaliteit van het bronsysteem wordt gevolgd.	Gegevens/proceseigenaar.
Is de organisatie bekend met de processen rondom datakwaliteit?	Nee, buiten scope.	Ja, expliciet benoemd.
Is de organisatie bekend met kwaliteit van de output van het datawarehouse?	Nee, technische verwerking is de norm.	Ja, op basis van kwaliteitseisen.
Wat is de verantwoordelijkheid van het projectteam?	Morele verantwoordelijkheid voor juistheid van data. Verantwoordelijkheden zijn niet duidelijk belegd.	Formele verantwoordelijkheid voor het valideren van de afspraken die zijn vastgelegd in de gegevensleveringsovereenkomst.
Wat is de verantwoordelijkheid van de beheerorganisatie?	Implementeren van business rules om de datakwaliteit te verbeteren.	Het uitbreiden van de gegevensleveringsovereenkomst met vernieuwde datakwaliteitinzichten en het implementeren van deze nieuwe validaties.
Lost deze aanpak de problemen rondom datakwaliteit op?	Nee.	Ja, bij juist acteren van de organisatie.

Tabel 1. Verschillen tussen GIGO en het gebruik van de validatiemodule.

De volgende stap is het uitvoeren van de inhoudelijke validaties. Voor verschillende bronbestanden is veelal eenzelfde soort validatie nodig. De validatiemodule beschikt over generieke, parametergestuurde validaties. Dit zijn validaties in een vorm zoals in afbeelding 3 te zien is. Voor elke bronaanlevering is in de metadata vastgelegd wat de parameters zijn die voor deze aanlevering gebruikt moeten worden. Deze technische controle levert als resultaat alle rijen die niet voldoen aan de opgegeven kwaliteitseisen uit de gegevensleveringsovereenkomst.

> Met de traditionele GIGO-aanpak werd de datakwaliteit onvoldoende bewaakt

Wanneer de technische controle onjuiste gegevens signaleert, wordt op basis van de in de metadata vastgestelde regels de aanlevering afgekeurd en in de uitvalbak geplaatst, of, indien expliciet vermeld in de gegevensleveringsovereenkomst, toch doorgegeven aan de staging-processen van het datawarehouse. Door het inzetten van de validatiemodule groeit het kwaliteitsbesef binnen een organisatie enorm. Dit is niet altijd een gemakkelijk proces. Men verwacht of hoopt dat kwaliteit geen issue is, of op zijn minst dat het het probleem van iemand anders is. Elke partij in de keten van brongegevens, via datawarehouse, naar gebruik, heeft een eigen verantwoordelijkheid. Doordat deze verantwoordelijkheden nu ook in duidelijke afspraken worden vastgelegd en gecontroleerd, wordt men ook aangesproken op de eigen verantwoordelijkheden.

Nee zeggen

Het optimaliseren en garanderen van hoge datakwaliteit is verschoven van een 'could have' naar een 'must have' volgens de MOSCOW-analyse. Om een optimale datakwaliteit te kunnen waarborgen is 'nee zeggen' tegen brondata die niet overeenkomstig zijn met de opgestelde specificaties wél klantvriendelijk. 'Nee zeggen' tegen de brondata van de klant is uiteraard niet vanzelfsprekend, vandaar dat het functionele validatieproces een verschuiving van verantwoordelijkheden met zich meebrengt.

Met de traditionele GIGO-aanpak werd de datakwaliteit onvoldoende bewaakt, lagen de verantwoordelijkheden vaak bij de verkeerde mensen en werd het validatieproces onvoldoende vastgelegd. Bij de aanpak met gebruik van de validatiemodule zijn alle verantwoordelijkheden duidelijk vastgelegd en bekend bij de betrokkenen. Op deze manier is het mogelijk om 'nee' te zeggen tegen een bronaanlevering en kan men er de broneigenaar op aanspreken.

Robert Allewijn, Bas Pruijn en Vincent Wylenzek
Drs. R. Allewijn is senior projectmanager bij Ordina VisionWorks.
Drs. B. Pruijn is senior consultant (architect) bij Ordina VisionWorks.
Ing. V. Wylenzek is medior consultant (BI) bij Ordina VisionWorks.